A Multiple Test Correction for Streams and Cascades of Statistical Hypothesis Tests

Geoffrey I. Webb
Faculty of Information Technology
Monash University
Clayton, Australia
geoff.webb@monash.edu

François Petitjean
Faculty of Information Technology
Monash University
Clayton, Australia
francois.petitjean@monash.edu

ABSTRACT

Statistical hypothesis testing is a popular and powerful tool for inferring knowledge from data. For every such test performed, there is always a non-zero probability of making a false discovery, i.e. rejecting a null hypothesis in error. Familywise error rate (FWER) is the probability of making at least one false discovery during an inference process. The expected FWER grows exponentially with the number of hypothesis tests that are performed, almost guaranteeing that an error will be committed if the number of tests is big enough and the risk is not managed; a problem known as the multiple testing problem. State-of-the-art methods for controlling FWER in multiple comparison settings require that the set of hypotheses be predetermined. This greatly hinders statistical testing for many modern applications of statistical inference, such as model selection, because neither the set of hypotheses that will be tested, nor even the number of hypotheses, can be known in advance.

This paper introduces Subfamilywise Multiple Testing, a multiple-testing correction that can be used in applications for which there are repeated pools of null hypotheses from each of which a single null hypothesis is to be rejected and neither the specific hypotheses nor their number are known until the final rejection decision is completed.

To demonstrate the importance and relevance of this work to current machine learning problems, we further refine the theory to the problem of model selection and show how to use Subfamilywise Multiple Testing for learning graphical models.

We assess its ability to discover graphical models on more than 7,000 datasets, studying the ability of Subfamilywise Multiple Testing to outperform the state of the art on data with varying size and dimensionality, as well as with varying density and power of the present correlations. Subfamilywise Multiple Testing provides a significant improvement in statistical efficiency, often requiring only half as much data to discover the same model, while strictly controlling FWER.

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939775

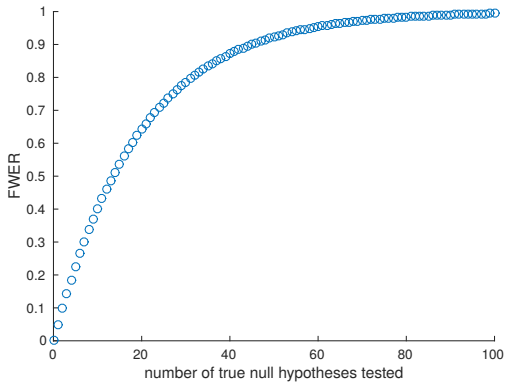

Figure 1: Evolution of the FWER as a function of the number of tested true null hypotheses if no multiple testing correction procedure is used with significance level $\alpha = 0.05$.

Keywords

Hypothesis testing, multiple testing, model selection

1. INTRODUCTION

Statistical hypothesis testing was developed in a context of assessing a single proposition with respect to some data while strictly controlling the risk of type 1 error – the risk of rejecting the null hypothesis, and thus *incorrectly* accepting the hypothesis of interest. This is an extremely powerful approach to knowledge discovery from data. However, in the digital age, where data are almost omnipresent, it is often desirable to employ this approach to test not just a single proposition but massive numbers of propositions. In this case it is not sufficient to simply control in isolation the risk of type one error for each null hypothesis, as even if the individual risk is low, the cumulative risk rapidly approaches near certainty; as shown in Figure 1 which plots the evolution of the FWER as the number of independent true null hypotheses tested m increases. This has led to multiple testing procedures, that can directly control the Familywise Error Rate (FWER), the cumulative risk of any type 1 error when testing a set of null hypotheses [2, 3, 9, 14].

However, these powerful techniques are restricted in that the set of null hypotheses (or at least an upper bound on their number) must be known in advance. This limits their applicability in a growing range of applications where there

is a need to test some hypotheses before other hypotheses are known (*hypothesis streams*) and the special case of these streams where the rejection of one hypothesis changes the null hypotheses to be tested subsequently (*hypotheses cascades*).

One context in which hypothesis streams are encountered is online controlled experiments, where many tests are conducted in parallel, with new tests commencing while others are still in progress [19].

Hypothesis streams and cascades also confront scientific research [34]. Scientists increasingly accumulate large datasets at great expense, and it is imperative that these be used to maximum advantage. As a result, it is important to be able to undertake statistical testing of initial hypotheses, and then follow these up with further tests that arise from the initial discoveries.

A further complication is that many datasets are continually growing. It is not possible to wait until all astronomical data are collected, for example, because new observations are made all the time. Nonetheless, we want to be able to derive conclusions from the data. We also want to be able to reconsider hypotheses in the light of additional data as it becomes available. Further, conclusions that we reach with respect to initial hypotheses are likely to lead to the formulation of further hypotheses that we also want to test. If we just keep testing null hypotheses without controlling for the multiple testing problem, then eventually some will be rejected by chance. There is a pressing need for statistical methods that can support science in the age of big data.

Hypothesis streams and cascades are also serious challenges in many data science contexts, including forward stepwise feature selection [8], backward stepwise feature elimination [8], association discovery [38, 39] and model selection [27]. We use this latter application as a case study.

Let us motivate why hypotheses cascades can occur in model selection. Graphical model forward selection traditionally starts with a reference model over n nodes (one node per variable) with no edges. At step s of the process, the reference model contains s edges, and the addition of one edge among e_s has to be considered.[1] The number of possible steps is equal to the maximum number of edges $M = n \cdot (n-1)/2$, and at each step s, there are $M - s + 1$ edges to consider. In consequence, the greatest number of models that we may need to test is $M \cdot (M+1)/2$, which means that discovering a model with $n = 1,000$ leads to the possibility of more than 100 billion tests to be performed, and thus to control against.

This highlights several key motivations of this work:

1. With more than 100 billion tests, even using a recommended flat critical value $\alpha = 0.001$ [17], this is potentially 100 million hypotheses that could correspond to false discoveries. Therefore, controlling for multiple testing is critical.

2. It would be more natural to consider the set of hypotheses as a stream than as a complete set, because many of the hypotheses might never be considered. We are in fact very likely to stop at step $s_k, k \ll M$, which means that controlling for all the hypotheses that might have been present after step s_k is needlessly strict and as we will show, often prevents the acceptance of many true discoveries. We only want to control for the risk with respect to hypotheses that are actually assessed.

   In consequence, the multiple testing procedure should be able to consider new hypotheses to be tested that were not present at the start of the process, without having any knowledge about the hypotheses and their number in advance.

This paper introduces Subfamilywise Multiple Testing (SMT): a multiple testing correction procedure that strictly controls FWER in settings where we have no prior knowledge about the set of hypotheses that will be tested or of their number.

[1]It is important to note that the addition of edge (a, b) at step s and at step s' are two separate hypotheses, because the reference models to which (a, b) would be added is different; we will detail this element in Section 4.

2. SUBFAMILYWISE TESTING

2.1 Problem statement

Let $\mathcal{F} = \bigcup_{i=1}^{n} \mathcal{F}_i$ be a family of n subfamilies of null hypotheses. For any null hypothesis $h \in \mathcal{F}$, let κ_h be an associated test statistic and $p_h = Px(\kappa_h \mid istrue(h))$ be the probability of obtaining the test statistic κ_h or more extreme if h were true, where $Px(\kappa_h \mid istrue(h))$ represents the probability of obtaining the test statistic or more extreme under some predetermined probability distribution for h. Classical hypothesis testing rejects each null hypothesis h if $p_h \leq \alpha$, where α is a predefined critical value.

The probability of making a false discovery when testing a unique null hypothesis h (*i.e.*, the probability of rejecting h when it is true) is no greater than $Pr(h)p_h$, the prior probability that the null hypothesis is true times the p-value p_h. Thus, p_h is an upper bound on the probability of a false discovery when a single hypothesis is tested.

Definition 1. *The familywise error rate (FWER), is the probability of making at least one false discovery — or Type I error — when testing family $\mathcal{F}$. Let $\mathcal{R} \subseteq \mathcal{F}$ be the rejected null hypotheses and $\mathcal{T} \subseteq \mathcal{F}$ be the true null hypotheses. We have*

$$FWER = Pr(\mathcal{R} \cap \mathcal{T} \neq \emptyset) = 1 - Pr(\mathcal{R} \cap \mathcal{T} = \emptyset). \quad (1)$$

A multiple testing correction procedure provides *strict control over the FWER* if it is a function from families of null hypotheses (together with their associated test statistics) and a critical value to a subset of the family, $\mathcal{F}, \alpha \rightarrow \mathcal{R} \subseteq \mathcal{F}$ such that no matter which subset of the null hypotheses is true, the probability of rejecting any true null hypothesis is less than α —

$$Pr(\mathcal{R} \cap \mathcal{T} \neq \emptyset) \leq \alpha. \quad (2)$$

We wish to control FWER while iterating over the subfamilies of $\mathcal{F}$, selecting a single null hypothesis for rejection from each subfamily. When considering subfamily $\mathcal{F}_i$ we have knowledge of the subfamilies previously considered, but no information about subsequent subfamilies to be encountered. We allow that the selection of the null hypothesis for rejection in $\mathcal{F}_i$ might determine the null hypotheses that are contained in subsequent subfamilies, as is the case with forward sequential model selection, where the components considered for subsequent inclusion in a model may depend on the components already included.

2.2 SMT Procedure

THEOREM 1. *Let $h_i^{\min}$ be the null hypotheses in $\mathcal{F}_i$ with the lowest p-value, $h_i^{\min} = \arg\min_{h \in \mathcal{F}_i}(p_h)$. Let $\mathcal{R}$ contain the null hypothesis with the lowest p-value from each of the first r subfamilies, $\mathcal{R} = \{h_1^{\min}, \ldots, h_r^{\min}\}$. Let $p_i^{\min} = p_{h_i^{\min}} = \min_{h \in \mathcal{F}_i}(p_h)$ be the minimum p-value of a null hypotheses in $\mathcal{F}_i$.*

$$Pr(\mathcal{R} \cap \mathcal{T} \neq \emptyset) \leq \sum_{i=1}^{r} p_i^{\min} \cdot |\mathcal{F}_i|. \tag{3}$$

PROOF.

$$Pr(\mathcal{R} \cap \mathcal{T} \neq \emptyset) = Pr\left(\bigvee_{i=1}^{r} \bigvee_{h \in \mathcal{F}_i \cap \mathcal{T}} p_h \leq p_i^{\min}\right) \tag{4}$$

$$\leq \sum_{i=1}^{r} \sum_{h \in \mathcal{F}_i \cap \mathcal{T}} Pr(p_h \leq p_i^{\min}) \tag{5}$$

$$\leq \sum_{i=1}^{r} \sum_{h \in \mathcal{F}_i \cap \mathcal{T}} p_i^{\min} \tag{6}$$

$$\leq \sum_{i=1}^{r} \sum_{h \in \mathcal{F}_i} p_i^{\min} \tag{7}$$

$$= \sum_{i=1}^{r} p_i^{\min} \cdot |\mathcal{F}_i|. \tag{8}$$

□

Comments. (4) recasts the probability of FWER as the probability of any p-value being less than or equal to $p_i^{\min}$ for any true null hypothesis in any subfamily $\mathcal{F}_i$ from which a null hypothesis is rejected. (5) follows from the consequence of the general disjunction rule that the probability of a disjunction of events cannot exceed the sum of the probabilities of the individual events. For any true null hypothesis h, the probability that the p-value from a valid test procedure is less than or equal to a given p is no greater than p. This justifies (6). The sum over a set of probabilities can be no less than the sum over a superset of those values, justifying (7). (8) re-expresses the equation in terms of $|\mathcal{F}_i|$.

The Subfamilywise Multiple Testing Procedure (SMT) provides strict control over the FWER in a multiple testing situation where there are successive pools of null hypotheses from each of which a single null hypothesis is to be rejected. The procedure has two steps. Step 1 finds

$$r^\star = \arg\max_{r}\left(\sum_{i=1}^{r} p_i^{\min} \cdot |\mathcal{F}_i| \leq \alpha\right). \tag{9}$$

Step 2 rejects $h_1^{\min}, \ldots, h_{r^\star}^{\min}$. That this procedure strictly controls FWER to be no greater than α follows directly from Theorem 1.

3. MONTE CARLO SIMULATIONS

To elucidate the statistical power of the technique, we conducted Monte Carlo simulations, generating sets of null hypotheses, which were randomly assigned to be either true or false and were randomly assigned simulated p-values. These simulations were governed by three parameters — *subfamilySize*: the size of each subfamily; *pTrue*: the probability that a null hypothesis should be designated to be true; and *maxFalsePVal*: the maximum simulated p-value to be assigned to a false null hypothesis.

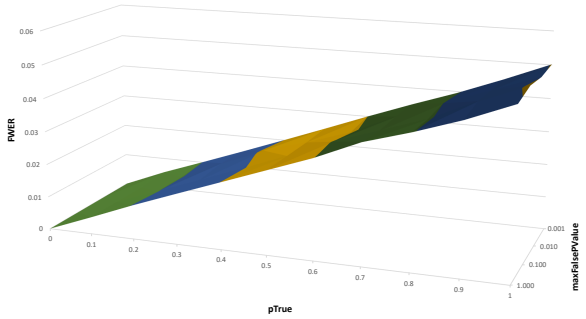

Figure 2: The FWER of SMT as the relative frequency of true to false null hypotheses is increased and the relative p-values of false relative to true null hypotheses is decreased

The following is the procedure used for each simulation.

$\mathcal{R} \leftarrow \emptyset$
$sumP \leftarrow 0.0$
$i \leftarrow 0$
repeat
  $i \leftarrow i + 1$
  Generate $\mathcal{F}_i$
  $sumP \leftarrow sumP + |\mathcal{F}_i| \cdot p_i^{\min}$
  if $sumP \leq \alpha$ **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{h_i^{\min}\}$
  end if
until $sumP > \alpha$

To generate each $\mathcal{F}_i$, *subfamilySize* simulated null hypotheses were generated. Each was designated as either true or false, with probability *pTrue* of being designated true. Each true null hypothesis was assigned a simulated p-value drawn uniformly at random from $[0, 1]$ and each false null hypothesis was assigned a simulated p-value drawn uniformly at random from $[0, maxFalsePVal]$. Having lower p-values for false null hypotheses simulates the use of a test statistic that is useful for discriminating between true and false null hypotheses.

We performed two experiments. In the first experiment *subfamilySize* was set to 100, *pTrue* was varied from 0.00 to 1.00 in steps of 0.1 and *maxFalsePVal* was set to each of the values 1.0, 0.1, 0.01 and 0.001, creating a total of 44 treatments. 100,000 Monte Carlo simulations were conducted for each treatment and the FWER and average number of true discoveries per simulation determined.

Figure 2 presents a surface chart that shows the impact on FWER as the relative frequency of true to false null hypotheses is increased and as the p-values of false null hypotheses decrease relative to those of true null hypotheses. As the probability of a true null hypothesis increases, so too does the FWER. This is to be expected, as FWER must be zero when all null hypotheses are false and should be more likely when all null hypotheses are true. Indeed, when all null hypotheses are true, FWER only occurs for family $\mathcal{F}$ if a null hypothesis is rejected from its first subfamily, $\mathcal{F}_1$.

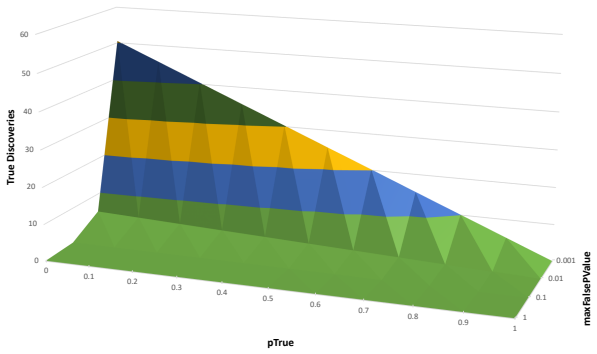

Figure 3: The true discoveries of SMT as the relative frequency of true to false null hypotheses is increased and the relative p-values of false relative to true null hypotheses is decreased

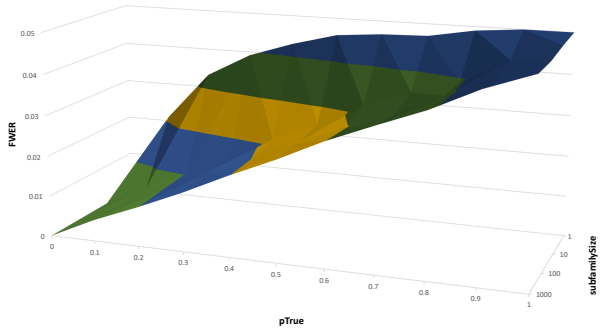

Figure 4: The FWER of SMT as the relative frequency of true to false null hypotheses is increased and the relative p-values of false relative to true null hypotheses is decreased

As the test for the first subfamily is a conventional Bonferroni correction, the FWER must be strictly controlled when all null hypotheses are true.

In contrast, the relative p-values of true and null hypotheses (that is, the power of the simulated test statistic) has little impact on FWER.

Figure 3 shows the effect on true discoveries of the same factors. As can be seen, true discoveries increase as the ratio of false to true null hypotheses increases and as the relative p-value of a false null hypothesis decreases. Recall that true discoveries occur when a false null hypothesis is rejected and hence the true alternative hypothesis is accepted. Clearly there can be no true discoveries when all null hypotheses are true. Conversely, when all null hypotheses are false, all rejected null hypotheses will be true discoveries. In all cases where there are false null hypotheses, the number that are rejected will be determined by their relative p-values, which is why the number of rejections rises as $maxFalsePVal$ decreases.

As $maxFalsePVal$ had negligible effect on FWER, in the second experiment $maxFalsePVal$ was set to 0.01, $pTrue$ was again varied from 0.00 to 1.00 in steps of 0.05 and $subfamilySize$ was set to each of the values 1, 10, 100 and 1,000, again creating a total of 44 treatments. As in the first experiment, 100,000 Monte Carlo simulations were conducted for each treatment and the FWER and average number of true discoveries per simulation determined.

Figure 4 presents a surface chart showing the effect on FWER as the relative frequency of true to false null hypotheses is increased and as the subfamily size varies. Again, when pTrue is 1.0 and FWER is determined by whether a null hypothesis is rejected for the first subfamily or not, the probability of FWER is strictly controlled by the equivalent of a Bonferroni correction for the first subfamily. Like the first experiment, and for the same reasons, FWER falls as the proportion of false null hypotheses rises.

However, the speed at which FWER falls is greatly affected by the subfamily size. Consider a scenario in which each subfamily contains only a single null hypothesis. If the first null hypothesis is true, the probability of FWER will be exactly α. If the first null hypothesis is false and the second is true, the probability of FWER will be exactly $\alpha/2$, and so on. Hence, when the probability of each null hypothesis being true is high, FWER will be relatively high.

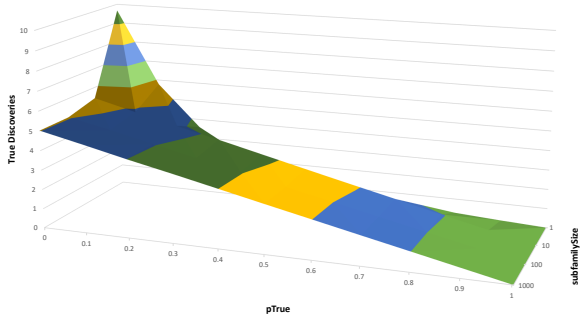

Figure 5: The true discoveries of SMT as the relative frequency of true to false null hypotheses is increased and the relative p-values of false relative to true null hypotheses is decreased

In contrast, as the size of the subfamilies increases, the probability of a subfamily containing false as well as true null hypotheses increases, and the probability increases that the hypothesis with the lowest p-value is false. As familywise error only occurs when the lowest p-value for a subfamily belongs to a true null hypothesis, the FWER correspondingly decreases.

Figure 5 shows the effect on true discoveries of the same factors. Again, true discoveries increase as the ratio of false to true null hypotheses increases. When the probability of each null hypotheses being false is low, true discoveries are highest when the subfamilies are largest, as this gives the greatest chance of each subfamily containing some false null hypotheses to correctly reject. However, as the ratio of false to true null hypotheses increases, the smaller subfamilies begin to reject the greatest number of false null hypotheses, as most subfamilies will include false null hypotheses for rejection, but the smaller subfamilies result in much higher adjusted critical values, making it more likely that the false null hypotheses will be rejected.

These simulations demonstrate the power of SMT, and show that it is most effective when the ratio of false to true hypotheses is highest and the relative p-values of false null

hypotheses is lowest. The impact of subfamily size is less straightforward, however, depending on the other factors.

4. SMT FOR MODEL SELECTION

We demonstrate the power of SMT in the context of hill-climbing search for graphical model selection, specifically, where it is used to select interactions for inclusion in a log-linear model within the Chordalysis system [27, 25, 26]. We have made the source-code available for SMT within Chordalysis at https://github.com/fpetitjean/Chordalysis.

4.1 Background

The use of statistical tests and multiple correction methods for learning the structure of graphical models responds to the need for explainable models (see e.g. [22] for genomics). Although long recognized by the data mining community, this highlights again that building an *explanatory* model from data has a different objective to building a *predictive* model [31]. Learning graphical models for which we place greater weight on being confident in the structure is often called log-linear analysis in the statistical community [4], which comes from the fact that the search is often performed among log-linear models. Log-linear models that are graphical are equivalent to Markov Random Fields (or Markov Networks).

Both statistical and machine learning communities have studied the use of statistical methods to learn the structure of graphical models, from the well-known PC algorithm [18, 32], to recent work scaling statistical procedures to high-dimensional data [5, 27, 40].

It is here also interesting to mention MML/MDL methods [28, 37] which, by building on Shannon theory of information, often produce an explanatory model. This is intuitively because every additional parameter has to be explained away by enough data. Several works have studied these approaches for learning Bayesian Networks and Markov Random Fields [1, 6, 25, 30].

It is finally interesting to mention methods based on ℓ_1-regularizers, because they are often claimed to produce understandable models, by biasing the search towards models for which many parameters are zero. Different configurations have been studied: performing a logistic regression for every variable independently [36], focusing on a reduced subset of features [20] or finding a set of variables that best divides the graph [11]. Note that these methods place great weight on the predictivity of the model, which often leads to a significant number of false discoveries (see for example the precision trend depicted in [36] – Section 6). For this reason, we will not consider these methods as being directly related to statistical testing methods.

4.2 Using SMT for model selection

The refined algorithm is given in Table 1. We start with a reference model $\mathcal{M}$ corresponding to the independence model (no edges). We then collect the set E of all potential edges at this stage, *i.e.*, all the $V(V-1)/2$ edges. We then iterate until either there are no edges in E that can be added to the current reference model $\mathcal{M}$ (which happens if all edges have been added), or the addition of the best edge does not satisfy our multiple correction. The process can then be compared to the use of a budget of FWER risk. Starting with the budget being α, at each iteration, we remove from the budget the p-value associated with the accepted edge times the number of concurrent hypotheses that were assessed at the same time as the best edge.

Table 1: Forward selection with SMT

Require: α: the requested FWER
Require: $\mathcal{D}$: a dataset over V variables
Require: $pval(\mathcal{M}, edge)$: a statistical testing procedure returning the p-value associated with the addition of an edge to a model
$\mathcal{M} \leftarrow (V, \emptyset)$ {independence model (empty graph)}
$E \leftarrow \{(v_1, v_2) : \forall v_1, v_2 \in V, v_1 \neq v_2\}$
$budget \leftarrow \alpha$
while $E \neq \emptyset$ **do**
  bestEdge $\leftarrow \arg\min_{e \in E} pval(\mathcal{M}, e)$
  $p \leftarrow pval(\mathcal{M}, bestEdge)$
  $\alpha' = budget/|E-1|$
  if $p > \alpha'$ **then**
    return $\mathcal{M}$ {would not be valid to add bestEdge}
  end if
  $E \leftarrow E \setminus bestEdge$
  $\mathcal{M} \leftarrow \mathcal{M} \cup bestEdge$
  $budget \leftarrow budget - p \cdot |E|$
end while

4.3 Experiments with synthetic data

Assessing the quality of model selection techniques requires having knowledge about the multi-way interactions that take place in data. Therefore we start by evaluating the discovery with data that is sampled from known distributions (sets of interactions and associated probability tables). This allows us to compare the discovered interactions to the true structure from which the data was sampled.

Then, we show some results on real-world data.

4.3.1 Description of experiments

The task is, given some categorical data, to find the set of correlations that were planted in the distribution from which the data was drawn; all in an unsupervised manner. There are two components to generating the data; first we have to choose a graph structure that describes those correlations, and then we have to parametrize the graphical model. Note that, to ensure reproducibility of the experiments described below, we have made the source-code for generating the parameterized models and data available at http://bit.ly/SourceDataGeneration.

Generation of the data.

To form graphs that are representative of the real-world, we generate random scale-free graphs using the Barabási–Albert (BA) model. This has desirable properties including that the degree distribution follows a power law. BA models are parameterized with (1) the number of nodes (i.e. the number of variables) and (2) the degree, which controls the edge density in the graph. We add an additional condition that the graph be chordal, in order to match the class of models explored by the model selection method that we use: Chordalysis.

Having the graph structure, all we need to generate data is to parameterize the associated graphical model. Chordal graphs each correspond to an equivalence class of Bayesian Networks (BN); we can then use standard procedures for BN parametrization. Each line of every Conditional Probability

Table (CPT) in the network encodes a multinomial distribution. All we need is to control for the strength of the encoded correlations. To this end, we use a flat Dirichlet prior with concentration S (Dir($\vec{S}$) for each multinomial). The higher the value of α, the closer the multinomial will be to a uniform distribution, and hence the more subtle the correlation will be.

Finally, the last parameter corresponds to the number of samples to generate. Each data sample is generated independently by sampling the associated Bayesian Network and using a secure pseudo-random number generator (SHA1PRNG).

We thus have four parameters for each experiment:

1. V the number of variables
2. D the degree/density of the edges in the graph
3. S the subtlety of the correlations (inverse of strength)
4. N the number of samples

To assess the statistical efficiency of our proposed SMT, we vary each of those parameters in turn, having set the others to values we assessed as being reasonable: $V = 100$, $D = 3$, $S = 200$ and $N = 50,000$.

Evaluation measure.

The task is unsupervised recovery of the edges[2] that were introduced in the graphical model that generated the data. The number of true/false positives/negatives are thus sufficient statistics for the evaluation; we can then study the *precision*, which is the proportion of discoveries that are true discoveries (rejected null hypotheses that are false), and *recall*, which is the proportion of false null hypotheses that are rejected. The precision obtained across all experiments is consistently 100%; for clarity we thus only report the recall and made all precision graphs available at http://bit.ly/SyntheticResults for cross-check.

We generated over 7,000 datasets in order to study the influence of each parameter on statistical efficiency; each experiment is run 20 times. We report mean and standard deviation.

We compare our novel multiple testing procedure (SMT) to the state of the art:

1. Bonferroni correction [3]
2. *layered critical values* introduced in [39] and integrated in Chordalysis in [27]; we then use different layering 'consuming', at each step either 50%, 1% or 1‰ of the budget. Note that this reformulation in terms of budget is one of the novel contributions of this paper; layered critical values typically uses 50% [27].
3. an MDL/MML scoring technique [25]; MDL and MML are known for their low false-discovery rates, because intuitively, each parameter has to be justified in terms of gain in the compression of the data.

We use a critical value of $\alpha = 0.05$; note that we have also ensured that our results hold for $\alpha = 0.001$ (as per recent recommendations in [17]).

[2]Note that the original graph structure is undirected, actually forming an equivalence class of Bayesian Networks. The use of any BN within this class would result in the same joint probability under maximum likelihood.

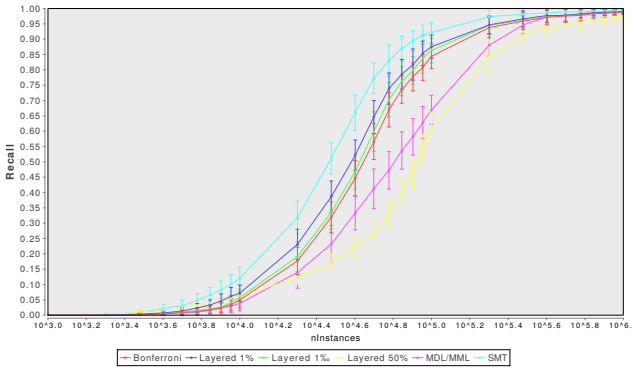

Figure 6: Study of the variation on the number of instances. Note that the x-axis is in log-scale.

4.3.2 *Varying the number of instances*

We vary N from $1,000$ to $1,000,000$. Results are depicted in Figure 6. Recall generally increases as the data quantity increases, as more data provides more evidence of the planted correlations. SMT uniformly dominates state-of-the-art methods by recovering significantly more of the true structure. The performance of the remaining approaches is uniformly ordered Layered-1%, Layered-1‰, Bonferroni, MML/MDL and then Layered-50%.

We can observe from these results that most parameterizations of the layered correction can beat the full Bonferroni correction. This was to be expected given that the Bonferroni gives a uniform weight to any test in the search, while layering puts more weight on the start of the search. Setting the right layering value is however extremely complicated, because it corresponds to giving a prior as per the number of edges to be found and the distributions of their p-values. For instance, if only few edges are to be discovered, then a high layering value will perform well, while low values should be preferred if many edges are to be discovered. This however raises an important obstacle to the discovery, because there is no way to correctly set this value. By contrast, our proposed SMT does not require to set any parameter, and also significantly dominates all other methods. We will see in the next section that this superiority is substantial for real-world data where the amount of data is fixed.

4.3.3 *Varying the number of variables*

We vary V from 10 to $1,000$. Results are depicted in Figure 7. Recall generally decreases with the number of variables, which is due to the fact that the density is kept constant; a higher number of variables translates into more and higher-order correlations to be found.

Here again, SMT dominates Layered and MML, with recall declining as the number of variables increases. It is however interesting to note that while SMT uniformly dominates, the relative positioning of other methods differs depending on the number of variables. For instance, MML/MDL performs quite poorly in this experiment when the number of variables is low, but substantially beats Layered-50% from 40 ($10^{1.6} \approx 40$) variables. Similarly, Layered-1% performs well for medium size graphs, but is outperformed by Layered-1‰ when the number of variables is high. This is because with more variables and the density kept constant, there are more correlations to find, which

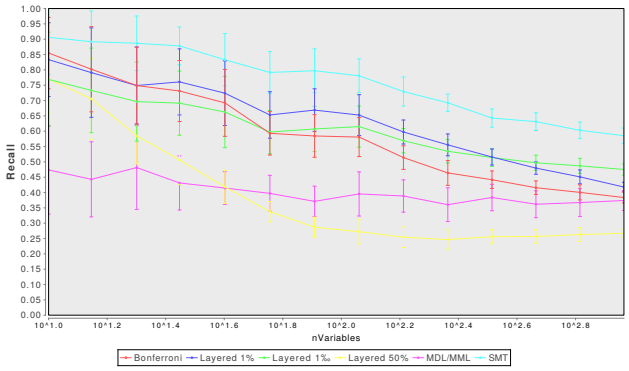

Figure 7: Study of the variation on the number of variables. Note that the x-axis is in log-scale.

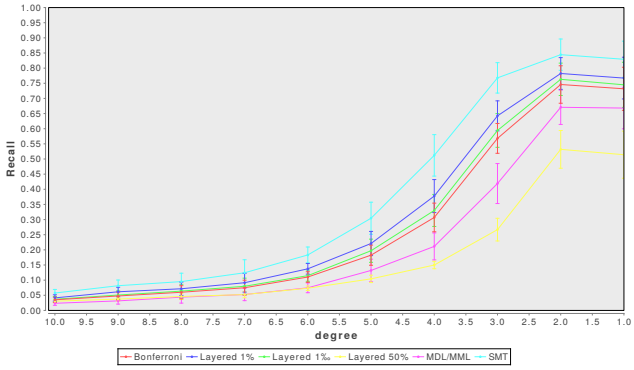

Figure 8: Study of the variation of the density; the x-axis corresponds to the degree in the Barabási–Albert (BA) graph model. High value means more complex graph.

calls for a more 'spread out' use of the budget. Here again, SMT not only outperforms all approaches, but does so without having any parameter to set.

4.3.4 *Varying the density of edges in the graph*

We vary D from 10 to 1 in steps of 1. Note that D controls for the size of the largest cliques in the graph, which means that the complexity of finding edges grows exponentially with D. Results are depicted in Figure 8. Again, SMT uniformly dominates all other methods; observations are similar to the experiment varying the number of instances.

4.3.5 *Varying the subtlety of the correlations*

We vary S from 50 to 500 in steps of 50; recall that S corresponds to the concentration parameter of a Dirichlet prior put on the multinomial probabilities of the parameterized model. Intuitively, low values of S lead to peaked multinomial distributions (thus easier to find), while high values lead to multinomials that are close to uniform (thus difficult to find). Results are depicted in Figure 9. SMT uniformly dominates other methods. Observations are similar to the experiment varying the number of instances with the slight difference for $S \geqslant 450$ where Layered-50% slightly outperforms MML/MDL approaches. This is because the absolute value of the recall is very low for high S. We posit that it might then favor statistical approaches that are able to

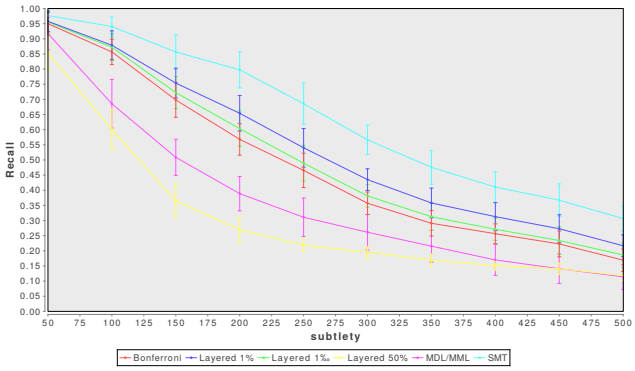

Figure 9: Study of the variation on the strength of correlation. Low values in the x-axis represent strong correlations (thus easier to find).

"spend most of the risk budget," even though one single edge might not be extremely significant. In contrast, MML/MDL approaches decide upon acceptance without looking at the risk that has been spent at previous steps: a model is accepted if it is more probable without any consideration of the multiple testing problem.

4.3.6 *General comment*

It is clear from these experiments that SMT significantly outperforms state-of-the-art approaches for explanatory model selection (also known as log-linear analysis). We have also drawn a few interesting observations as per the statistical efficiency of other methods. We showed that the setting of the layering parameter makes the layered search method range from very poor performance to second-best. Such parametrization is however problematic because it equates to a prior on the distribution of the p-values for the hypotheses that will be tested. This shows a second advantage to SMT, which outperforms all other methods, without requiring tuning of any parameter.

4.4 Real-world data

We use a broad range of real-world datasets, with both various number of variables and various quantities of data:

Mushroom the classical mushroom dataset, 22 variables, 8k examples [21].

EPESE epidemiological study of the elderly, 25 variables, 14k examples [33].

Internet demographic information on internet users, 70 variables, 10k examples [13].

CoIL2000 insurance customer management, 86 variables, 6k examples [35].

MITFace face recognition dataset, discretized to 4 bins using equal frequency, 362 variables, 31k examples [23].

Finance stock performance of the companies listed in the S&P500 over 20 years of trading, 500 variables.

Protein Multiple alignment of the Serpin family of proteins, 750 variables, 212 proteins [16].

Table 2: Number of edges found in the resulting models for real-world datasets.

Name	Number of interactions found			
	MDL	Bonferroni	Best Layered	SMT
Mushroom	21	76	78	**79**
EPESE	26	50	50	**53**
Internet	137	219	230	**247**
CoIL2000	67	168	169	**173**
MITFace	722	1,449	1,456	**1,487**
Finance	864	1,320	1,465	**1,640**
Protein	4	321	399	**471**
Orphamine	**650**	283	394	506
ABC	1,408	1,842	2,102	**2,426**
NYT	9,352	15,778	15,429	**17,741**

Orphamine Frequency of occurrence of 1,260 symptoms for 2,600 rare diseases (1,260 variables and 2,600 examples) [24].

ABC Use of the 500 most interesting (as per $tf \star idf$) words in all the news articles about Melbourne published by the Australian Broadcasting Network (ABC), 500 variables, 35k examples.

NYT Use of the 2,000 most interesting words in 10% of the articles published by the New York Times from 1987 to 2007, 2,000 variables, 180k examples [29].

Where licensing restrictions permit us to do so, we have made these datasets available at http://bit.ly/RealWorldResults.

We report in Table 2 the number of interactions found by the Chordalysis framework using the SMT, layered critical values and MML/MDL frameworks. Note that for clarity of the presentation, we report the best result obtained by the three parameterizations of the layered framework; note that this is significantly 'helping' it.

The results show that SMT significantly outperforms state-of-the-art approaches that control for FWER.

The only possibly surprising result is for the *Orphamine* dataset where the MDL/MML approach retrieves more edges than SMT. It is here important to remind the reader that we included MML/MDL approaches for reference only. MDL/MML approaches neither explicitly control for FWER, nor offer guarantees about it, as opposed to layered critical values and SMT. We thus posit that for the *Orphamine* dataset, the distribution of the successive tested hypotheses lead to MDL/MML having a high FWER. *Orphamine* is indeed a particular dataset with high number of variables and few examples; this translates into a high number of weak associations to find, because the number of samples is too low to have strong indications of the interactions. This has different consequences for statistical and MDL/MML frameworks:

- For SMT, it means that a significant part of the 'risk budget' is consumed for each edge that is found. Quickly after 500 edges, all of the budget has been consumed.

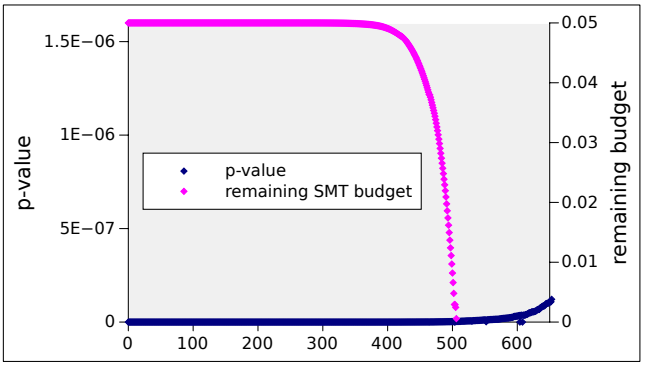

Figure 10: Study of SMT's use of the statistical significance 'budget' compared to p-values of the interactions up to 650 edges.

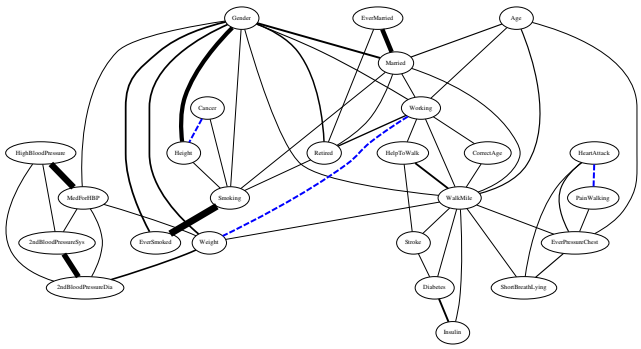

Figure 11: Compared graphs of correlations extracted from the EPESE dataset by the all statistical methods and SMT. Edges found by both frameworks are depicted in black. The three edges found by SMT but not using layered critical values are depicted in dashed blue lines. Edge weights correspond to the decrease in entropy associated with their addition — a proxy for strength of correlation.

- Conversely for MDL/MML, the decision of accepting an edge does not take into account any of the previous decisions. This means that they do not control for the fact that many weak correlations might have already been accepted in the search, resulting in a potential high FWER.

This interpretation is confirmed by Figure 10, where we plot the p-value of the first 650 steps of the process (without stopping criterion on FWER). We can see that the p-values of edges accepted at step 506 (when SMT stops) are not substantially different to the ones that would be accepted at step 650 (when MML/MDL stops) and increase slowly. SMT doesn't stop because of the significance of the current edges, but only because of the previous actions that have consumed the budget. Because MML/MDL decides upon acceptance without looking at what has been done in the previous steps, it naturally follows that it would be able to keep finding edges.

We make all the graphs obtained for all methods and all datasets at http://bit.ly/RealWorldResults. We give an example of results on the medical dataset EPESE in Figure 11.

EPESE corresponds to an epidemiological study of the elderly conducted by [33] in the U.S. between 1981 and 1987,

and named *Established Populations for Epidemiologic Studies of the Elderly* (EPESE). The study goals were to describe and to identify predictors of mortality, hospitalization, and placement in long-term care facilities and to investigate risk factors for chronic diseases and loss of functioning. The survey elicited information from persons 65 years of age and older in four geographic locations in the U.S. The publicly available baseline data covers demographic characteristics (age, sex, race, income, education, marital status, number of children, employment, and religion), height, weight, social and physical functioning, chronic conditions, related heath problems, health habits, self-reported use of dental, hospital and nursing home services, and depression.

Many of the multi-way relationships that have been retrieved have supporting evidence. For example, the obvious interaction between being married and having been married is included in the selected model. Many high-order interactions including age and gender and a third variable have been identified by Chordalysis. This is for example the case for the `Married` variable. The corresponding interpretation is simple: it is more likely to be married for an old patient than a young one, and it is well-known that women get married earlier than men. More generally, if the patient is older, then there has been a longer period over which they have had the opportunity to have smoked, been married or have retired (and hence be not working). Moreover, many medical conditions or social behaviors depend upon the gender. It is thus consistent that our approach identified many high-order interactions including age and gender.

Most identified interactions also have a direct medical interpretation. This is the case for the relationship between diabetes and taking insulin, between smoking (or having smoked) and having had cancer, between ability to walk and having had a stroke, between diastolic and systolic blood pressure, between having high blood pressure and taking medications for it, or between having had a heart attack and experiencing shortness of breath.

More interestingly, we can note the direct correlation between smoking and being married, which finds some supporting evidence in [7, 15]; and the one between diabetes and stroke, which is now suggested by several medical studies (see http://bit.ly/DiabetesStroke), because untreated diabetes tends to narrow blood vessels.

Finally, SMT finds 3 more edges than any other method:

- correlation between weight and working; this is supported by recent studies suggesting a tendency to gain weight for women when retiring [10].
- correlation between heart attack and having pain walking; Manesh Patel, MD. says that 'some forms of leg pain can be the first sign of heart disease [...] when leg pain occurs each time you engage in exercise or movement, and it stops soon after you stop, it could be a sign of peripheral arterial disease' (see http://bit.ly/LegPainHeartAttack).
- height and cancer; although this might seem like a very unexpected correlation, a recent large-scale study in the Lancet journal of Oncology found that, for women, 'every 10 cm increase in height [...] risk increased for 15 of the 17 cancers studied' [12].

5. CONCLUSIONS

We have introduced the first multiple testing correction for streams and cascades of statistical hypothesis tests. Monte Carlo simulations demonstrate the statistical power of the approach. A case study of its application in model selection demonstrates its practical utility.

This work opens up multiple avenues of research. Is it possible to improve the power of the technique? Can it be extended to rejecting multiple null hypotheses for each subfamily? Is it possible to allow the decision of whether to reject a null hypothesis to be revisited in the light of null hypotheses and evidence encountered when assessing subsequent subfamilies?

It would also be interesting to study how to guarantee the FWER with MML/MDL approaches. Our experiments have repeatedly shown that such approaches lead to a significantly reduced statistical efficiency. We believe that this could be due to the fact that information theoretic methods implicitly set the significance threshold very low, and hence do not use the whole "risk budget" that one might be willing to spend. We believe that exploring how such a tolerance can be integrated in Bayesian discovery processes constitutes a very promising avenue of research.

Finally, our new type of multiple test correction opens the possibility of new scientific experimental designs, where multiple hypotheses are identified for assessment, that are then assessed in order, with the exact hypotheses to be assessed being determined by which are rejected as the process unfolds. For example, a social science experiment might first test the hypotheses that either a) a particular group suffers from a specific disadvantage, or b) that the group enjoys a specific advantage. If one of the associated null hypothesis is rejected, then the researchers would assess alternative sets of hypotheses that may reveal reasons for the advantage or disadvantage that has been revealed in the first set of tests. The results of these tests might open up further sets of more detailed hypotheses to be explored in turn.

6. ACKNOWLEDGMENTS

This research has been supported by the Australian Research Council under grant DP140100087. We are grateful to Bart Goethals for helpful comments and suggestions.

7. REFERENCES

[1] S. Altmueller and R. Haralick. Approximating high dimensional probability distributions. In *IEEE Int. Conf. on Pattern Recognition*, pages 299–302, 2004.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[3] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[4] R. Christensen. *Log-Linear Models and Logistic Regression Second Edition*. Springer, 1997.

[5] C. Dahinden, M. Kalisch, and P. Bühlmann. Decomposition and model selection for large contingency tables. *Biometrical Journal*, 52(2):233–252, 2010.

[6] L. M. De Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *The Journal of Machine Learning Research*, 7:2149–2187, 2006.

[7] E. W. Doherty and W. J. Doherty. Smoke gets in your eyes: Cigarette smoking and divorce in a national sample of American adults. *Families, Systems, & Health*, 16(4):393, 1998.

[8] N. Draper and H. Smith. *Applied regression analysis*. John Wiley, 1981.

[9] O. J. Dunn. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, 30(1):192–197, 1959.

[10] V. L. Forman-Hoffman, K. K. Richardson, J. W. Yankey, S. L. Hillis, R. B. Wallace, and F. D. Wolinsky. Retirement and weight changes among men and women in the health and retirement study. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 63(3):S146–S153, 2008.

[11] V. Gogate, W. Webb, and P. Domingos. Learning Efficient Markov Networks. In *Advances in Neural Information Processing Systems*, pages 748–756, 2010.

[12] J. Green, B. J. Cairns, D. Casabonne, F. L. Wright, G. Reeves, V. Beral, and Million Women Study collaborators. Height and cancer incidence in the million women study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk. *The Lancet Oncology*, 12(8):785–794, 2011.

[13] S. Hettich and S. Bay. UCI KDD archive, 1999.

[14] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[15] G. G. Homish and K. E. Leonard. Spousal influence on smoking behaviors in a us community sample of newly married couples. *Social science & medicine*, 61(12):2557–2567, 2005.

[16] J. Irving, R. Pike, A. Lesk, and J. Whisstock. Phylogeny of the serpin superfamily: Implications of patterns of amino acid conservation for structure and function. *Genome Research*, 10(12):1845–1864, 2000.

[17] V. E. Johnson. Revised standards for statistical evidence. *Proc. of the National Academy of Sciences of the USA*, 110(48):19313–19317, 2013.

[18] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

[19] R. Kohavi and R. Longbotham. Online controlled experiments and a/b testing. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*. Springer US, in-press.

[20] S.-I. Lee, V. Ganapathi, and D. Koller. Efficient Structure Learning of Markov Networks using ℓ_1-Regularization. In *Advances in Neural Information Processing Systems*, pages 817–824, 2006.

[21] M. Lichman. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2013.

[22] S. F. Meisel, D. A. Carere, J. Wardle, S. S. Kalia, T. A. Moreno, J. L. Mountain, J. S. Roberts, R. C. Green, P. S. Group, et al. Explaining, not just predicting, drives interest in personal genomics. *Genome Medicine*, 7(1):1–7, 2015.

[23] MIT Center For Biological and Computation Learning. CBCL Face Database #1. http://www.ai.mit.edu/projects/cbcl, 2000.

[24] Orphanet. An online database of rare diseases and orphan drugs. http://www.orpha.net, 2014.

[25] F. Petitjean, L. Allison, G. Webb, and A. Nicholson. A statistically efficient and scalable method for log-linear analysis of high-dimensional data. In *IEEE Int. Conf. on Data Mining*, pages 480–489, 2014.

[26] F. Petitjean and G. Webb. Scaling log-linear analysis to datasets with thousands of variables. In *SIAM Int. Conf. on Data Mining*, pages 469–477, 2015.

[27] F. Petitjean, G. Webb, and A. Nicholson. Scaling log-linear analysis to high-dimensional data. In *IEEE Int. Conf. on Data Mining*, pages 597–606, 2013.

[28] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[29] E. Sandhaus. The New York Times Corpus. https://catalog.ldc.upenn.edu/LDC2008T19, 2008.

[30] M. Schmidt, A. Niculescu-Mizil, K. Murphy, et al. Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pages 1278–1283, 2007.

[31] G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[32] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

[33] J. Taylor, R. Wallace, A. Ostfeld, and D. Blazer. Established Populations for Epidemiologic Studies of the Elderly, 1981-1993. http://dx.doi.org/10.3886/ICPSR09915, 1998.

[34] A. Terada, M. Okada-Hatakeyama, K. Tsuda, and J. Sese. Statistical significance of combinatorial regulations. *Proc. National Academy of Sciences*, 110(32):12996–13001, 2013.

[35] P. van der Putten and M. van Someren. A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000. *Machine Learning*, 57(1–2):177–195, 2004.

[36] M. Wainwright, P. Ravikumar, and J. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems*, pages 1465–1472, 2007.

[37] C. S. Wallace and D. M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

[38] G. I. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007.

[39] G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2–3):307–323, 2008.

[40] X. Wu, D. Barbará, and Y. Ye. Screening and interpreting multi-item associations based on log-linear modeling. In *Int. Conf. on Knowledge Discovery and Data Mining*, pages 276–285, 2003.