Preservation of Statistically Significant Patterns in Multiresolution 0-1 Data

Prem Raj Adhikari and Jaakko Hollmén

Aalto University School of Science and Technology
Department of Information and Computer Science
P.O. Box 15400, FI-00076 Aalto
Espoo, Finland
prem.adhikari@tkk.fi, jaakko.hollmen@tkk.fi

Abstract. Measurements in biology are made with high throughput and high resolution techniques often resulting in data in multiple resolutions. Currently, available standard algorithms can only handle data in one resolution. Generative models such as mixture models are often used to model such data. However, significance of the patterns generated by generative models has so far received inadequate attention. This paper analyses the statistical significance of the patterns preserved in sampling between different resolutions and when sampling from a generative model. Furthermore, we study the effect of noise on the likelihood with respect to the changing resolutions and sample size. Finite mixture of multivariate Bernoulli distribution is used to model amplification patterns in cancer in multiple resolutions. Statistically significant itemsets are identified in original data and data sampled from the generative models using randomization and their relationships are studied. The results showed that statistically significant itemsets are effectively preserved by mixture models. The preservation is more accurate in coarse resolution compared to the finer resolution. Furthermore, the effect of noise on data on higher resolution and with smaller number of sample size is higher than the data in lower resolution and with higher number of sample size.

Keywords: Multiresolution data, statistical significance, frequent itemset, mixture modelling.

1 Introduction

Biological experiments performed with high throughput and high resolutions techniques often produce data in multiple resolutions. Furthermore, International System for human Cytogenetic Nomenclature (ISCN) has defined five different resolutions of the chromosome band: 300, 400, 550, 700 and 850[1]. In other words, chromosomes are divided into 862 regions in resolution 850 (fine resolution) and 393 regions in resolution 400 (coarse resolution). Thus, data are available in different resolutions and methods needs to be devised to work with multiple resolutions of the data. However, current standard algorithms only work with a single resolution of data. So, sampling in different resolutions possesses

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 86–97, 2010.

high importance. In this paper, we model multiresolution data and use statistical significance testing on data generated by generative models. Finite mixture models are generative models [2,3] able to generate the potentially observable data. Over the years, finite mixture models have been extensively used in many application domains including model based clustering, classification, image analysis, and collaborative filtering in analysis of high dimensional data because of their versatility and flexibility. In spite of the wide application areas of mixture models, the evaluation of mixture models are often based on the likelihood of the model on the original data, not by testing the data generated by the generative models.

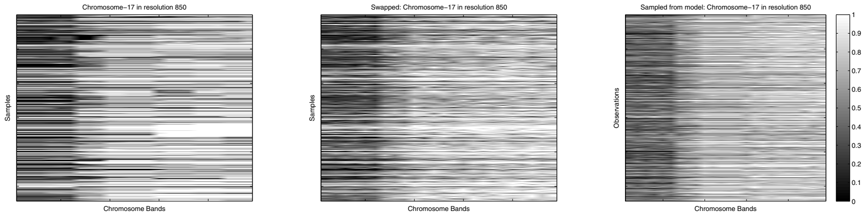
In [4], the authors used HMO (Hypothetical Mean Organism) motivated from Bacteriology [5] and maximal frequent itemsets[6] to define the data to the domain experts in a compact and understandable manner. Furthermore, in [7], the authors also compared the frequent itemsets [8,9] extracted from each cluster to that extracted globally showing that the frequent itemsets were significantly different. However, the authors failed to consider the significance of the itemsets and their preservation by generative models. Study of patterns generated by the generated models has received little interest. However, preserving patterns from the original data should be essentially an important property of mixture models and if properly designed can be one of the benchmarks for selecting better mixture models. In this paper, we experiment with finite mixture models of multivariate Bernoulli distribution to test whether the statistically significant itemsets are preserved by mixture models. We also extend the ideas in [10] to observe if the significant itemsets are preserved by the sampling in different resolutions.

Novelties in this paper are determination of presence of statistically significant itemsets with respect to sampling different resolutions and especially by the data generated through the generative mixture models. Furthermore, we experiment the mixture model with different levels of noise showing that the trained mixture models are robust to noise in lower resolution and when there is significant amount of data to train and constrain the mixture model thus showing the importance of working in multiple resolutions which is useful for database integration.

Rest of the paper is organized as follows: Section 2 presents the dataset used in the experiments. Section 3 reviews the theoretical framework for experiments including sampling, randomization and mixture modelling. Section 4 explicates the experiments performed on the data and discusses the obtained results. Section 5 draws conclusions from the experimental results.

2 DNA Copy Number Amplification Dataset

The dataset used in the experiments defines DNA amplifications in different chromosomes. Amplification is the special case of duplication where the copy number increases more than 5 [11]. The data was collected by bibliomics survey of 838 journal articles during 1992-2002 by hand without using state-of-the-art



(a) Original Data (b) Randomized Data (c) Sampled from Model

Fig. 1. DNA copy number amplifications in chromosome-17, resolution 850. $\overline{\underline{\boldsymbol{X}}} = (X_{ij})$, $X_{ij} \in \{0, 1\}$. Each row represents one sample of the amplification pattern for a patient and each column represents one of the chromosome bands.

text mining techniques [4,12]. The dataset contained information about the amplification patterns of 4590 cancer patients in resolution 400. There was another set of similar data but in resolution 850 with higher sample size. The dataset shown in Figure 1 contains the original data in resolution 850, the randomized version and sampled from the mixture model. Each row describes one sample of cancer patient while each column identifies one chromosome band(region). The amplified chromosome regions were marked with 1 while the value 0 defines that the chromosome band is not amplified. Patients whose chromosomal band had not shown any amplification for specific chromosome were not included in the experiments since we are interested in modelling the amplifications, not their absence.

3 Theoretical Framework

Determining the significance of the results obtained by any algorithm or method is an actively researched area. Statistical significance testing have often been implemented to determine the significance of the results. In this paper, we implement our statistical significance testing on data in multiple resolutions and data generated by mixture models.

3.1 Sampling Resolutions

We have recently in [10] suggested three downsampling and a simple upsampling technique for 0-1 data and performed experiments on them showing that the methods are fairly similar. Upsampling is the process of changing the resolution of data from coarse resolution to finer resolution and downsampling is the process of changing the resolution of data from fine resolution to coarse resolution. Upsampling makes multiple copies of similar chromosome bands in higher resolution. Downsampling, in turn, proceeds with one of three different methods: OR-function, Majority decision and Weighted Downsampling. In OR-function downsampling, a cytogenetic band in lower resolution is amplified if any of the bands in higher resolution which combines to form the cytogenetic band in the lower resolution is amplified. In majority decision downsampling method, the cytogenetic band in lower resolution is amplified if majority of the cytogenetic band in higher resolution are amplified. In weighted downsampling method, the length of the cytogenetic bands are considered. The cytogenetic band in lower resolution is amplified if the total length of amplified band is higher than that of the unamplified band.

3.2 Randomization

Statistical significance testing on datasets are not trivial as the data belongs to a class of empirical distributions thus integrating over the PDF(Probability Density Function) to calculate the $p-$values is often not possible. Furthermore, given the data set $\mathcal{D}$, its PDF or true generating model is often unknown. It is trivial to integrate over the empirical distribution where a null distribution can be fixed and samples can be drawn from the null distribution. Randomization [13] is one of the method to sample from null distribution and it has been proposed with some plausible results and implemented in various application areas such as redescription mining [14]. Comparing segmentations of genomic sequences [15] among many others.

Consider a 0-1 dataset, $\mathcal{D}$ with m rows and n columns. Let $\mathcal{D}_1, \mathcal{D}_2 \ldots \mathcal{D}_n$ be the randomized data produced using the randomization approach repeated n times. Also, consider a data mining algorithm $\mathcal{A}$, for instance frequent set mining and mixture modelling in our case which is run on the data $\mathcal{D}$ with the result $\mathcal{A}(\mathcal{D})$. The result $\mathcal{A}(\mathcal{D})$ determines the structural measure of the dataset $\mathcal{D}$, the frequencies of frequent itemset and likelihood in our case. The randomized datasets $\mathcal{D}_1, \mathcal{D}_2 \ldots \mathcal{D}_n$ are also subjected to the algorithm $\mathcal{A}$ producing results $\mathcal{A}(\mathcal{D}_1), \mathcal{A}(\mathcal{D}_2) \ldots \mathcal{A}(\mathcal{D}_n)$. The task is then to determine whether the result on the original data is different from the results on the randomized data. Empirical $p-$values can be used for the same purpose.

Null Distribution: Given a binary dataset $\mathcal{D}$, the null distribution considered in the paper are all the datasets satisfying all the following properties:

1. The dataset of the same size i.e. number of rows and columns of randomized data is equal to the number of rows and columns of the original data.
2. The dataset with same row and column margins. Margins here describes the sums. Thus, row and column sums are exactly fixed. This automatically preserves the number of ones in the dataset i.e. the number of amplifications.

As the the constraints discussed above increases, the randomization is becomes more conservative. However, the main focus is to compare the results obtained with the original dataset with closely related datasets. Furthermore, the number of datasets satisfying the above constraints are still significantly high. Generally, the application area determines the constraints of the randomization. Maintaining row and column margins in this case is adapted from the idea in [13] which seems relevant in our case considering the fact that most of the binary datasets especially in the field of biology such as the amplification data discussed in Section 2 are often spatially dependent and sparse. On the other hand, if the randomization is not subjected to the constraints discussed above then any result of an algorithm turns out to be relevant. With lesser constraints, the number of randomized datasets to sample for convergence discussed in Section 4.1 increases which consequently increases the computational complexity of the approach. Experimental results in [13] have shown that complexity of using a data mining algorithm $\mathcal{A}$ on a dataset has significantly higher computational complexity compared to the generation of randomized dataset under the constraints discussed above. Similar to [13], the data is randomomized in the with repeated 0-1 swaps until convergence. The null hypothesis H_0 throughout this paper is that for all datasets $\mathcal{D}$ that satisfies the given constraints, the test statistic follows the same distribution. Test statistic used here is frequency or the support (α) in case of frequent itemset and sample likelihood in case of mixture models.

***p*—Values:** p-value can be defined as probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true [16,17]. Let $\hat{\mathcal{D}} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_k\}$ be the randomized versions, sampled i.i.d from the null distribution, of the original data $\mathcal{D}$. The one-tailed *empirical $p-$value* of $\mathcal{A}(\mathcal{D})$ for $\mathcal{A}(\mathcal{D})$ being large is

$$\tilde{p} = \frac{1}{n+1}\left(\sum_{i=1}^{n} I(\mathcal{A}(\mathcal{D}_i) \geq \mathcal{A}(\mathcal{D})) + 1\right), \tag{1}$$

where $i \in \{1, 2 \ldots k\}$ and I is the indicator variable.

The Equation 1 gives the fraction of randomized dataset whose structural measure, itemset frequency (support) in case of frequent itemset and sample likelihood in case of mixture models, is greater than the original data $\mathcal{A}(\mathcal{D})$. In one-tailed $p-$value small value of $\mathcal{A}(\mathcal{D})$ are interesting and can be defined similarly for the two-tailed test. In this paper the randomized datasets are produced using Markov Chain Monte Carlo(MCMC) approach. The samples produced by MCMC are not independent thus diminishing the reliability of the $p-$values. To mitigate this problem and guarantee the ex-changeability of samples, we implement forward-backward approach discussed in [18]. The basic idea is to run the chain, a number of defined steps, say J backwards and forward after reaching J. In other words, given the original dataset $\mathcal{D}$, a dataset $\hat{\mathcal{D}}$ is obtained such that the path length between $\mathcal{D}$ and $\hat{\mathcal{D}}$ is J. The desired number of $\mathcal{K}$ samples of randomized data is obtained by running the chain J steps forward and obtaining the samples $\hat{\mathcal{D}_i}$ thus producing $\mathcal{D}, \hat{\mathcal{D}_1} \ldots \hat{\mathcal{D}_k}$ as the set of exchangeable samples. Furthermore, the $p-$values were adjusted for multiple hypothesis testing using the Holm-Bonferroni test correction[19].

3.3 Mixture Models of Multivariate Bernoulli Distribution

Cancer is not a single disease but a collection of several diseases. Furthermore, the amplification data discussed in Section 2 being high dimensional binary data, finite mixtures of multivariate Bernoulli distribution was selected as the model to model the amplification data. The finite mixture of multivariate Bernoulli distributions is defined as:

$$p(\mathcal{D}|\boldsymbol{\Theta}) = \sum_{j=1}^{\mathrm{J}} \pi_j \prod_{i=1}^{\mathrm{d}} \theta_{ji}^{x_i} (1 - \theta_{ji})^{1-x_i}, \tag{2}$$

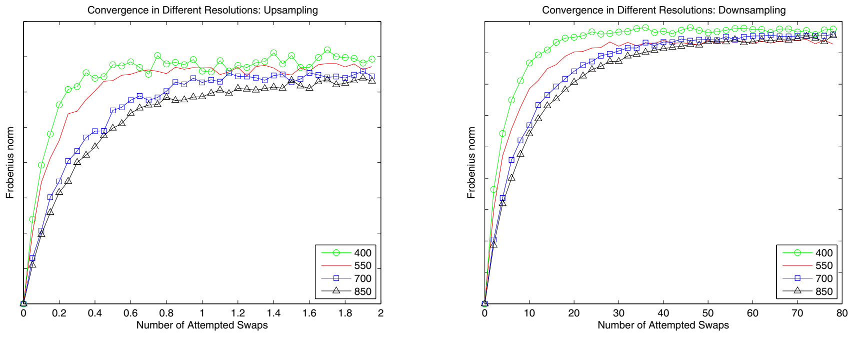
where the data is assumed to originate from the known number of components J. The mixture proportions π_j satisfy the properties such as convex combination such that $\pi_j \geq 0$ and $\sum_{j=1}^{\mathrm{J}} \pi_j = 1, \forall j = 1, \ldots J$. The model parameters $\boldsymbol{\Theta}$ is composed of $\theta_1, \theta_2, \theta_3 \ldots \theta_d$ for each component distribution.

Model selection in finite mixture modelling refers to the process of selecting number of mixture components, J in the data. 10-fold cross-validation [20,21] is used to select the optimal number of components taking parsimony into account. The process of model selection employed is similar to [4,10,22]. Since the mixture models are complex and sample size of data was small to constrain it, chromosome-wise mixture modelling was performed for data in different resolutions. Expectation Maximization algorithm [23,24] was used to train the mixture models using BernoulliMix[25] which is an open source program package for finite mixture modelling of multivariate Bernoulli distribution.

4 Experiments

4.1 Convergence Analysis of the Swaps

In order to determine the optimal number of swaps to be performed, convergence test for the randomized data was performed. In our experiments, the process of randomization is said to converge when the distance between the the original data and the randomized data changes the least with respect to the predefined difference measure. Similar to [13] and [26], the distance measure used here is the Frobenius norm between the original and the randomized matrix. In order to test the convergence, first the number of attempted swaps is fixed to 1 and increased by the step size of 1. The approach used here differs from [13] and [26] because they set the initialization point to $\mathcal{K}$ equal to the number of ones in the data and increase the number of attempts in multiples of $\mathcal{K}$. Such approach could prove beneficial in large datasets but since amplification dataset is small, it was very easy to compute the swaps thus making it easier to initialize number of attempted swaps to 1. Furthermore, similar dataset was available in resolution 850 with higher sample size. Thus, the convergence test was performed for both the data and their upsampled and downsampled versions as shown in Figure 2. Ten different instances of the swaps are performed and the mean of the results is taken as the final convergence test. Similar, convergence analysis was also performed for combined data and the sampled data. Convergence of sampled data was similar to the original data from which the model was trained. However, in case of combined data, convergence required relatively higher number of swaps i.e. 700000 swaps. Figure 2 shows that the swap converges when the number of



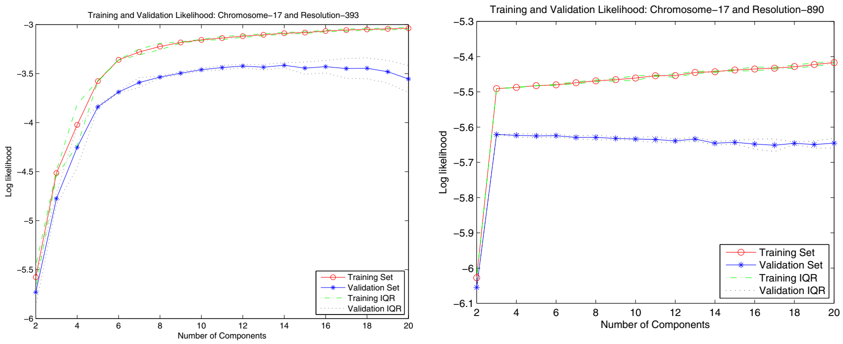
(a) Resolution 400 and Upsampled (b) Resolution 850 and Downsampled

Fig. 2. Convergence analysis for randomization with respect to 0-1 swaps

attempted swaps is approximately 16000 for original data in resolution 400. From the Figure 2 it can also be seen that the Frobenius norm increases rapidly until certain number of attempted swaps and then tends to stabilize. The stabilizing point is taken as the convergence. As discussed in Section 2, the sample size of data in resolution 850 was high thus taking longer time to converge. The number of swap attempted to get the randomized data in this case is 600000.

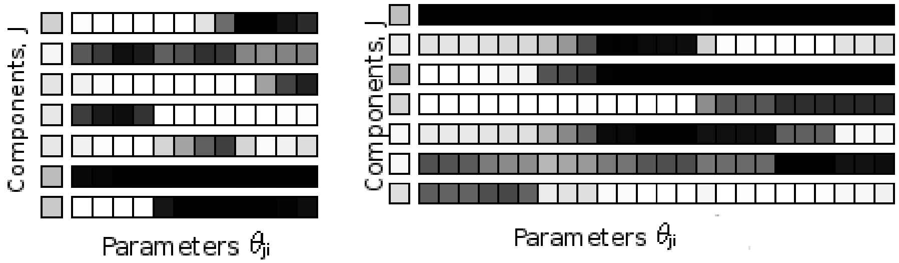
4.2 Model Selection in Mixture Model

Model selection in the context of mixture modelling is the selection of number of components of the mixture model. It is often recommended to repeat cross-validation technique a number of times, at least 10, because a 10-fold cross-validation can be seen as a "standard" measure of the performance whereas ten 10-fold cross-validations would be a "precise" measure of performance[27]. In addition, EM-algorithm is highly sensitive to initializations and the global optimum is not often guaranteed [28]. Therefore, the cross-validation procedure was repeated 50 times. Since the analysis was performed chromosome-wise, the data dimension was relatively less. Thus, the number of mixture components were varied between 2 and 20. Using higher number components can overfit the data. Furthermore, our major goal, as in [4], was to generate compact and parsimonious models. The log-likelihood was averaged for each component and the interquartile range(IQR) was calculated. Furthermore, the model selection procedure was also performed for the randomized data. In Figure 3a, both training and validation likelihood are smoothly increasing curves with low variation in IQR. The number of components selected in this case is 7, taking the parsimony into account. We also performed similar model selection procedure on the randomized data as shown in Figure 3b. It was found that there is no well defined clustering structure present in the data with respect to the mixture models. Furthermore, the results on randomized data also proves that the data is not a random data but there is a well-defined structure present in the data which mixture model is able to extract.



(a) Original combined Resolution 400 (b) Randomized combined Resolution 400

Fig. 3. Model Selection procedure and Model visualization: Example case in combined data of Chromosome-17 in resolution-400 and its corresponding randomized version. Corresponding IQR (Inter Quartile Range) for each training and validation run has also been plotted.



(a) Resolution 400 (b) Resolution 850

Fig. 4. Two different models for the combined data trained in resolution 400 and 850

After selecting the number of components, ten different models were trained to convergence and best of the trained models were used to calculate the likelihood on data as shown Figure 5b. The model was also used to sample the data to calculate the significant itemsets in the sampled data. Figures 4a and 4b are the final models trained to convergence for combined data in resolution 400 and 850 respectively. Similarity of the models can be tracked visually from the model visualization as in Figure 4. For example, component 6 in Figure 4a corresponds to component 1 in 4b.

4.3 Significance of Frequent Itemsets and Data Samples

In the experimental setup, first the frequencies of the itemsets of the size two were determined from the original data. The itemsets of size three and above were discarded from the experiments for simplicity and space constraints for explaining the results. However, results in [10] has shown that generally the frequent itemsets in the amplification data discussed in Section 2 are large and consecutive. The core of the work was to determine if the statistically significant itemsets were preserved in different resolutions and by the generative mixture model. First the itemsets of size two which had a frequency or support($\alpha \approx 0.5$) were determined and the original data was then subjected to randomization. Randomization produces 100000 samples of randomized dataset. Larger number of random samples are chosen because Holm-Bonferroni [19] used to correct for multiple hypothesis requires higher number of samples for plausible results. The structural measure used to calculate the $p-$values in our case is the support or the frequency of the itemsets. The choice of frequency or support($\alpha \approx 0.5$) is arbitrary but motivated by majority voting protocol and constraining the number of frequent itemsets thus making it easier to interpret and report. Furthermore, itemsets with very low support but statistically significant are not highly interesting. The samples of data were generated equal to the number of samples in the original data. Similarly, the data generated from the trained mixture models were also subjected to randomization to determine the statistically significant itemsets.

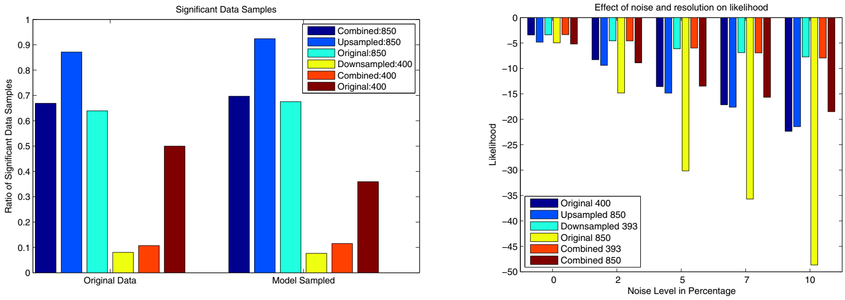
Table 1. Itemsets of size 2 with their frequency (support) in original as well as sampled resolution. Results of Downsampling have been omitted because of space constraints. The symbol ${}^{item}_{n}\boldsymbol{C}^{item}_{r}$ suggests combination where subscript n and r determines n choose r in the combination and superscript determines the item to start and end the combination.

Significant itemsets of Size 2 at $\alpha = 0.05$			
Data	**Support**	**Original Data**	**Model Sampled**
Original 393	.4	{9,10}, {11,12}	{9,10}, {11,12}
Upsampled 850	.4	${}^{10}_{5}\boldsymbol{C}^{14}_{2}$, ${}^{15}_{4}\boldsymbol{C}^{18}_{2}$, ${}^{19}_{6}\boldsymbol{C}^{24}_{2}$	${}^{10}_{5}\boldsymbol{C}^{14}_{2}$, ${}^{19}_{6}\boldsymbol{C}^{24}_{2}$
Combined 393	.6	{ 5, 7}, { 5, 12}, ${}^{8}_{6}\boldsymbol{C}^{12}_{2}$	{ 5, 7}, { 5, 12}, { 7, 12}, ${}^{8}_{6}\boldsymbol{C}^{12}_{2}$
Combined 850	.6	${}^{10}_{6}\boldsymbol{C}^{15}_{2}$, {12,16}, {12,17}, {12,18}, {13,16}, {13,17}, {13,18}, {14,16}, {14,17}, {14,18}, ${}^{15}_{10}\boldsymbol{C}^{24}_{2}$	${}^{10}_{6}\boldsymbol{C}^{15}_{2}$, {12,16}, {12,17}, {12,18}, {12,20}, {13,16}, {13,17}, {13,18}, {13,20}, {14,16}, {14,17}, {14,18}, {14,20}, ${}^{15}_{10}\boldsymbol{C}^{24}_{2}$

The $p-$values were calculated to test the significance of the itemsets. The statistically significant itemsets computed at significance level (α)= 0.05 in the original data and the sampled data from the model is compared and analyzed. Table 1 shows that significant itemsets are approximately but not exactly preserved by the generative mixture model as well as the sampling of resolutions. Difference is subtle in higher resolution. The itemsets in lower resolution correspond to itemsets in higher resolution. For example, itemset {11,12} in resolution 400 corresponds to itemset ${}^{19}_{6}\boldsymbol{C}^{24}_{2}$ in resolution 850. It is to be noted that not all frequent itemsets are significant and not all significant itemset are frequent. For example, in case of combined resolution 400, itemset {1,2} is significant where as it is not frequent. Furthermore, itemset {7,12} is frequent but not significant.

We also determined the number of significant data samples in different resolutions and from the sampled model. Figure 5a suggests that numbers of significant data vectors are preserved in the generative models. During our experiments, we also determined the indices of the significant data vectors and it was seen that indices of the significant vectors are not preserved i.e. generated of samples of data are not arranged in similar manner to original data. Furthermore, it was also seen that finer resolution has higher number of significant data samples because with increasing dimension the uniqueness of the rows increases and the 0-1 swap strategy used in the randomization ceases to function properly. However, this has little or no significance because of i.i.d assumption for each data sample.

4.4 Effect of Noise on the Likelihood



(a) Significant Data Samples

(b) Effect of Noise

Fig. 5. Ratio of significant data samples to the number of samples in the left panel and effect of noise and resolution on the likelihood in right panel

We added random noise to the data. Since the data was binary data, adding noise is simply flipping the bits i.e. changing ones to zeros and zeros to ones. Addition of 5% noise means that 5% of total data items in the dataset are flipped. Figure 5b shows that the effect of noise will be significantly higher for data in finer resolution than the data in the lower resolution. Furthermore, when the number of samples is low (Cases: Original 400 and Upsampled to 850), the difference in the likelihood is large because the number of samples are too low to constrain the mixture model. However, when the number of samples are increased, as in case of combined datasets, the variation in likelihood is not significant. Nevertheless, likelihood for the data in the higher resolution deviates significantly even when the sample size is increased.

5 Summary and Conclusions

We use statistical significance testing on data in different resolutions and on data generated by the generative mixture models using randomization. From the experiments we conclude that finite mixtures of multivariate Bernoulli distribution retains the significant itemsets and the significant data vectors in the original data even when the mixture model is trained parsimoniously. Furthermore, experiments with different levels of noise on the data shows that models parsimonious models in coarse resolution are more robust to noise. Nevertheless, when there is adequate amount of data to constrain the mixture model, the effect of noise diminishes significantly even in higher resolution.

References

1. Shaffer, L.G., Tommerup, N.: ISCN 2005: An International System for Human Cytogenetic Nomenclature(2005) Recommendations of the International Standing Committee on Human Cytogenetic Nomenclature. Karger (2005)
2. McLachlan, G.J., Peel, D.: Finite mixture models. In: Probability and Statistics – Applied Probability and Statistics Section, vol. 299. Wiley, New York (2000)
3. Everitt, B.S., Hand, D.J.: Finite mixture distributions. Chapman and Hall, Boca Raton (1981)
4. Hollmén, J., Tikka, J.: Compact and understandable descriptions of mixtures of bernoulli distributions. In: R. Berthold, M., Shawe-Taylor, J., Lavrač, N. (eds.) IDA 2007. LNCS (LNAI), vol. 4723, pp. 1–12. Springer, Heidelberg (2007)
5. Gyllenberg, M., Koski, T.: Probabilistic models for bacterial taxonomy. International Statistical Review 69, 249–276 (2000)
6. Burdick, D., Calimlim, M., Gehrke, J.: Mafia: A maximal frequent itemset algorithm for transactional databases. In: ICDE, pp. 443–452 (2001)
7. Hollmén, J., Seppänen, J.K., Mannila, H.: Mixture models and frequent sets: Combining global and local methods fordata. In: SDM (2003)
8. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 207–216. ACM, New York (1993)
9. Mannila, H., Toivonen, H., Verkamo, A.I.: Efficient algorithms for discovering association rules. In: Fayyad, U.M., Uthurusamy, R. (eds.) AAAI Workshop on Knowledge Discovery in Databases (KDD-94), Seattle, Washington, pp. 181–192. AAAI Press, Menlo Park (1994)
10. Adhikari, P.R., Hollmén, J.: Patterns from multiresolution 0-1 data. In: UP '10: Proceedings of the 16th ACM SIGKDD. ACM, New York (to appear, 2010)
11. Bishop, J.F.: Cancer facts: a concise oncology text. Harwood Academic Publishers, Amsterdam (1999)
12. Myllykangas, S., Tikka, J., Böhling, T., Knuutila, S., Hollmén, J.: Classification of human cancers based on DNA copy number amplification modeling. BMC Medical Genomics 1, 15 (2008)
13. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Transactions on Knowledge Discovery from Data 1(3), 14 (2007)
14. Gallo, A., Miettinen, P., Mannila, H.: Finding subgroups having several descriptions: Algorithms for redescription mining. In: SDM, pp. 334–345 (2008)
15. Haiminen, N., Mannila, H., Terzi, E.: Comparing segmentations by applying randomization techniques. BMC Bioinformatics 8(1), 171 (2007)

16. Schervish, M.J.: P values: What they are and what they are not. American Statistician 50(3), 203–206 (1996)
17. De La Horra, J., Rodriguez-Bernal, M.T.: Posterior predictive p-values: What they are and what they are not. Test 10(1), 75–86 (2001)
18. Besag, J., Clifford, P.: Generalized monte carlo significance tests. Biometrika 76(4), 633–642 (1989)
19. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70 (1979)
20. Geisser, S.: A predictive approach to the random effect model. Biometrika 61(1), 101–107 (1974)
21. Monsteller, F., Tukey, J.: Data analysis including statistics. In: Lindzey, G., Aronson, E. (eds.) Handbook of Social Psychology, vol. 2. Addison-Wesley, Reading (1968)
22. Tikka, J., Hollmén, J., Myllykangas, S.: Mixture modeling of DNA copy number amplification patterns in cancer. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 972–979. Springer, Heidelberg (2007)
23. Wolfe, J.H.: Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research 5, 329–350 (1970)
24. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38 (1977)
25. Hollmén, J.: BernoulliMix: Program package for finite mixture models of multivariate Bernoulli distributions (May 2009), `http://www.cis.hut.fi/jHollmen/BernoulliMix/`
26. Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., Mannila, H.: Tell me something I don't know: randomization strategies for iterative data mining. In: KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 379–388. ACM, New York (2009)
27. Gay, S.D.: Datamining in proteomics: extracting knowledge from peptide mass fingerprinting spectra. PhD thesis, University of Geneva, Geneva (2002)
28. Mclachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 1st edn. Wiley Interscience, Hoboken (November 1996)