Maximum Entropy Modelling for Assessing Results on Real-Valued Data

Kleanthis-Nikolaos Kontonasios
University of Bristol,
Intelligent Systems Laboratory
kk8232@bristol.ac.uk

Jilles Vreeken
University of Antwerp,
Advanced Database Research and Modelling
Jilles.Vreeken@ua.ac.be.com

Tijl De Bie
University of Bristol,
Intelligent Systems Laboratory
tijl.debie@gmail.com

***Abstract*—Statistical assessment of the results of data mining is increasingly recognised as a core task in the knowledge discovery process. It is of key importance in practice, as results that might seem interesting at first glance can often be explained by well-known basic properties of the data. In pattern mining, for instance, such trivial results can be so overwhelming in number that filtering them out is a necessity in order to identify the truly interesting patterns.**

In this paper, we propose an approach for assessing results on real-valued rectangular databases. More specifically, using our analytical model we are able to statistically assess whether or not a discovered structure may be the trivial result of the row and column marginal distributions in the database.

Our main approach is to use the Maximum Entropy principle to fit a background model to the data while respecting its marginal distributions. To find these distributions, we employ an MDL based histogram estimator, and we fit these in our model using efficient convex optimization techniques. Subsequently, our model can be used to calculate probabilities directly, as well as to efficiently sample data with the purpose of assessing results by means of empirical hypothesis testing. Notably, our approach is efficient, parameter-free, and naturally deals with missing values. As such, it represents a well-founded alternative to swap randomisation.

I. Introduction

Clearly, any data analyst is only interested in results that identify real structure in the data, and does not want to invest time analyzing complex results that can be trivially explained by well-known basic properties of the data. Hence, arguably one of the most important sub-tasks of data mining is assessing whether a discovered result is significant or not. However, as important as such assessment is, it has received relatively little attention by the data mining community.

From traditional statistics we know that randomization is a powerful approach for significance testing. The main idea is that a result is only interesting if it is highly unlikely to be obtained from random data of the same basic properties as the original (i.e. when a result has a low p-value). Randomization is a very general technique, of which the effectiveness mainly depends on how efficiently we can sample data, and which constraints we can maintain.

Recently, Gionis et al. [5] introduced swap randomization for assessing data mining results on binary data, where they give an algorithm to randomize data while keeping the row and column margins intact. A generalization for real-valued data proposed by Ojala et al. [13], [12], preserving row and column margins approximately, made the approach more generally applicable.

Essentially, swap randomization is a very simple process for sampling randomized datasets: starting with the original data, we randomly find a submatrix of certain properties, *swap* the values of the cells, and repeat this until the data is randomized. Although straightforward, it is computationally taxing. Randomly finding a swap that maintains the background information is not trivial, and, as very many such swaps are required to randomize the data, in practice the number of randomized datasets we can sample is rather limited—and consequently, so is the resolution of the empirical p-values. With low resolution, however, we cannot properly rank results, nor can we reliably identify the most significant result, as often many results will be measured to be equally significant. Moreover, as no theoretic results are available for the required number of swaps, heuristics are used, leaving the probability that the assessment is biased.

Alternatively, instead of perturbing the original data, we can construct an analytical model of the data by the Maximum Entropy principle [8], which in turn we can use to sample randomized data quickly, as well as to calculate exact probabilities and p-values. Recently, De Bie [3], showed that modelling by the Maximum Entropy principle provides a well-founded and fast alternative to swap randomization when considering column and row sums as background knowledge. However, to date, the flexibility of maximum entropy modelling in terms of data types, as well as in terms of types of constraints, is not as general as for swap randomization. This paper is meant to close this gap by introducing new maximum entropy models for general real-valued data, and subject to constraints that go much further than simple row and column sums.

As such, we propose a fast and well-founded approach for assessing the statistical significance of data mining results on real-valued data, providing a number of advantages over swap randomization. Our approach employs the Maximum Entropy principle to fit a background model to

the data that respects its marginal distributions. To find these distributions, we employ a parameter-free histogram estimator [10] based on the Minimum Description Length principle [6], and we fit these in our model using efficient convex optimization techniques. Alternatively, our approach can also consider the mean and variance over the rows and columns as constraints, as even more straightforwardly interpreted background knowledge.

Assessing data mining results with respect to our model allows us to identify as uninteresting any structure that is the (trivial) result of individual row or column distributions, such that only results due to dependencies in the data remain. Notably, the resulting model can both be used for efficiently sampling data, as well as for direct analytical assessment, and is naturally able to deal with missing values.

Experiments show our approach is an order of magnitude faster than the state of the art, while it provides at the least equally strong assessment. We show it is generally applicable for empirical p-value testing, as well as for calculating exact probabilities and p-values. As an example of the latter, we discuss how to analytically calculate p-values for the weighted relative accuracy in order to identify significant subgroups. As such, we show our approach represents an efficient and well-founded alternative that goes further than what is currently achievable with swap randomization, while doing this in a rigorous manner without convergence issues.

The roadmap of this paper is as follows. First, we discuss related work on statistical assessment in more detail. Then, in Section III we cover the main theory of our method, which we empirically evaluate in Section IV. We round up with discussion and conclusions in Sections V and VI.

II. Related Work

The assessment of data mining results was pioneered in the work of Gionis et al. on swap randomization [5], which were initially developed for binary databases and row and column marginal constraints. In this context, a swap is a local randomization operation on a binary database that leaves the row and column sums unaltered. A Markov chain of randomly chosen swaps thus randomizes the data completely, and in principle allows one to uniformly sample from the set of all databases that satisfy the prior information. After sampling a sufficiently large number (typically at least 1000) of such randomized databases, empirical hypothesis testing can be used to compute empirical p-values for patterns found in the database.

Soon after its introduction, the swap randomization approach was extended to real-valued data [13]. Furthermore, alternative swapping schemes were developed for types of background information different from just row and column marginals, such as itemset frequencies, and clusters in the data. This allowed the use of swap randomization in iterative data mining settings [7], where newly discovered patterns are added to the background information in an iterative manner. These developments have made the swap randomization approach a comprehensive one for the assessment of data mining results in general, and pattern mining in particular, in a wide range of contexts.

Swap randomization does suffer from a number of problems, though. The convergence of a Markov chain of random swaps to its stationary distribution, which is required for the method to work well, is hard to study theoretically. Thus, heuristic rules need to be used to determine when to stop swapping. This number of required swaps is typically so large that sampling one database requires a significant computational effort. As a result, the number of randomized databases that can practically be generated is limited, and therefore also the resolution of the empirical p-values is poor. These disadvantages essentially stem from the fact that the randomized databases are defined in a procedural manner, rather than in an analytical manner.

An approach that remedies these problems explicitly models the background information into an analytically specified background distribution, obtained as the maximum entropy distribution subject to constraints that are implied by the background information. This methodology has been demonstrated for binary, positive integer-valued, and positive real-valued data, with constraints on the expected row and column sums [3]. In each of these cases the resulting background distribution is an exponential family model that can be fitted remarkably efficiently for database of potentially very large sizes. Sampling from the maximum entropy distribution has also been shown to be highly efficient. Still, the benefits of maximum entropy background distributions as compared to swap randomization were perhaps demonstrated most clearly by the development of analytically specified interestingness measures [3], [11], which seems impossible to do with swap randomization, and makes repeated sampling unnecessary altogether.

III. Theory

In this section we formally introduce our approach. We start by giving a gentle introduction to the Maximum Entropy principle, after which we present our approach to modelling a database of real-valued attributes. We give two different ways of incorporating background knowledge on the marginal distributions into the model.

A. Notation

We are concerned with m-by-n real-valued databases D. $D_{i,j}$ denotes the (i, j) element of database D. The set $I = \{1, 2, ..., m\}$ denotes the set of rows and $J = \{1, 2, ..., n\}$, the set of columns.

B. The Maximum Entropy Principle

Our approach to assessing patterns is based on modelling basic background information about the data into a background model, and subsequently contrasting results with

this model, e.g. by means of (empirical) hypothesis testing. Typically, several probabilistic models exist that conform to the background information, so that a criterion is needed to decide which one to use. In order to avoid introducing any undue biases in selecting this model, it has been argued that the distribution of maximum entropy [8] is the most appropriate one to use [3], [11].

Simply put, by modelling by entropy maximization, we ensure that the probabilistic model we obtain uses the provided background information optimally, but is completely unbiased any further. As such, these models are superior to any other probabilistic model.

Entropy maximization problems are extremely well-suited to deal with background information on expected values of certain properties of the data. For general data D, such a problem is of the form:

$$\max_P \quad -\sum_D P(D)\log\left(P(D)\right) \ , \tag{1}$$

$$\text{s.t.} \qquad \sum_D P(D) f_s(D) = d_s, \quad \forall s \ , \tag{2}$$

$$\sum_D P(D) = 1 \ . \tag{3}$$

where the functions f_s compute the properties of which the expected values are known to be equal to d_s as part of the background information. Entropy maximization subject to such constraints is a convex problem, which can often be solved efficiently. Furthermore, the resulting distributions are known to belong to the exponential family of distributions, the properties of which are very well understood [17]. In particular, the maximum entropy distribution is of the form:

$$P(D) \quad = \quad \frac{1}{Z}\exp\{\sum_s \lambda_s f_s(D)\} \ , \tag{4}$$

where Z is a normalization factor known as the partition function and λ_s are the Lagrange multipliers for the constraints of problem 1-3, to be optimized by solving the dual optimization problem.

C. Modelling a Real-Valued Data Matrix

Maximum entropy modelling of *positive* real-valued data matrices has been discussed before by De Bie [3]. There the maximum entropy model was shown to reduce to a product distribution of exponential distributions defined for each matrix element, if constraints on the expected row and column means are used. Here, we significantly broaden the scope to models for general real-valued data. Furthermore, we consider different and richer types of background information, instead of just the means of each row and column.

In particular, our goal is to maintain the entire distribution of the elements in each row and in each column, rather than just the means. Doing this would be useful in those cases where patterns that arise due to interactions between rows and columns are of interest, but patterns that trivially result from the distribution of values in individual rows and columns are to be discarded.

Obviously, maintaining the entire distribution of values in a given row or column requires us to first infer it from the data. Below we consider two ways of doing that. The first approach assumes that a distribution is specified sufficiently by fixing its mean and its variance only.[1] The second approach is more principled, and specifies the distributions by means of histograms of optimal resolution, as determined using the Minimum Description Length (MDL) approach of Kontkanen and Myllymäki [10].

Remark. A basic assumption in our framework is that the summations of the elements or the histogram distributions along the rows and columns in the database have a meaningful and intuitive interpretation. If columns are considered as features, this is usually the case since the column sums can encapsulate information for the mean of the feature value. However summing over values of different features makes sense only when the features are similar in nature. Gene expression data is a prime example of such data.

To summarize, we formulate and solve two different modelling problems defined over the prior knowledge presented above. For the first, we assume the data analyst has easy access to the means and variances of the rows and columns.

Problem 1: Means and Variances. Given a real-valued database D, construct a probabilistic model P for D that preserves in expectation the means and the variances of the elements in each row and column in the database D.

By this problem, we construct probabilistic models that assign higher probabilities, and hence higher p-values, to results that follow from the means and variances of the rows and columns. These basic statistics, however, are possibly not the most informative easily accessible background information. Instead, we can consider histograms per row and column, which leads us to the following problem.

Problem 2: Histograms. Given a real-valued database D and a set of histograms derived from each row and column in D, construct a probabilistic model P for D that preserves in expectation the histogram information.

In the remainder of this section we specify the functions f_s and the constants d_s for these respective problems and present the shape of the resulting distributions. Details about the derivation of these solutions are omitted due to space restrictions. Generally speaking, the approach for solving both problems is the same and is summarized as follows.

First, the Karush-Kuhn-Tucker (KKT) stationarity conditions can be used to derive the mathematical form of the solution [2]. In both cases, this yields an exponential family type distribution [17] parameterized by Lagrange multipliers, one for each of the constraints. Subsequently, the value

[1]Note that this method could be extended by also fixing higher order moments of the distribution.

of the Lagrange multipliers can be found by solving the unconstrained dual optimization problem. This can be done using established and provably efficient techniques from convex optimization [2]. We discuss our choices for the optimizers, and the resulting complexities in Section III-F.

D. Using Mean and Variance

For specifying the general formulation of Eq. 2 to the context of modelling means and variances, the following equations are used as functions f_s.

$$f_i(D) = \sum_j D_{i,j} \quad \forall i \in I \ , \tag{5}$$

$$f_{m+j}(D) = \sum_i D_{i,j} \quad \forall j \in J \ , \tag{6}$$

$$f_{m+n+i}(D) = \sum_j D_{i,j}^2 \quad \forall i \in I \ , \tag{7}$$

$$f_{2\cdot m+n+j}(D) = \sum_i D_{i,j}^2 \quad \forall j \in J \ . \tag{8}$$

These functions compute the row/column sums (Eqs. 5, 6) and the row/column square sums (Eqs. 7, 8) for a database D of dimensions m-by-n. Preserving sums and square sums is equivalent to preserving means and variances. The total number of constraints used is $2mn$.

The values of d_s in Eq. 2 are calculated using Eqs. 5–8 in the original database, where missing values can simply and safely be ignored—they do not provide any bias to the maximum entropy distribution.

Shape of the solution. Using the Lagrange multipliers theory and the KKT conditions it can be derived that the probability distribution of the database factorizes as a product of independent distributions, one for each database entry. More formally, it can be shown that

$$P(D) = \prod_{i,j} P_{ij}(D_{i,j}) \ , \tag{9}$$

and the probability density distribution for the database entry at row i and column j is a normal distribution of the form:

$$P_{ij}(D_{i,j}) = \sqrt{\frac{\mu_i^r + \mu_j^c}{\pi}} \exp\left\{ \frac{-\left[D_{i,j} + \frac{1}{2}\frac{\lambda_i^r + \lambda_j^c}{\mu_i^r + \mu_j^c}\right]^2}{\frac{1}{\mu_i^r + \mu_j^c}} \right\} \ ,$$

where $\lambda_i^r / \lambda_j^c$ are Lagrange multipliers defined for the constraints on the sums of row i, and column j, correspondingly. Analogously, μ are Lagrange multipliers for constraints over the sums of squares.

In other words the MaxEnt distribution is product of mn normal distributions with mean equal to

$$-\frac{1}{2} \times \frac{\lambda_i^r + \lambda_j^c}{\mu_i^r + \mu_j^c}$$

and variance equal to

$$\frac{1}{\sqrt{2(\mu_i^r + \mu_j^c)}} \ .$$

It is easy to see that the probability distribution for each database entry $D_{i,j}$ depends only on the Lagrange multipliers of the row i and column j that the entry belongs to.

E. Using Histogram Bin Information

Although easily accessible and interpretable, means and variances of rows, or columns, only provide limited insight regarding the distributions. Histograms, however, while almost as easily interpretable, provide much more detailed views. Here we formulate Eq. 2 such that the resulting model respects the histograms of the rows and columns.

Our formulation is general, i.e., the binning results of any histogram estimation technique can be used. However, many estimation techniques require the user to choose some parameters beforehand, like the number or width of bins. Typically, we do not know these values, and consequently may obtain overly simple or complex histograms. To this end, we propose to employ a well-founded and parameter-free, variable-bin-width histogram estimation technique [10], based on the Minimum Description Length principle [6].

The basic idea behind it is rather simple, namely, that the best histogram is the histogram that gives the optimal lossless compression. This means that the resulting histogram will not be overly complex, nor overly simple, for then another histogram exists that can compress the data better. Kontkanen and Myllymäki [10] use the normalized maximum likelihood distribution, which has several theoretic optimality properties, to model the values within a bin, and give an efficient dynamic programming algorithm for discovering the MDL optimal histogram automatically. For more detail, we refer to the original publication [10].

Preserving the row and column histograms in a probabilistic model is essentially equivalent to preserving their expected bin content. This requirement can be formalized using indicator functions as follows,

$$f_{s_{r_i,bin_{r_i}}}(D) = \sum_j I^{r_i}_{bin_{r_i}}(D_{i,j}) \quad \forall i \in I \ , \ \forall bin_{r_i} \ , \tag{10}$$

$$f_{s_{c_j,bin_{c_j}}}(D) = \sum_i I^{c_j}_{bin_{c_j}}(D_{i,j}) \quad \forall j \in J \ , \ \forall bin_{c_j} \ , \tag{11}$$

where $bin_{r_i} \in \{1, 2, ...,$ number of bins in r_i histogram$\}$ is an index to the bin of the histogram of row i. Correspondingly bin_{c_j} refers to the bins of column j's histogram. The function $I^{r_i}_{bin_{r_i}}(D_{i,j})$ indicates the inclusion/exclusion of element $D_{i,j}$ from the 'bin_{r_i}'th bin of the histogram of row i. Analogously $I^{c_i}_{bin_{c_j}}$ is defined for bins in the column histograms. Clearly the number of constraints for *Problem 2* is equal to the number of the bins in all histograms.

The values of d_s in Eq. 2 are calculated using Eqs. 10, 11 in the original database. Like for Means-Variance modelling,

when computing the constraint values, we can compute them simply based on the non-missing values—as so the missing values will not give any bias to the resulting distribution.

Shape of the solution. Solving the MaxEnt problem under these settings produces again a probability distribution which decomposes into a product of independent components for each database entry, as described by Eq. 9.

The probability distribution for each database entry is

$$P_{ij}(D_{i,j}) = \frac{1}{Z_{i,j}} \exp\Big\{ \sum_{bin_{r_i}} \lambda^{r_i}_{bin_{r_i}} I^{r_i}_{bin_{r_i}}(D_{i,j}) + \sum_{bin_{c_j}} \lambda^{c_j}_{bin_{c_j}} I^{c_j}_{bin_{c_j}}(D_{i,j}) \Big\} , \quad (12)$$

where $Z_{i,j}$ is a normalization factor. This distribution P_{ij} is piece-wise contant between the cut-points of the histograms for row i and column j; i.e. it is a histogram itself.

F. Solving the models

Given the above formalizations, in order to obtain the maximum entropy models, we arbitrarily chose two convex optimization techniques for the two different problems. For each of these techniques, the complexity directly depends on the number of Lagrange multipliers $\#\lambda$.

For Problem 1 we made use of Newton's method with the Armijo rule to determine the step size. At the core of Newton's method is the solution of a linear system of $\#\lambda$ equations which has computational complexity $O(\#\lambda^3)$ and space complexity $O(\#\lambda^2)$. However, the number of steps required by Newton's algorithm is logarithmic in the required accuracy and in practice always very small (only ten to twenty in our experiments).

For Problem 2 we used the conjugate gradient method, which requires only gradient computations and vector-vector multiplications that require $O(\#\lambda)$ time and space. The gradient itself can be computed in $O(\#\lambda^2)$ operations. Although the number of iterations is typically much larger for conjugate gradient, and no strong results bounding the number of required iterations in terms of accuracy are available, we will see in the experiments below that it turns out to be a better trade-off than Newton's method.

Note that many other techniques for convex optimization exist, and it is likely that the techniques we used in this paper can be significantly improved upon [2]. Studying this in detail, however, is clearly beyond the scope of this paper.

G. Usages of explicit probabilistic models

Explicit probabilistic models are very flexible in terms of their usage in assessing data mining results.

Sampling Data: First of all, maximum entropy models are very efficient for sampling complete randomized databases. The relatively expensive step to attain the maximum entropy model of a dataset needs only to be taken once. Given that model, sampling randomized databases is very cheap, and essentially comes down to drawing mn random numbers from a normal distribution, or histogram.

Table I
CHARACTERISTICS OF THE DATASETS USED IN THIS PAPER. GIVEN ARE THE NUMBER OF ROWS, NUMBER OF COLUMNS, AND THE AVERAGE NUMBER OF BINS IN ROW AND COLUMN HISTOGRAMS.

Dataset	*# of rows*	*# of columns*	*Avg. # of row bins*	*Avg. # of column bins*
Random	100	100	1.97	1.97
Gaussian	1 000	100	3.01	8.60
Component	1 000	50	5.25	5.91
Cluster	1 117	100	5.03	7.58
Alon	2 000	62	4.74	10.80
Gene	1 375	60	4.12	10.25
Thalia	734	69	2.81	8.25

Exact Probabilities: Second, explicit models can be used for calculating analytically *exact* probabilities and *exact* p-values. The complexity of calculating these strongly depends on the complexity of the structure we are interested in. While calculating the expected support of a simple pattern comes down to a straightforward linear combination, it is not immediately apparent how to, for example, calculate the probability that a particular cluster will be discovered.

Despite the fact that analytically calculating p-values for complex structures is difficult in principle, it can be highly useful in settings where empirical hypothesis testing is expensive. In the context of this paper, as an example of such usage, we use our models to calculate exact p-values for assessing subgroup discovery results. In particular, we evaluate the statistical significance of the Weighted Relative Accuracy (WRAcc) [16] of a rule.

Interestingness: Third, maximum entropy models can be used to define Information-Theoretic interestingness measures of structure using means other than hypothesis testing. For example, De Bie [3] showed how maximum entropy models can be used to assess the interestingness of tiles in binary data. Extending these ideas in the context of real-valued data, with the goal of identifying the most interesting numeric patterns, is a realistic goal for future work.

IV. EXPERIMENTS

In this section we empirically evaluate our methods. We first provide a toy example of our modelling approaches in order to give an intuition about the obtained results. Next we examine execution times for the modelling and sampling procedures in artificial and real datasets. Execution times are compared with the methods presented in [12], [13]. The statistical significance of global structural measures in the data, such as the k-means clustering error, the fraction of variance explained by the first five principal components and the maximum correlation between columns, using empirical hypothesis testing follows. Finally, we show how our explicit models can be used for calculating exact p-values for assessing subgroup discovery results.

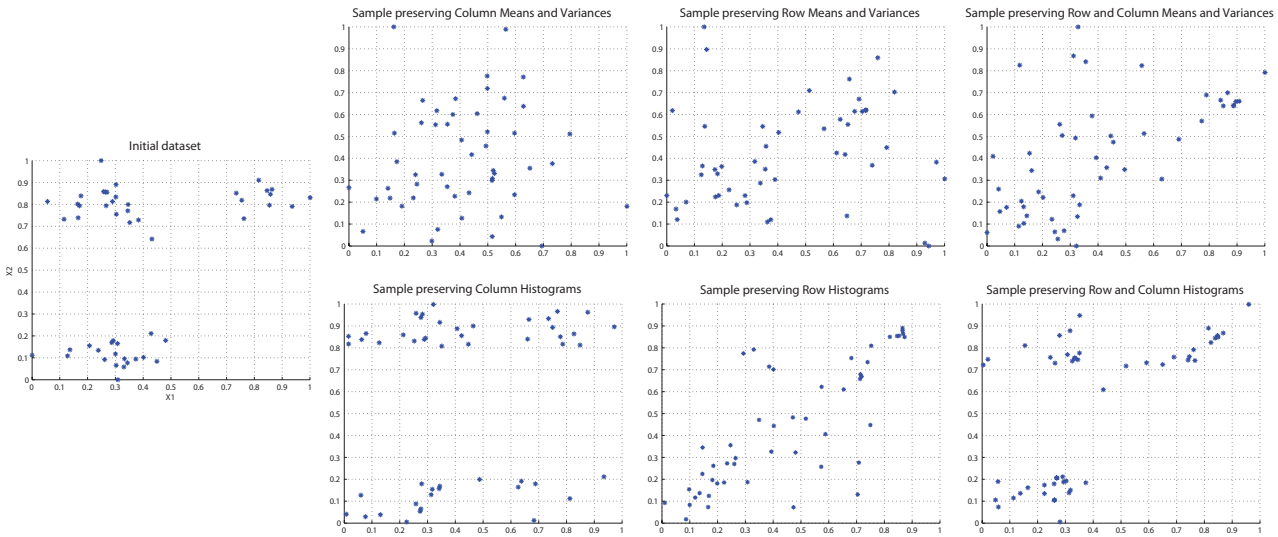

Figure 1. Means-Variances (top row) and Histogram modelling (bottom row) for a 50×2 artificial dataset (far left) using Column information (col. 2), Row information (col. 3), and both Row and Column information (col. 4) as constraints.

A. Setup

We implemented our methods in C++, and make our code available for research purposes.[2] All experiments were performed on a 2.66GHz processor/2.87GB RAM machine.

We tested our methods on 4 synthetic and 3 real datasets. We employ the artificial *Random*, *Gaussian*, *Component* and *Cluster* datasets as used by Ojala et al. in [13] and [12], to analyze performance on data with known structure.

To see how our methods fare on real data, we take three publically available gene expression datasets, resp. *Gene* [15], *Thalia* [14] and *Alon* [1]. For each of these datasets, we use a matrix representation such that rows correspond to genes and columns to conditions. The basic characteristics of the datasets are depicted in Table I. For the Histograms method, we obtained the MDL optimal histograms per row and per column using the implementation of Kontkanen and Myllymäki [10], and the average number of bins per row and per column are reported in Table I.

B. An illustrative example

In our first experiment we create and sample from models that successively preserve only column information, only row information and finally both row and column information. Our aim here is to give a basic intuition about the evolution of the samples generated by our models using different, increasingly involved, prior knowledge under our proposed modelling schemes.

The dataset used is an artificial dataset of 50 two-dimensional points containing three clusters. Two clusters contain 20 points each and a third one 10 points. Points within each cluster are drawn randomly from normal distributions. Figure 1 depicts the initial dataset (leftmost plot), as well as datasets sampled for every combination of modelling scheme (rows) and prior information (columns).

The second column of Figure 1 presents dataset samples drawn from the maximum entropy models preserving only column information. Beginning with Mean-Variance modelling on the top row, we observe that the sampled points using only column information resemble a single normal distribution. Indeed, it is easy to prove theoretically that the maximum entropy distribution in this case is a normal distribution with mean equal to the overall mean, and a diagonal covariance matrix with diagonal elements equal to the variances along the two individual dimensions [8]. Clearly, the true cluster structure of the original data is lost.

For the case of modelling Histograms, on the bottom row, two histogram bins were used for each of both columns. Compared to Mean-Variance modelling, more structural information is retained. However, it is clear that retaining only column information still preserves relatively little structural information. Indeed, samples appear in the lower right part of the plot, where they were absent in the original data.

Samples generated from models that instead only preserve row information are depicted in the third column of Figure 1. For both modelling schemes, row modelling introduces some correlation between the columns of the dataset, which is most clear for Histogram modelling. This is a natural consequence of the imposed background knowledge on the row distributions. Indeed, database elements in different columns in the same row are then sampled from the same distribution, and from a different distribution than the elements in different rows. Thus, different database elements within the same row will generally be more similar.

When modelling row and column information simultaneously (rightmost column of Figure 1), most of the true cluster

[2]http://www.tijldebie.net/software/maxentRV

Table II
MEANS-VARIANCE MODELLING. NUMBER OF LAGRANGE MULTIPLIERS, NUMBER OF ITERATIONS, AND RUNNING TIME FOR MODELLING, AS WELL AS FOR SAMPLING 10 000 DATASETS.

	Model Statistics		**Time**	
Dataset	*# of λ's*	*# of it's*	*Model*	*10k Samples*
Random	400	12	7.9s	18.3s
Gaussian	2 020	13	1 451.8s	19.0s
Component	2 100	15	1 806.2s	82.0s
Cluster	2 434	14	2 419.9s	177.4s
Alon	4 124	20	17 881.5s	198.2s
Gene	2 872	15	4 670.5s	132.3s
Thalia	1 606	15	839.3s	82.9s

Table III
HISTOGRAM MODELLING. NUMBER OF LAGRANGE MULTIPLIERS, NUMBER OF ITERATIONS, AND RUNNING TIME FOR MODELLING, AS WELL AS FOR SAMPLING 10 000 RANDOMIZED DATASETS.

	Model Statistics		**Time**	
Dataset	*# of λ's*	*# of it's*	*Model*	*10k Samples*
Random	394	240	1.7s	9.9s
Gaussian	3 460	632	15.8s	14.2s
Component	5 547	881	107.4s	96.7s
Cluster	6 381	1 062	339.4s	217.0s
Alon	10 168	2 024	830.3s	341.2s
Gene	6 287	1 510	442.8s	196.0s
Thalia	2 638	1 294	162.3s	139.4s

structure is reconstructed. As expected, Histogram modelling approximates the original data most closely. This shows that the cluster structure in this database can be explained to a large extent by the row and column distributions only. Note that the correlation effect is now less prominent, apparently moderated by the column constraints.

C. Modelling and sampling times

As mentioned in Sec. III, for computing the solutions of the Mean-Variance and Histogram problems stated in Section III, we respectively used Newton's method and conjugate gradient descent [2]. Newton's method is more costly per iteration but requires significantly less iterations, while conjugate gradient needs more iterations that are however significantly cheaper. The number of Lagrange Multipliers (which to a large extent determine the computational complexity), the number of iterations, and the modelling and sampling times are shown in Tables II and III for the Mean-Variance and Histogram modelling methods respectively.

For both Mean-Variance and Histogram modelling, we see that the time solving the model greatly outweighs the sampling time. While for Mean-Variance we see that the modelling time can be significant, for Histograms modelling it only takes up to a few minutes. The difference is mostly due to the choice for Newton's method for the Mean-Variance modelling and conjugate gradient for Histogram modelling, leaving significant room to further speed up the Mean-Variance modelling. Sampling from the resulting MaxEnt models is very fast, requiring only up to milliseconds per randomized dataset.

Table IV
COMPARISON OF THE RUNNING TIMES FOR SAMPLING 10 000 DATASETS, FOR TWO SWAP-BASED RANDOMIZATION METHODS, AND OUR MAXIMUM ENTROPY MODELS (TIMINGS INCLUDE MODELLING).

	Swap-based methods		**MaxEnt methods**	
Dataset	*Swap-con.*	*Swap-dis.*	*Mean-Var.*	*Histogram*
Alon	33 281s	40 844s	18 080s	**1 171s**
Gene	13 422s	13 766s	4 802s	**638s**
Thalia	7 562s	9 953s	921s	**301s**

Next, we compare our running times to swap-randomization. Table IV shows running times for sampling 10 000 databases using two swap randomization methods, namely *SwapConstrained* and *SwapDiscretized*, and our two maximum entropy based variants. For swap randomization we used the publicly available implementation[3] by the authors of [13], [12]. For MaxEnt, the reported timings include both modelling and sampling.

The timings recorded in the table show that overall both our modelling approaches are considerably faster than the state of the art. At the worst case, using Means-Variances modelling with the *Alon* data, sampling 10 000 randomized datasets still takes only half the time required by the swap-based methods. The Histograms based method, while modelling the data in most detail, is consistently more than one order of magnitude faster, requiring only minutes instead of hours. Note additionally that sampling more databases is cheaper than using swap randomizations, as most effort goes into the modelling step which needs to be done only once.

D. Assessing global patterns in data

Next we use the models computed for assessing the statistical significance of patterns using empirical hypothesis testing. In particular we follow the experimental framework proposed in [13], [12] and we evaluate the significance of the values of three structural measures in a dataset, namely k-means clustering error, the fraction of variance explained by the five dominant principal components, and the maximum correlation between the columns of the dataset.

Formally, for a given dataset D the value of a structural measure $M(D)$ is calculated. Next, a set of sampled databases $\mathcal{S}$ is generated and the same measure is calculated on every database $S \in \mathcal{S}$. As measure of statistical significance we use *empirical* p-values. It is defined as the ratio of the number of sampled databases that have a structural measure value less or equal to the value calculated in the original database over the total number of sampled

[3] http://users.ics.tkk.fi/mrojala/randomization

Table V
P-VALUES ASSESSING THE SIGNIFICANCE OF 3 DIFFERENT GLOBAL STRUCTURAL MEASURES, BY MAXIMUM ENTROPY MODELLING USING THE MEANS-VARIANCES OF ROWS AND COLUMNS AS CONSTRAINTS.

	Means-Variances modelling		
Dataset	*k-Means clustering error*	*PCA*	*Maximum correlation*
Random	0.788	0.918	0.575
Gaussian	0.999	0.999	0.999
Component	**0.001**	**0.001**	**0.001**
Cluster	**0.001**	**0.001**	**0.001**
Alon	0.999	**0.001**	**0.001**
Gene	**0.001**	**0.001**	**0.001**
Thalia	**0.001**	**0.001**	**0.001**

Table VI
P-VALUES ASSESSING THE SIGNIFICANCE OF 3 DIFFERENT GLOBAL STRUCTURAL MEASURES, BY MAXIMUM ENTROPY MODELLING USING HISTOGRAMS OF THE ROWS AND COLUMNS AS CONSTRAINTS.

	Histogram modelling		
Dataset	*k-Means clustering error*	*PCA*	*Maximum correlation*
Random	0.999	0.999	0.937
Gaussian	0.999	0.999	0.999
Component	**0.001**	**0.001**	**0.001**
Cluster	**0.001**	**0.001**	**0.001**
Alon	**0.001**	**0.001**	**0.001**
Gene	**0.001**	**0.001**	**0.001**
Thalia	0.999	0.975	**0.001**

databases. Formally,

$$p(M(D)) = \frac{|\{S \in \mathcal{S} \mid M(S) \leq M(D)\}| + 1}{|\mathcal{S}| + 1} \ .$$

Essentially, a p-value approximates the probability that the structural measure M has value equal to or less than the observed value in the original database D. A small p-value means that the examined result has a low probability of stemming from the background model, that is, a low p-value indicates the structure is significant when contrasted with the encoded background knowledge. (We note that in the above discussion we assumed that S is more structured than D according to the measure M when $M(S) \leq M(D)$. For structure measures that are larger when the structure is stronger, the inequalities should be reversed.)

Tables V and VI present p-values for the three structural measures under consideration, respectively for Mean-Variance and Histogram modelling. Many results are not surprising and simply validate our approach (e.g. the non-significant p-values for the Random and Gaussian datasets, and the significant p-values for the Component and Cluster datasets that do indeed exhibit genuine structure). On the Alon dataset, it is remarkable that the k-means clustering error for Histogram modelling is significant, while it is not for Mean-Variance modelling. This may suggest that the structure in this dataset is better explained by simple means and variances than by histograms. Analogously, the p-values for Thalia for k-Means and PCA are significant for Mean-Variance modelling but not for Histogram modelling.

Note that quite a few p-values are 0.999, which indicates that all randomizations have a stronger value for the structure measure than the given data. This is possible since modelling the row distributions will result in correlations between the columns, as explained above. In other words, the background knowledge is so rich that the randomizations will contain some structure. Thus, the test would be significant only if the structure is stronger than what can be expected by chance. If the structure is weaker than what can be expected by chance, the p-value will be close to 1.

E. Assessing local patterns in data

Next, we give an example of how our models can be used for ranking and pruning results of subgroup discovery.

Subgroup discovery is a supervised branch of pattern mining, i.e., one is interested in finding patterns that correlate with a given target, that typically considers real-valued data. Like all areas of pattern mining, also subgroup discovery suffers from the pattern explosion: typically prohibitively many results are returned, many of which are variations of the same theme, and hence redundant [9]. By assessing the statistical significance of the returned patterns with regard to the background knowledge, and discarding insignificant results, we can strongly alleviate this problem.

A complicating factor when assessing many results empirically is the multiple testing effect, which will cause many p-values to be small even if there is no true structure. A common approach to resolve this is the Bonferroni correction, which multiplies the p-values by the number of patterns considered, in order to arrive at a meaningful p-value that can be compared against at standard significance levels.

Because of this, empirical p-values [13], [12] are unsuited to evaluate significance of local patterns. The smallest computable empirical p-value is one divided by the number of samples, and the significances of all patterns with an actual p-value smaller than this number are indistinguishable. Additionally, after Bonferroni correction, the smallest empirical p-value that can be computed is the ratio of the number of patterns and the number of randomizations. Thus, if a significance level of one percent is required, we would have to use a number of randomizations equal to 100 times the number of patterns, which is prohibitive in most practical cases. In contrast, our modelling strategies allow the analytical computation of p-values, with infinite resolution and range, and at very limited computational cost.

In subgroup discovery one of the most prominently used interestingness measures is Weighted Relative Accuracy (WRAcc) [16]. WRAcc measures the predictive accuracy of a classification rule $B \rightarrow H$, where the body, B, is a set of conjunctive conditions and the head, H, a binary target

Figure 2. Exact analytical p-values (left plot) and empirical p-values (right figure) obtained from 10 000 database samples against sorted weighted relative accuracy (WRAcc) values. The horizontal axis is log-scaled. The dashed orange line indicates the maximal (uncorrected) precision for the empirical p-values (i.e. 10^-4, which corresponds to 1 dataset out of the 10 000 samples). Further, note the correlation between analytical p-values and WRAcc.

label. Each condition specifies a lower or a higher value of an attribute A, eg. $A \leq c$, where c a constant. Formally the WRAcc measure is defined as

$$WRAcc(B \to H) \quad = \quad P(H,B) - P(H)P(B). \quad (13)$$

It is our goal to compute a p-value for each subgroup, quantifying the significance of the WRAcc as a test statistic, with the background model as null hypothesis. Since subgroup discovery is not the main focus of this paper, we will only give an outline of how this can be done.

Observe that $P(H,B)$ can be computed on any given dataset as the average of an indicator variable over the rows (where the indicator variable is one if both H and B are true). In a random dataset, $P(H,B)$ can thus be formalized as the average of a number of Bernoulli random variables, one for each row, and these probablities can be computed directly from our background model. Similarly, $P(B)$ is the average of indicator variables indicating whether B is true. On a random dataset, it will be the average of corresponding Bernoulli random variables. Note further that $P(H)$ is a constant (the proportion of rows in the target class) as we do not randomize the labels.

Assuming that these averages are taken over a sufficiently large number of variables, the central limit theorem allows us to approximate their distribution by a Gaussian distribution. The means $\mu_{P(H,B)}$ and $\mu_{P(B)}$, and variances $\sigma^2_{P(H,B)}$ and $\sigma^2_{P(B)}$ of these averages can be computed from the means and variances of the individual Bernoulli random variables.

Combining the random averages $P(H,B)$ and $P(B)$ as in the definition for the WRAcc, taking care to account for the dependencies between these random variables, we obtain the following results for the mean and variance of the WRAcc for a given rule:

$$\begin{aligned} \mu_{WRAcc} &= \mu_{P(H,B)} - P(H) \times \mu_{P(B)} \ , \\ \sigma^2_{WRAcc} &= (1 - 2P(H)) \times \sigma^2_{P(H,B)} - P(H) \times \sigma^2_{P(B)} \ . \end{aligned}$$

This specifies the Gaussian distribution approximating the WRAcc distribution, allowing us to analytically compute tight approximations to the exact p-values.

To put this to practice, for the *Alon* dataset, we mined the top-1 000 subgroups of $1, 2$, and 3 conditions using the publicly available open source subgroup discovery tool Cortana [4]. For each of the returned patterns, we calculated exact p-values as described above, using Histograms modelling to construct the model. In addition we calculated empirical p-values from 10 000 databases sampled from the same model.

Figure 2 presents p-values for sorted WRAcc values. Regarding the leftmost plot for exact p-values, we observe that generally p-values increase as WRAcc values decrease. However, the rule with the smallest p-value does not present the largest WRAcc value. In general, for other rules as well, it is easy to see that sorting according to the WRAcc values is not identical with sorting with respect to exact p-values. For comparison, the right plot shows the p-values one could obtain using empirical hypothesis testing, with 10 000 randomizations. Clearly, the resolution is limited, and this approach is unable to rank the most interesting subgroups according to p-value as they are all smaller than $1/10\,000$.

V. Discussion

The experiments showed that maximum entropy modelling indeed provides a well-founded, fast, and versatile alternative to swap randomization.

Easily accessible, yet informative, structural background knowledge such as means and variances, or histograms, of rows and columns are efficiently incorporated into our models. Subsequently sampling datasets from our models is highly efficient, and for sampling 10 000 datasets takes over one order of magnitude less time than by swap randomization. As our implementation of our models are not quite optimized, we expect significant further speed-ups to be attainable—and are particularly interested in evaluating different solving techniques, as for Histograms we saw that despite requiring many more iterations, conjugate gradient descend greatly outperformed Newton's method.

[4]Cortana: http://datamining.liacs.nl/cortana.html

Structure that follows trivially from the considered marginal distributions was correctly identified in the artificial data. Furthermore, for the Thalia gene expression dataset we saw that the discovered clusters, as well as 5 principal components, follow from the MDL optimal histograms.

As an example, we formalized how to calculate exact p-values for the weighted relative accuracies in subgroup discovery, and showed our model can be successfully applied to strongly reduce redundancy in the discovered set of patterns. Analogously, formulating the expected probabilities and p-values for other interestingness and/or similarity measures hence makes for promising future work, making assessment applicable both for filtering results afterwards, as well as for integration into the mining process. Application areas of particular interest include subgroup discovery and subspace clustering, as for both fields statistically sound redundancy reduction is an important open problem.

Last, but not least, we regard the further generalization of maximum entropy modelling as the most important line of future work, both to non-comparable real-valued attributes, and most importantly to databases containing both discrete and real-valued attributes. Such extensions could also include more complex background information, which in turn would allow for iterative data mining [7], [4].

VI. Conclusion

In this paper, we proposed a well-founded approach for assessing results on real-valued rectangular databases. We introduced maximum entropy models for general real-valued data that can be used to assess whether or not a discovery may be the trivial result of the row and column marginal distributions in the database. We gave theory for incorporating such distributions in the form of histograms over the rows and columns, as well as means and variances, as background knowledge into our models using well-founded convex optimization techniques.

Experiments showed our approach provides reliable statistical assessment, while being an order of magnitude faster than the state of the art. We showed its general applicability in empirical p-value testing, as well as for calculating exact probabilities. As an example of the latter, we discussed ranking statistically significant results in subgroup discovery.

Acknowledgements

JV is supported by a Postdoctoral Fellowship of the Research Foundation-Flanders, TDB and KNK by the EPSRC grant EP/G056447/1, the European Commission through the PASCAL2 Network of Excellence (FP7-216866) and a University of Bristol Centenary Scholarship.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Data pertaining to the article broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proc. Natl. Acad. Sci.*, volume 96, pages 6745–6750, 1999.

[2] D.P. Bertsekas. *Nonlinear Programmng*. Athena Scientific, 1999.

[3] T. De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, 23(3):407–446, 2011.

[4] T. De Bie, K.-N. Kontonasios, and E. Spyropoulou. A framework for mining interesting pattern sets. *SIGKDD Expl.*, 12(2):92–100, 2010.

[5] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *ACM Trans. Knowl. Discov. Data*, 1(3), 2007.

[6] P.D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[7] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila. Tell me something I don't know: randomization strategies for iterative data mining. In *Proc. KDD'09*, pages 379–388, 2009.

[8] E.T. Jaynes. On the rationale of maximum-entropy methods. *Proc. of the IEEE*, 70(9):939–952, 1982.

[9] A.J. Knobbe and E.K.Y. Ho. Maximally informative k-itemsets and their efficient discovery. In *Proc. KDD'06*, pages 237–244, 2006.

[10] P. Kontkanen and P. Myllymäki. MDL histogram density estimation. In *Proc. AISTATS'07*, 2007.

[11] K.-N. Kontonasios and T. De Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proc. SDM'10*, pages 153–164, 2010.

[12] M. Ojala. Assessing data mining results on matrices with randomization. In *Proc. ICDM'10*, pages 959–964, 2010.

[13] M. Ojala, N. Vuokko, A. Kallio, N. Haiminen, and H. Mannila. Randomization of real-valued matrices for assessing the significance of data mining results. In *Proc. SDM'08*, pages 494–505, 2008.

[14] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissemand, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinf.*, 22(9):1122–1129, 2006.

[15] U. Scherf, D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, L. Tanabe, K.W. Kohn, W.C. Reinhold, T.G. Myers, D.T. Andrews, D.A. Scudiero, M.B. Eisen, E.A. Sausville, Y. Pommier, D. Botstein, P.O. Brown, and J.N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, 24:236–244, 2000.

[16] L. Todorovski, P. Flach, and N. Lavrač. Predictive performance of weighted relative accuracy. In *Proc. PKDD'00*, pages 255–264, 2000.

[17] M. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foun. Trends Mach. Learn.*, 1(1-2):1–305, 2008.