Working Paper Series

A Note on Topological Concepts for Complexity Reduction of Binary Survey Data

Christian Buchta
Sara Dolnicar
Robert Köck
Edith Skriner

Working Paper No. 62
January 2000

Working Paper Series

January 2000

SFB
'Adaptive Information Systems and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

A Note on Topological Concepts for Complexity Reduction of Binary Survey Data

Christian Buchta†, *Sara Dolnicar*†, *Edith Skriner*†, *Robert Köck*†

† Department of Tourism and Leisure Studies,
Vienna Univ. of Economics and Business Administration,
Augasse 2-6, A-1090 Vienna, Austria/Europe.
e-mail: firstname.lastname@wu-wien.ac.at

Abstract

Recently in Steiner 1999 [7] the use of topological concepts was suggested for deciding on the level of complexity reduction of continuous data sets by quantisations. Level of complexity is synonymous with the number of classes or class representing means (centroids) of a data partition. In this paper a simulation study of the applicability of this concept to binary data is presented which is comparable with the results in Weingessel et al. 1999 [9]. The class of quantisation methods used contains traditional *k-means* and three robust approaches to partitioning, as suggested in Pötzelberger and Strasser 1997 [4].

1 Introduction

The need of reducing the level of complexity of consumer survey data can be motivated by two prominent aims. One is the classical aim of marketing to segment consumers into homogenous and targetable groups. Another is the statistical or data analytic aim to characterise a consumer population. Practically, such analysis can operate only on a reduced event set definition, i.e. on the classes of a data partition, see Strasser 1999 [8]. Thus *level of complexity* is synonymous with the number of classes (groups, segments, clusters) of a data partition, and membership of a consumer to a class is defining analysable events. It is no surprise, then, that a first question concerning the analysis of survey data is, which level of complexity to choose. Typically, the hope is that the data itself will give the analyst a hint. A well known approach to this end is the computation of indexes for choosing an optimal class (cluster) number, see Milligan and Cooper 1985 [5].

In Weingessel et al. 1999 [9] a comprehensive simulation study of the ability of 14 cluster indexes to indicate the number of consumer types mod-

elled into artificial binary data sets is presented. Similarly, in Steiner 1999 [7] a topological concept for deciding on the level of complexity reduction of continuous data sets is presented. Though this index is based on a different concept than the cluster indexes it can be assessed in the same way by its ability to suggest a level which is indicative of the known structural characteristics of test data. In this paper a simulation study of the behaviour of the topologically motivated index on artificial binary data sets is presented which is comparable with the results for the cluster indexes. The class of quantisation methods used contains traditional *k-means* and three robust approaches to quantisations as suggested in Pötzelberger and Strasser 1997 [4].

2 Binary Data Scenarios

The simulations of this study are based on the 12-dimensional binary data sets as documented in Dolnicar et al. 1998c [3]. As the main aim of the analysis of artificial data is to gain insight into quantisation methods, features typical of empirical data are included one by one in the data sets in order to enable a systematic analysis. Except for the random scenario, there are 6 response-types modelled. Each type is modelled as a combination of 4 groups of a total of 12 indicator/measurement variables with a characteristic pattern of low/high (0.2/0.8) probabilities of, say, observing that a respondent finds a statement of a questionnaire to express his/her perception of a brand. In table (1) the scenarios are summarised. *Name* are the names of the scenarios, *Dist* indicates the distributional model of the indicators (i/d independence/dependence of the indicator variables), and *Bayes* indicates the Bayes classification rate. The latter shows one aspect of the level of difficulty of a scenario. High rates have the interpretation 'easy' and low rates 'difficult', see Dolnicar et al. 1998a [1] for details.

Name	Scen	Description	Dist	Bayes
0	0	Random	i	16.67
1a	1	Basic	i	82.98
1ad	1	Basic	d	69.52
1b	5	Niche Segment	i	88.85
1bd	5	Niche Segment	d	78.23
2	2	Unequal Latent Variable	i	82.99
2d	2	Unequal Latent Variable	d	69.88
3a	3	Medium Importance	i	48.93
3ad	3	Medium Importance	d	41.87
3b	6	Answer Tendencies	i	81.25
3bd	6	Answer Tendencies	d	71.08
4d	4	Bad Indicator	d	81.20
7	7	Asymmetric	i	81.99
8	8	Extreme Segment Size	i	87.71

Table 1:

Scenario 1a is the basic scenario. Each group has 3 indicators and the patterns modelling the types are well separated in the sense that the group patterns are not adjacent (topologically), see table (2). Further the types are represented by equally sized samples (1000). In scenario 2 different numbers of indicators per group (5-4-2-1) are used. In scenarios 3b and 7 separation of the group patterns is abandoned. In scenarios 1b and 8 unequal sample sizes are modelled for the types, (1000,300,700,3000,500,500) and (300,300,300,1700,1700,1700). And in scenario 3a low probabilities of indicator groups are replaced by medium probabilities resulting in more noise. Finally the independent scenarios are characterised by independence of the indicators of a group whereas in the dependent scenarios they are positively correlated. In scenario 4d the correlation of one indicator with the two other indicators of a group is low. Between the variables of different groups there is always independence according to their interpretation as measurement variables connected to latent dimensions.

	i_1, i_2, i_3	i_4, i_5, i_6	i_7, i_8, i_9	i_{10}, i_{11}, i_{12}
	1a, 1b, 2, 3a, 8			
t_1	0.8	0.8	0.2	0.2
t_2	0.2	0.2	0.8	0.8
t_3	0.2	0.8	0.8	0.2
t_4	0.8	0.2	0.2	0.8
t_5	0.2	0.8	0.2	0.8
t_6	0.8	0.2	0.8	0.2
	3a			
t_1	0.8	0.8	0.5	0.5
t_2	0.5	0.5	0.8	0.8
t_3	0.5	0.8	0.8	0.5
t_4	0.8	0.5	0.5	0.8
t_5	0.5	0.8	0.5	0.8
t_6	0.8	0.5	0.8	0.5
	3b			
t_1	0.8	0.8	0.8	0.8
t_2	0.2	0.2	0.2	0.2
t_3	0.8	0.2	0.2	0.2
t_4	0.2	0.8	0.2	0.2
t_5	0.2	0.2	0.8	0.2
t_6	0.2	0.2	0.2	0.8
	7			
t_1	0.8	0.8	0.2	0.2
t_2	0.2	0.8	0.8	0.8
t_3	0.2	0.8	0.8	0.2
t_4	0.8	0.2	0.2	0.8
t_5	0.2	0.2	0.2	0.8
t_6	0.8	0.2	0.8	0.2

Table 2:

3 Quantisation Methods and Data Transformations

In Pötzelberger and Strasser 1997 [4] a general class of quantisation methods based on convex functions is presented. For the present study 4 convex functions $f(\cdot)$ were used. One, $f(x) = ||x||^2/2$, corresponds to ordinary *k-means* and the others are variants of more robust partitioning procedures. That is, the optimal partitions should be less influenced by the extreme parts of the underlying data distributions. Specifically these functions are, $f(x) = ||x||$, which will be referred to as *Kohonen*, $f(x) = 2\ln(\cosh(||x||/2))$, referred to as *logistic*, and $f(x) = \sqrt{1 + ||x||^2}$, referred to as *Masters*.

The fix-point algorithm proposed to use with the convex functions consists of the following steps: compute the means of a partition, map the means to prototypes (according to the convex function used), determine a new partition and their class means, and finally repeat the whole procedure until there is no improvement in the *information* associated with a partition. The information of a partition depends on the convex function and is part of the topological index on which, in turn, the proposals on the number of data representing class means are based. This will be explained in the next section.

For the simulations the same scheme as proposed in Weingessel et al. 1999 [9] was adopted in order to avoid instabilities of the fix-point method: 10 fix-point solutions were generated and that with the maximal information was chosen as solution. This process was repeated 100 times over a range of 2 to 18 classes. Initial solutions were drawn from the set of patterns contained in the data scenarios.

Except for *k-means* it is at least necessary to centre the data in order to define *outliers* as observations way off the overall mean of the data. Further we can standardise the indicator variables, as suggested in Steiner 1998 [6], and ask if that transformation has an influence on simulation results. For binary data the standard deviation of an indicator i is a function of the mean $\sqrt{\bar{x}_i(1-\bar{x}_i)}$, so that standardising implies magnifying a dimension relative to an indicator distributed with $\bar{x}_i = 0.5$. For all convex functions the simulations were made on standardised data, for *k-means* and *Kohonen* additional simulations were made on centred-only data.

4 Data Representation

In Steiner 1999 [7] a method for generating proposals for suitable numbers of data representing *prototypes* (class means, centroids) is presented. The idea

is based on the estimation of the *statistical dimension* of a data set. E.g. a curve in two dimensions is in fact topologically one-dimensional whereas the data describing the curve has *algebraic* dimension of two. On a logarithmic scale, the information of partitions based on different numbers of classes were shown to lie on a straight line for uniformly distributed data. Conversely data with specific distributional (normal) or topological (squares, rings, etc.) properties were shown to exhibit *breaks* in the statistical dimension line. These observations were combined to use the change in the slope of the dimension line as an indicator for suitable numbers of classes, and termed a *utility*. More specifically, only positive utilities are of interest because they indicate an upward bend or a higher dimensional view of the data. Further, the class numbers associated with the peaks of the utilities were shown to be in fact indicative of the *macro* structural properties of the data sets analysed, e.g. indicating that a straight line is topologically one-dimensional. The magnitude of the utilities was used to rank the proposals in order to have a best, second best, etc. proposal. Finally, the standard deviation of the utilities from a simulation was suggested as a measure of the certainty of a proposal.

In the present paper a further measure of the certainty of a suggestion is introduced: the number of times the utility is positive. Since the utilities of a uniform data set may be expected to vary around zero nearly the same number of positive and negative utility values might be expected in a simulation. Conversely, there may be the case of a high certainty but on average a low positive or high negative utility value.

For comparison with Weingessel et al. 1999 [9] the maximum positive utility over the comparable range of numbers of classes (3–12) is used to generate *proposals*. These can then be compared against the number of types modelled in the data scenarios so that it can be assessed if the utilities indicate their correct number (6).

Finally, similarities between partitions with k and $k+1$ data representing class means are computed. The similarity of two partitions i, j is defined as the maximal percentage of data points with pairs of matched class indexes $(i_1, j_1), (i_2, j_2), \ldots, (i_k, j_k)$. This measure informs a data analyst what difference it would make to use one more or one less class than suggested. Also the similarities might be indicative of structural breaks or if a data set can be regarded as rather 'homogenous'.

5 Recoverability

In Dolnicar et al. 1998a and 1998b [1, 2] the *reproduction capability* of different quantisation methods is investigated. A basic idea for assessing reproduction capability is mapping the type means of the scenarios and the class means of a partition to binary 'prototypes' and match them. Specifically this implies setting a one for a binary prototype if the type/class mean of an indicator variable is greater than 0.5 and setting a zero otherwise. In fact the binary prototypes are the *medians* of the classes. The prototypes are then considered to be matching if they are identical. Since empirical data can be expected to be less structured than artificial data sets, requiring a stronger definition of recoverability does seem a little bit too demanding.

Unfortunately, this idea cannot be applied to scenarios 3a and 3ad because indicators with probability 0.5 are not decidable to be above or below average. Thus these scenarios were left out from the simulations on recoverability. For the remaining scenarios simulations were made with *k-means* and *Kohonen* using 5, 6 and 7 classes. For 1b and 1bd further simulations were made with 4 classes. The rational behind these simulations was to gain insight into the relationship between breaks in the statistical dimension line and the recoverability of types. The number of types we hope to recover is 6 for all scenarios and the information of solutions with 5, 6, and 7 classes is needed to compute the utility for a data representation with 6 class means. For a utility to be positive the change in information for a shift from a representation with 5 to 6 class means must be greater than for changing from a representation with 6 to 7 class means because the information is increasing with increasing numbers of classes, see Pötzelberger and Strasser 1997 [4].

6 Results

6.1 Description

Tables (3–8) show the results for the artificial data sets by convex functions and centring/standardising of the data. Tables with heading *proposals* contain counts of how many times a specific number of classes is suggested. Figures in bold emphasise counts close to the total number of replications (100). Tables with heading *certainty* contain counts of how many times the utility value for a specific number of classes is positive. There are tables with *std* and *cen* in the headings, which indicates if the results are based on standardised or centred data. For comparison with Weingessel et al. 1999 [9] the range of the number of classes reported was reduced to 3–12.

Figures (1–16) contain each four curves and an overlaid bar plot for 2–18 classes (prototypes). The curves show the mean utility values (mu), the standard deviations of the utility values (su), the mean similarity values (ms), and the minimal similarity values (ls) over 100 replications. The bar plots show the bias in the certainty of the utility values (bu). The latter is the difference between the number of times the utility for a specific number of classes is positive and negative.

Finally, tables (9–10) show the results on the recoverability of types. The second column indicates the number of classes used. The figures to the left in the first block are counts of how many times no, one, two, etc. types of a scenario were found (recovered). The second block contains counts of how many times a specific type (t_i) was found.

6.2 Discussion

Scenarios 1a, 2, 3b, 4d, and 7 are with high certainty suggested to be best represented with 6 classes over different convex functions and irrespective of centring or standardising the data. An exception is scenario 3b where *Kohonen* and *Masters* suggest to use 5 classes. Further scenario 8 is consistently suggested to be best represented with 3 classes. For all other scenarios there is no such significant number, but a tendency to higher numbers of classes. For scenarios 1a, 2, 3b, 4d, and 7 it is remarkable to note that there are nearly no proposals with more or less than 6 classes, indicating that there is only one 'correct' representation of the data. It is also remarkable that for the corresponding dependent scenarios 1ad, 2d, and 3bd there are no consistent suggestions. For scenario 0 (random scenario) it is interesting to note that there are breaks for 4 and 8 classes, with 8 being more frequent. This is different to the results on uniformly distributed continuous data sets, as reported in Steiner 1999 [7].

In comparison with Weingessel et al. 1999 [9], for scenarios 1a,1b, and 2 (2a) the results are the same. Using the maximum utility value (over the range 3–12) as indicator we are able to recover the known number of types only under very limiting structural assumptions (1a, 2). Scenario 1b is an exception because in Weingessel et al. 1999 [9] there are two indexes which suggest to use 5 classes with high certainty.

If we consider the additional criteria suggested above we can at least state that for scenarios 1ad, 2d and 8 using *k-means*, and further for 3bd using *Kohonen*, *logistic*, and *Masters* there is a high certainty of observing a break in the statistical dimension line for 6 data representing class means. But there are other numbers where this is also true so that we are faced with a decision problem. Nevertheless, scenarios 1ad, 2, and 8 are then indicated to

be consistently representable by 6 classes. For scenario 1b, 4 and 5 classes seem to be a consistent choice but depending on the convex function used. For 1bd all convex functions do not 'suggest' more than 3 classes.

If we now turn to the plots, we see that the mean similarity curves of the scenarios based on independence models, 1a, 2, 3b, 7, and 8 peak after 6 data representing classes. For the dependent scenarios 1ad, 2d, 3bd, and 4d the mean similarity remains at the same level up to 10 classes. For 1b and 1bd there are a global and local peak after 5 classes. For both types of distributional models the level of similarity is high (0.8), so that using more than 6 classes could be interpreted as over-representing the data.

For the independent scenarios the standard deviation of the utilities shows a moderate increase with an increasing number of classes. For the dependent scenarios they are much higher for larger numbers of classes and show even dramatic increases for some number of classes where for 1ad, 3ad, 3bd and 4d they are also decreasing over different ranges. Similarly, the mean utilities are much higher and more erratic for the dependent scenarios than for the independent scenarios. Therefore proposals based on the mean utilities can be regarded as 'reliable' irrespective of the convex function used only for the independent scenarios 1a, 2, 3b and 7. An exception among the dependent scenarios is 4d.

For all scenarios except 1bd we observe significant breaks in the recoverability of types. The break occurs if 6 instead of 5 classes are used, except for 1b where the break occurs for 5 classes. After a break recoverability is the same or slightly better. An exception is again 1b using *k-means* where the niche type segment (2) is found consistently using 7 classes.

As a conclusion, it may be conjectured that a positive utility value for 6 (5) classes is for the most part 'due' to a break in recoverability. The breaks in recoverability coincide with the breaks in the statistical dimension line (compare the 'certainty' figures) but as discussed earlier the magnitudes of their utilities are dominating other breaks only for scenarios 1a, 2, 3b, 4d and 7. For scenario 8 there is a break in recoverability for 6 classes but the break in the statistical dimension line occurring for 3 classes is dominating.

7 Conclusions

The differential diagnosis attempted by using utilities, their standard deviations, certainty of a positive value and similarities on the one hand, and using the number of types recovered, on the other, leads to the conclusion that the binary test data used in this simulation study are topologically 'simple': either a prominent number of data representing classes is 'suggestible', or not.

Maybe therefore, the topological concept seems to be no superior approach in comparison to the simulation results for the indexes which suggest suitable numbers of clusters. Nevertheless, using breaks in the statistical dimension line which occur consistently in a simulation, we are faced with the difficulty of having no clear decision rule saying which number of classes to choose. It can be suspected that the cluster indexes are faced with the same problem on this type of data.

Since in general empirical binary data may be suspected to have no clear topological structure it might be worthwhile to consider the use of application specific criteria to decide on a suitable data representation. For instance, characterisability of classes typically deteriorates with a decreasing number of classes, or the class size (number of members) becomes economically infeasible with an increasing number of classes. Nevertheless if the aim is to find niche segments it is necessary to use larger numbers of classes. However, lack of a topological structure as indicated by the utilities does not necessarily imply that the data has no structure at all. There is evidence from empirical data sets that the utilities may show no significant peaks but that for a fixed number of classes the similarity of partitions is high. Further partitions with different numbers of class may contain similar means. This is not uncommon if specific data patterns have a high frequency and the remaining patterns are more uniformly distributed.

Future work should therefore report on empirical data as well as make attempts at more realistic data scenarios, especially the number of indicators is typically higher in empirical studies. Last but not least it should be studied if empirical data with interesting properties have also interesting topological properties. It can be suspected that 'separational' properties of the data patterns play a key role, but that for binary data such properties must be more 'pronounced' than for continuous data (such as in the basic scenario). If so, current practice of questionnaire design might be rethought and adapted to the 'special case' of 'binary' statements.

References

[1] **Sara Dolnicar, Friedrich Leisch, Andreas Weingessel, Christian Buchta, and Evgenia Dimitriadou.** *A Comparison of Several Cluster Algorithms on Artificial Binary Data Scenarios from Travel Market Segmentation*, April 1998a, Working Paper No. 7, Working Paper Series.

[2] **Sara Dolnicar, Friedrich Leisch, Gottfried Steiner, and Andreas Weingessel.** *A Comparison of Several Cluster Algorithms on Artificial Binary Data Scenarios from Travel Market Segmentation, Part II*, April 1998b, Working Paper No. 7, Working Paper Series.

[3] **Sara Dolnicar, Friedrich Leisch, and Andreas Weingessel.** *Artificial Binary Data Scenarios*, September 1998c, Working Paper No. 20, Working Paper Series.

[4] **Klaus Pötzelberger, and Helmut Strasser.** *Data Compression by unsupervised Classification*, December 1997, Report No. 10, Report Series.

[5] **Milligan, G. W. and Cooper, M. C.** *An Examination of Procedures for determining the Number of Clusters in a Data Set*, 1985, Psychometrika, 50(2), pp. 159–179.

[6] **Gottfried Steiner.** *COMPRESS – Data Compression and Neighbourhood Analysis*, April 1998, User Manual.

[7] **Gottfried Steiner.** *Data Compression and Reduction of Complexity – Theory, Algorithms and Experimental Results* September 1999, Doctoral Thesis, Vienna University of Economics and Business Administration.

[8] **Strasser Helmut.***Statistische Analyse der Wahrnehmung von Produktmarken*, September 1999, Technical Report.

[9] **Andreas Weingessel, Evgenia Dimitriadou, and Sara Dolnicar.** *An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets*, January 1999, Working Paper No. 29, Working Paper Series.

proposals std-k-means										
	3	4	5	**6**	7	8	9	10	11	12
0	0	14	0	0	1	63	5	1	5	11
1a	0	0	0	**100**	0	0	0	0	0	0
1ad	0	0	0	33	8	0	0	1	25	33
1b	0	0	1	0	0	5	11	30	26	27
1bd	0	0	0	0	2	3	7	18	37	33
2	0	0	0	**99**	0	0	0	0	0	1
2d	0	17	0	2	7	11	14	12	18	19
3a	0	1	0	5	7	5	11	20	28	23
3ad	0	0	0	0	0	0	10	23	32	35
3b	0	0	3	**96**	0	0	0	0	0	1
3bd	0	0	5	13	14	5	2	8	24	29
4d	0	0	0	**94**	0	0	0	0	1	5
7	0	0	0	**87**	1	0	0	1	5	6
8	**98**	0	0	0	0	0	0	2	0	0

certainty std-k-means										
	3	4	5	6	7	8	9	10	11	12
0	16	91	19	17	26	90	46	52	45	43
1a	0	0	0	100	7	35	40	45	41	50
1ad	0	3	9	91	49	24	21	35	54	52
1b	94	2	98	10	39	60	56	59	46	56
1bd	85	2	69	37	34	48	46	50	61	51
2	0	100	0	100	23	47	39	58	37	49
2d	0	100	20	80	57	49	63	54	50	54
3a	55	71	10	60	45	49	49	42	48	55
3ad	0	51	35	41	36	47	37	50	52	50
3b	0	0	100	100	53	34	45	36	53	48
3bd	0	1	63	47	56	44	58	41	46	47
4d	0	0	0	100	42	0	7	21	37	42
7	0	100	4	99	31	42	32	55	43	36
8	100	0	100	97	31	47	41	51	45	52

Table 3:

proposals std-Kohonen										
	3	4	5	**6**	7	8	9	10	11	12
0	0	30	0	0	0	55	0	1	3	11
1a	0	0	0	**100**	0	0	0	0	0	0
1ad	0	0	0	29	5	0	1	3	20	42
1b	0	0	0	0	4	8	22	25	28	13
1bd	0	0	0	1	1	4	6	23	32	33
2	0	0	0	**99**	0	0	0	0	0	1
2d	0	42	0	2	2	11	10	7	14	12
3a	3	11	2	3	6	11	11	11	23	19
3ad	0	0	0	0	0	0	3	19	44	34
3b	0	0	**99**	1	0	0	0	0	0	0
3bd	0	0	1	28	1	0	0	9	21	40
4d	0	0	0	**99**	0	0	0	0	0	1
7	0	0	0	**94**	0	0	0	0	2	4
8	**97**	0	0	0	0	0	0	2	1	0

certainty std-Kohonen										
	3	4	5	6	7	8	9	10	11	12
0	1	98	17	14	21	100	51	53	51	55
1a	0	0	0	100	20	29	44	46	52	45
1ad	0	54	8	92	69	15	39	41	49	57
1b	99	100	47	44	54	46	63	59	61	54
1bd	99	23	23	56	51	48	47	55	54	56
2	0	100	0	100	20	50	35	51	46	50
2d	0	100	12	88	46	56	61	45	53	46
3a	90	89	54	43	47	53	44	48	48	51
3ad	0	84	97	19	44	39	25	48	53	66
3b	0	0	100	99	29	40	47	47	47	59
3bd	0	0	76	88	34	28	30	56	46	57
4d	0	0	0	100	48	3	7	18	52	58
7	100	100	0	100	36	45	50	48	40	39
8	100	5	96	90	37	46	44	51	50	49

Table 4:

proposals std-logistic										
	3	4	5	**6**	7	8	9	10	11	12
0	0	24	0	0	0	53	1	0	7	15
1a	0	0	0	**100**	0	0	0	0	0	0
1ad	0	0	0	23	11	1	0	6	21	38
1b	0	0	0	0	3	9	14	25	27	22
1bd	0	0	0	1	1	4	7	30	38	19
2	0	0	0	**97**	0	0	0	0	0	3
2d	0	31	0	4	6	11	8	10	12	18
3a	0	1	0	5	9	4	18	20	21	22
3ad	0	0	0	0	0	0	2	22	42	34
3b	0	0	11	**88**	0	0	0	0	0	1
3bd	0	0	2	24	3	0	1	3	30	37
4d	0	0	0	**93**	0	0	0	0	1	6
7	0	0	0	**88**	0	0	2	0	4	6
8	**93**	0	0	0	0	0	0	1	3	3

certainty std-logistic										
	3	4	5	6	7	8	9	10	11	12
0	33	93	28	18	17	96	33	47	51	56
1a	0	0	0	100	11	40	38	51	48	46
1ad	0	26	22	78	65	19	27	43	47	53
1b	97	18	92	45	41	52	61	51	51	48
1bd	92	4	47	35	45	54	42	61	62	44
2	0	100	0	100	35	39	48	38	54	46
2d	0	100	21	79	60	52	54	49	47	58
3a	57	66	29	47	47	46	50	48	51	49
3ad	0	56	86	21	26	39	25	54	49	62
3b	0	0	100	100	37	36	52	50	36	57
3bd	0	0	48	85	30	41	33	49	57	56
4d	0	0	0	100	52	2	3	14	50	50
7	0	100	0	100	37	48	42	39	50	40
8	100	0	99	88	35	48	41	48	45	46

Table 5:

proposals std-Masters										
	3	4	5	**6**	7	8	9	10	11	12
0	0	19	0	0	0	63	2	0	6	10
1a	0	0	0	**100**	0	0	0	0	0	0
1ad	0	0	0	33	15	0	0	3	15	34
1b	0	0	0	0	0	5	27	24	18	26
1bd	0	0	0	2	0	4	6	21	30	37
2	0	0	0	**100**	0	0	0	0	0	0
2d	0	29	0	1	2	12	11	12	13	20
3a	1	5	0	2	3	8	21	20	21	19
3ad	0	0	0	0	0	2	1	15	42	40
3b	0	0	**99**	1	0	0	0	0	0	0
3bd	0	0	5	22	2	1	2	6	27	35
4d	0	0	0	**98**	0	0	0	0	0	2
7	0	0	0	**94**	0	0	0	0	1	5
8	**98**	0	0	0	0	0	0	0	0	2

certainty std-Masters										
	3	4	5	6	7	8	9	10	11	12
0	20	97	20	14	12	99	48	42	51	42
1a	0	0	0	100	12	33	45	44	47	45
1ad	0	43	19	81	60	29	29	39	49	59
1b	98	100	73	51	37	59	63	53	54	51
1bd	95	18	23	55	44	53	49	53	54	53
2	0	100	0	100	24	47	50	45	47	50
2d	0	100	26	74	64	58	49	53	47	50
3a	79	86	38	43	46	47	59	46	51	44
3ad	0	62	99	26	32	40	34	45	52	68
3b	0	0	100	100	33	41	47	44	42	54
3bd	0	0	69	87	26	38	36	43	53	54
4d	0	0	0	100	44	2	8	13	53	55
7	0	100	0	100	38	50	39	51	44	51
8	100	0	100	92	45	37	47	44	44	49

Table 6:

proposals cen-k-means										
	3	4	5	**6**	7	8	9	10	11	12
0	0	19	0	0	0	49	2	1	13	16
1a	0	0	0	**100**	0	0	0	0	0	0
1ad	0	0	0	47	7	0	1	2	6	37
1b	0	0	2	0	2	4	8	31	32	21
1bd	0	0	0	1	0	8	18	13	26	34
2	0	0	0	**99**	0	0	0	0	0	1
2d	0	27	0	1	5	11	12	10	14	20
3a	0	0	0	7	5	8	15	18	23	24
3ad	0	0	0	0	0	0	5	26	26	43
3b	0	0	0	**96**	0	0	0	0	1	3
3bd	0	0	6	12	10	8	3	8	27	26
4d	0	0	0	**96**	0	0	0	0	1	3
7	0	0	0	**95**	0	0	0	0	1	4
8	**91**	0	0	0	0	0	0	0	4	5

certainty cen-k-means										
	3	4	5	6	7	8	9	10	11	12
0	8	93	27	24	22	95	45	44	55	45
1a	0	0	0	100	24	27	42	45	50	46
1ad	0	5	15	85	50	30	27	41	42	55
1b	0	0	100	14	63	59	59	57	56	46
1bd	74	6	35	45	43	41	50	52	50	56
2	0	100	0	100	24	46	37	53	40	39
2d	0	100	15	84	45	65	50	45	52	47
3a	58	52	17	55	42	50	53	45	47	50
3ad	0	46	42	38	32	43	48	57	40	60
3b	0	0	100	100	42	45	42	43	49	41
3bd	0	5	54	55	51	51	45	55	50	44
4d	0	0	0	100	41	0	8	17	36	35
7	0	100	0	100	48	40	42	41	49	43
8	100	0	100	99	27	39	47	40	50	49

Table 7:

proposals cen-Kohonen										
	3	4	5	**6**	7	8	9	10	11	12
0	0	33	0	0	0	54	1	0	1	11
1a	0	0	0	**100**	0	0	0	0	0	0
1ad	0	0	0	27	10	0	0	2	22	39
1b	0	0	0	0	0	4	22	28	27	19
1bd	0	0	0	1	6	4	11	16	33	29
2	0	0	0	**99**	0	0	0	0	1	0
2d	0	45	0	0	0	10	12	11	16	6
3a	6	1	3	3	6	7	15	17	20	22
3ad	0	0	0	0	0	0	0	18	45	37
3b	0	0	**100**	0	0	0	0	0	0	0
3bd	0	0	6	27	4	1	0	6	29	27
4d	0	0	0	**98**	0	0	0	0	0	2
7	0	0	0	**97**	0	0	0	0	0	3
8	**99**	0	0	0	0	0	0	0	0	1

certainty cen-Kohonen										
	3	4	5	6	7	8	9	10	11	12
0	4	99	16	15	16	98	51	41	51	53
1a	0	0	0	100	17	27	46	40	44	50
1ad	0	66	15	85	69	24	29	40	49	60
1b	97	100	70	37	51	50	64	53	58	45
1bd	96	6	18	57	50	53	52	53	63	50
2	0	100	0	100	20	48	40	54	42	51
2d	0	100	19	81	56	56	50	54	56	38
3a	91	83	49	43	47	52	48	50	56	43
3ad	0	90	97	17	32	40	29	47	50	63
3b	0	0	100	99	33	39	49	41	44	51
3bd	0	0	70	81	43	37	33	52	65	43
4d	0	0	0	100	50	1	5	20	54	61
7	0	100	0	100	44	42	44	44	52	48
8	100	1	99	90	47	44	49	38	55	44

Table 8:

recoverability of binarised types std-k-means														
	#	0	1	2	3	4	5	6	t_1	t_2	t_3	t_4	t_5	t_6
1a	5	0	0	0	0	45	55	0	15	95	50	100	99	96
1a	6	0	0	0	0	0	0	100	100	100	100	100	100	100
1a	7	0	0	0	0	0	0	100	100	100	100	100	100	100
1ad	5	0	0	0	0	100	0	0	76	58	77	61	66	62
1ad	6	0	0	0	0	10	7	83	95	94	99	94	95	96
1ad	7	0	0	0	0	1	2	97	100	100	100	97	99	100
1b	4	0	0	0	28	72	0	0	100	0	72	100	28	72
1b	5	0	0	0	0	1	99	0	100	0	100	100	99	100
1b	6	0	0	0	0	0	56	44	100	44	100	100	100	100
1b	7	0	0	0	0	0	19	81	100	81	100	100	100	100
1bd	4	0	0	48	38	14	0	0	100	0	23	100	10	33
1bd	5	0	0	0	24	60	16	0	99	1	100	100	44	48
1bd	6	0	0	0	23	48	26	3	100	15	100	100	49	45
1bd	7	0	0	0	7	38	44	11	100	41	99	100	61	58
2	5	0	0	0	0	35	65	0	100	100	99	5	100	61
2	6	0	0	0	0	0	0	100	100	100	100	100	100	100
2	7	0	0	0	0	0	0	100	100	100	100	100	100	100
2d	5	0	0	1	0	99	0	0	100	100	94	5	94	5
2d	6	0	0	0	0	18	0	82	100	100	93	89	93	89
2d	7	0	0	0	0	3	0	97	100	100	99	98	99	98
3b	5	0	0	0	0	0	100	0	100	0	100	100	100	100
3b	6	0	0	0	0	0	0	100	100	100	100	100	100	100
3b	7	0	0	0	0	0	0	100	100	100	100	100	100	100
3bd	5	0	0	0	0	0	100	0	100	100	97	60	48	95
3bd	6	0	0	0	0	0	55	45	100	100	93	85	72	95
3bd	7	0	0	0	0	0	26	74	100	100	95	97	84	98
4d	5	0	0	0	0	100	0	0	69	87	51	70	73	50
4d	6	0	0	0	0	0	0	100	100	100	100	100	100	100
4d	7	0	0	0	0	0	0	100	100	100	100	100	100	100
7	5	0	0	0	0	0	100	0	100	100	100	0	100	100
7	6	0	0	0	0	0	0	100	100	100	100	100	100	100
7	7	0	0	0	0	0	0	100	100	100	100	100	100	100
8	5	0	0	0	0	75	25	0	98	22	5	100	100	100
8	6	0	0	0	0	1	3	96	98	98	99	100	100	100
8	7	0	0	0	0	0	1	99	100	99	100	100	100	100

Table 9:

recoverability of binarised types std-Kohonen														
	#	0	1	2	3	4	5	6	t_1	t_2	t_3	t_4	t_5	t_6
1a	5	0	0	0	0	16	84	0	33	98	54	100	100	99
1a	6	0	0	0	0	0	0	100	100	100	100	100	100	100
1a	7	0	0	0	0	0	0	100	100	100	100	100	100	100
1ad	5	0	0	0	0	99	1	0	78	49	85	58	80	51
1ad	6	0	0	0	0	11	3	86	94	95	97	96	96	97
1ad	7	0	0	0	0	1	0	99	99	100	100	100	99	100
1b	4	0	0	0	100	0	0	0	100	0	0	100	94	6
1b	5	0	0	0	1	6	93	0	100	0	99	100	93	100
1b	6	0	0	0	0	0	88	12	100	12	100	100	100	100
1b	7	0	0	0	0	0	52	48	100	48	100	100	100	100
1bd	4	0	0	45	55	0	0	0	100	0	3	100	13	39
1bd	5	0	0	3	34	49	14	0	100	0	93	100	37	44
1bd	6	0	0	0	47	41	12	0	100	2	99	100	29	35
1bd	7	0	0	0	25	51	23	1	100	12	99	100	36	53
2	5	0	0	0	0	0	100	0	100	100	99	1	100	100
2	6	0	0	0	0	0	0	100	100	100	100	100	100	100
2	7	0	0	0	0	0	0	100	100	100	100	100	100	100
2d	5	0	0	0	0	100	0	0	100	100	97	3	97	3
2d	6	0	0	0	0	18	0	82	100	100	96	86	96	86
2d	7	0	0	0	0	0	0	100	100	100	100	100	100	100
3b	5	0	0	0	0	0	100	0	100	0	100	100	100	100
3b	6	0	0	0	0	0	0	100	100	100	100	100	100	100
3b	7	0	0	0	0	0	0	100	100	100	100	100	100	100
3bd	5	0	0	0	0	0	100	0	100	100	98	91	24	87
3bd	6	0	0	0	0	0	14	86	100	100	97	99	91	99
3bd	7	0	0	0	0	0	3	97	100	100	100	100	98	99
4d	5	0	0	0	0	100	0	0	69	80	32	89	95	35
4d	6	0	0	0	0	0	0	100	100	100	100	100	100	100
4d	7	0	0	0	0	0	0	100	100	100	100	100	100	100
7	5	0	0	0	0	0	100	0	100	100	100	0	100	100
7	6	0	0	0	0	0	0	100	100	100	100	100	100	100
7	7	0	0	0	0	0	0	100	100	100	100	100	100	100
8	5	0	0	0	0	81	19	0	100	0	19	100	100	100
8	6	0	0	0	0	6	5	89	99	90	94	100	100	100
8	7	0	0	0	0	0	3	97	100	97	100	100	100	100

Table 10:

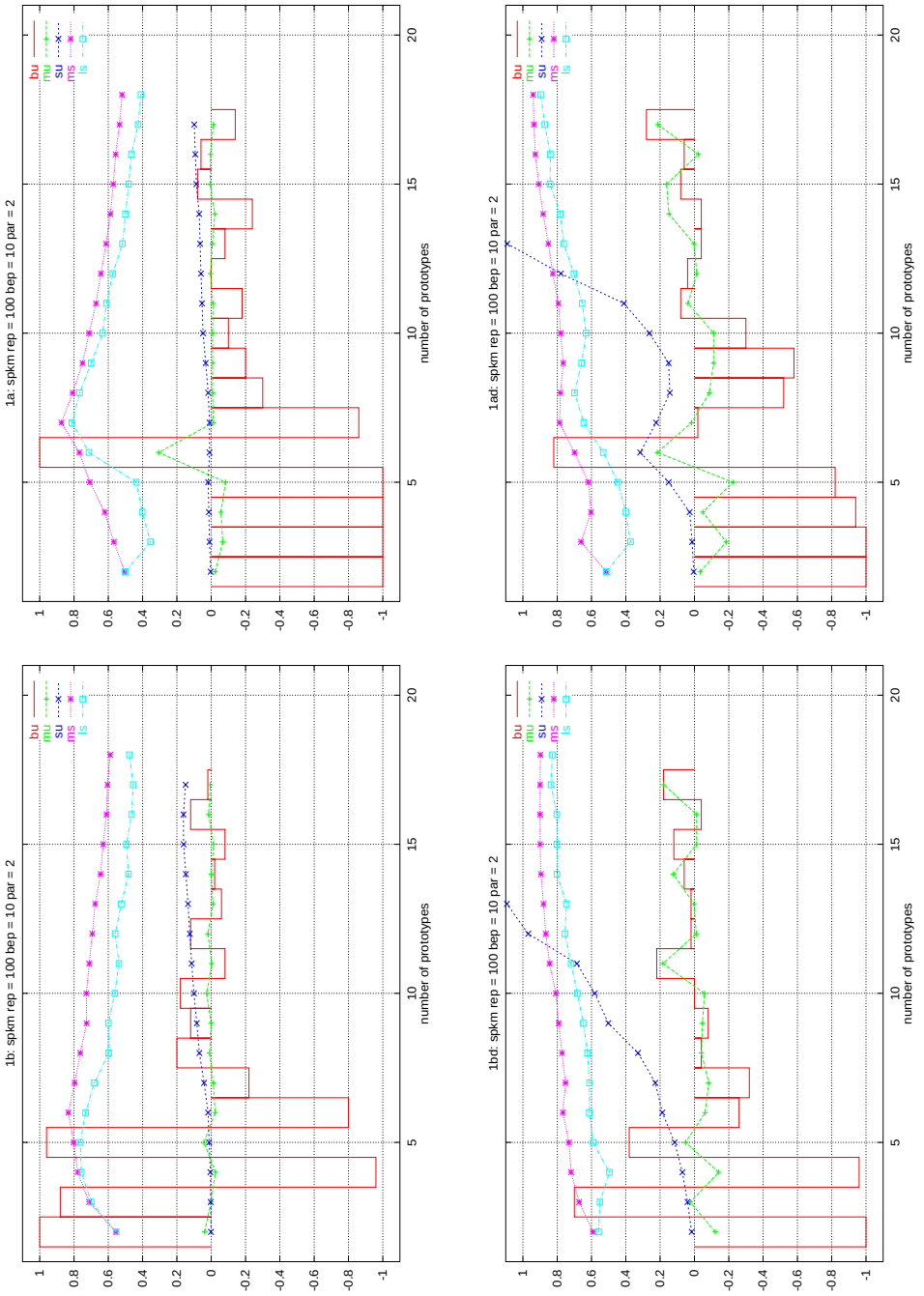
1a: spkm rep = 100 bep = 10 par = 2
1ad: spkm rep = 100 bep = 10 par = 2
1b: spkm rep = 100 bep = 10 par = 2
1bd: spkm rep = 100 bep = 10 par = 2
bu
mu
nu
su
ms
ls
number of prototypes

Figure 1:

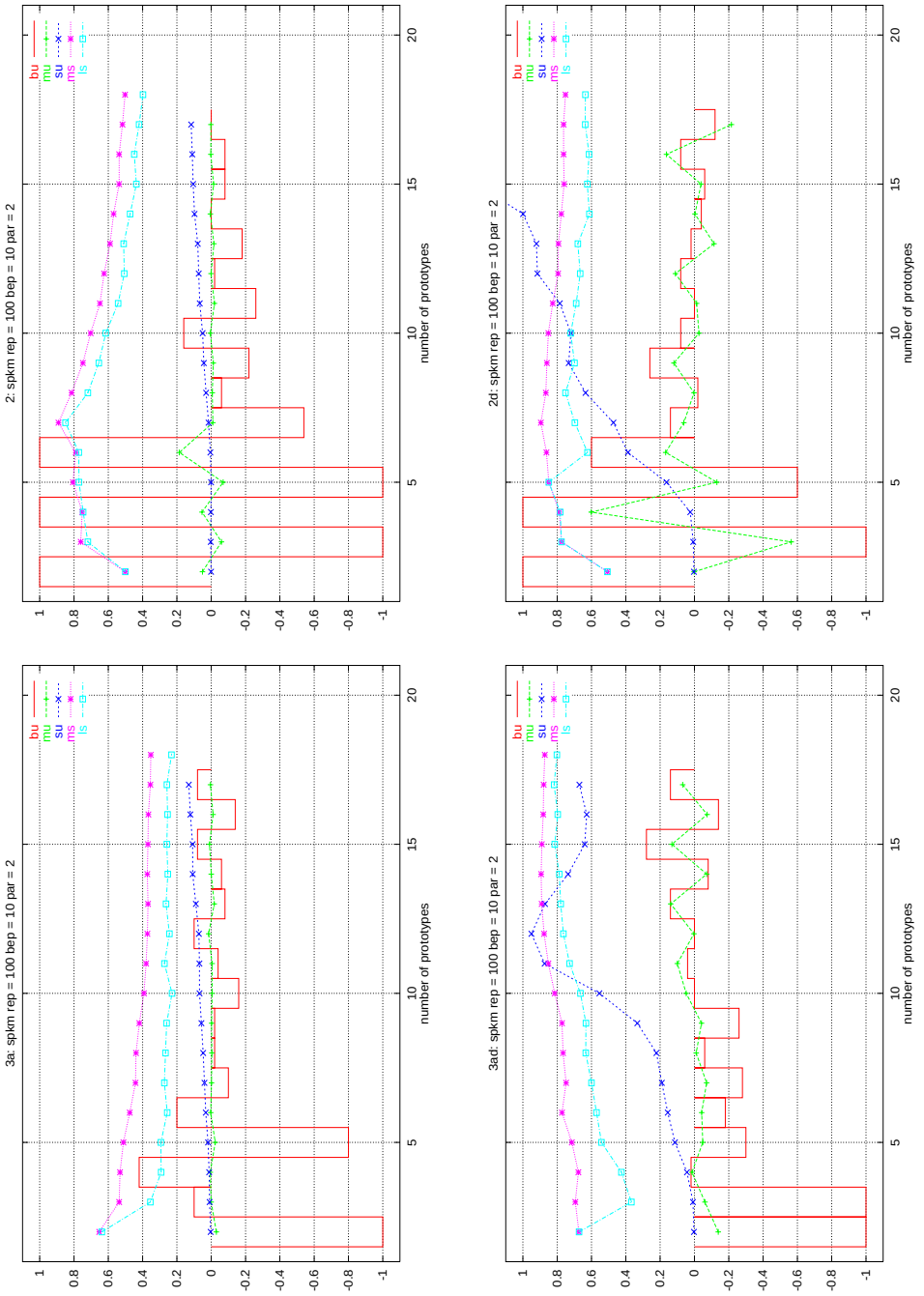
2: spkm rep = 100 bep = 10 par = 2
2d: spkm rep = 100 bep = 10 par = 2
3a: spkm rep = 100 bep = 10 par = 2
3ad: spkm rep = 100 bep = 10 par = 2
number of prototypes
bu
mu
nu
su
ms
ls

Figure 2:

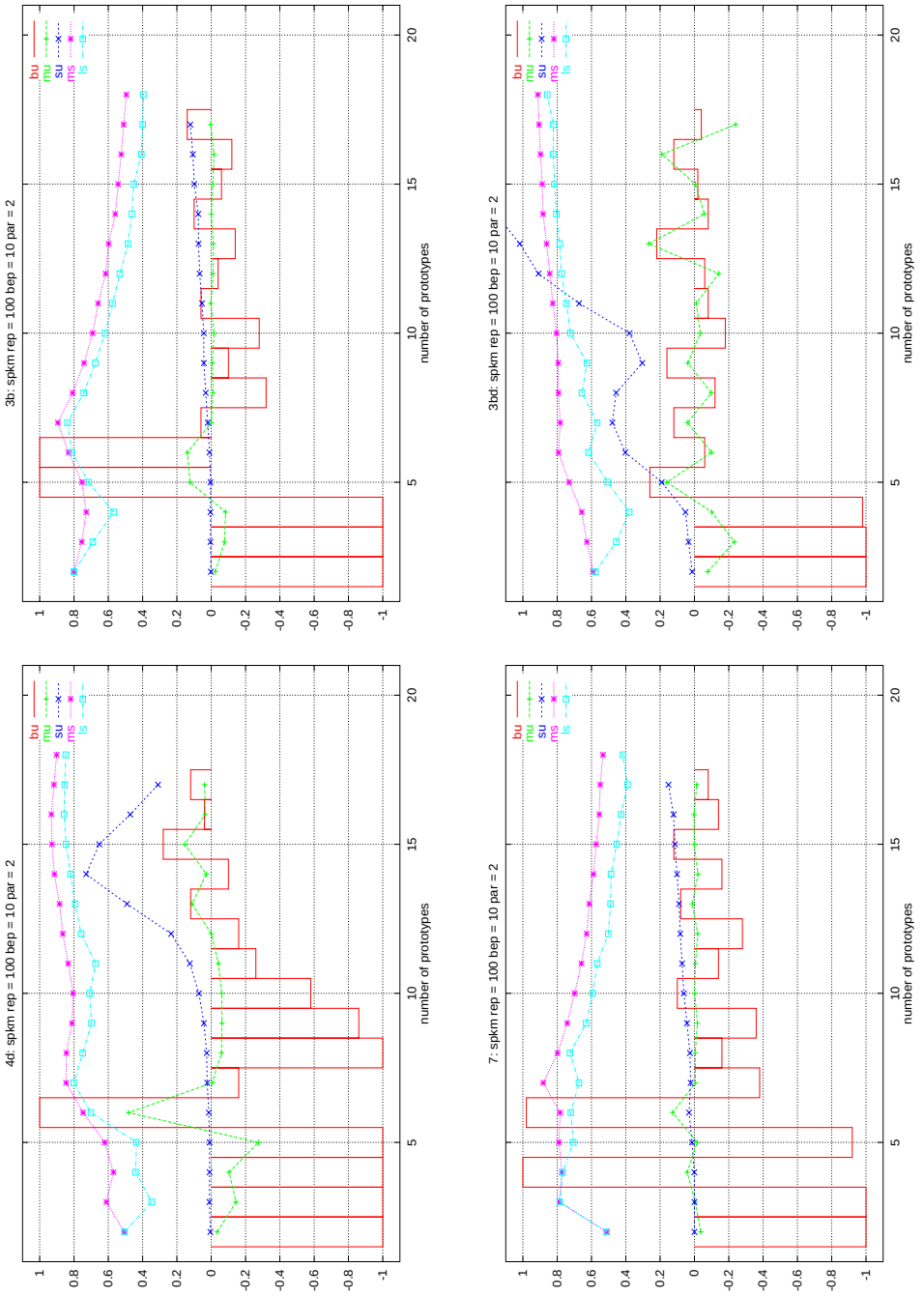
3b: spkm rep = 100 bep = 10 par = 2
3bd: spkm rep = 100 bep = 10 par = 2
4d: spkm rep = 100 bep = 10 par = 2
7: spkm rep = 100 bep = 10 par = 2
bu
mu
nu
su
ms
ls
number of prototypes

Figure 3:

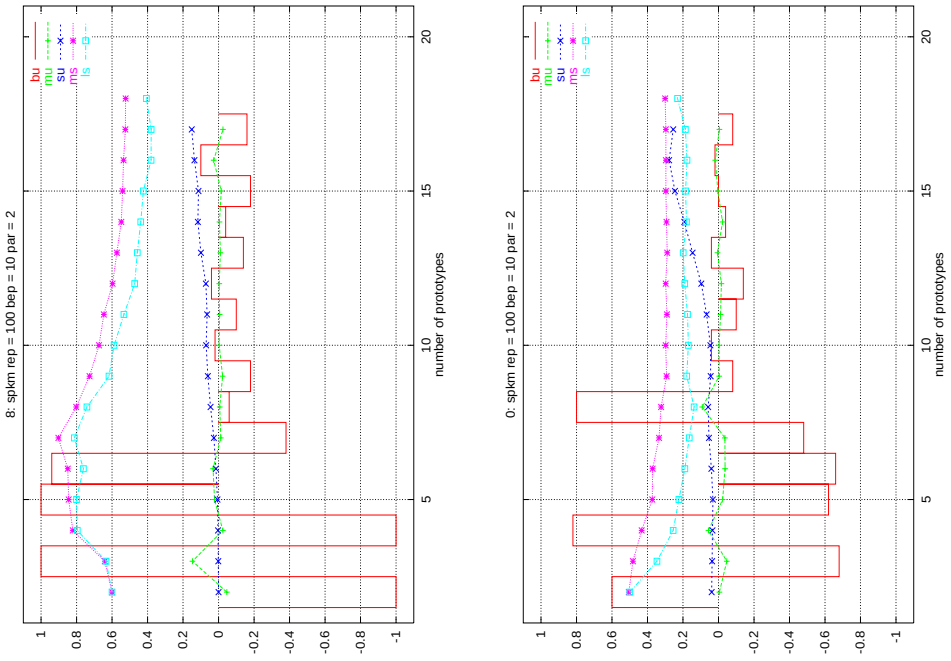
8: spkm rep = 100 bep = 10 par = 2
0: spkm rep = 100 bep = 10 par = 2
bu
mu
su
ms
ls
number of prototypes
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
5
10
15
20

Figure 4:

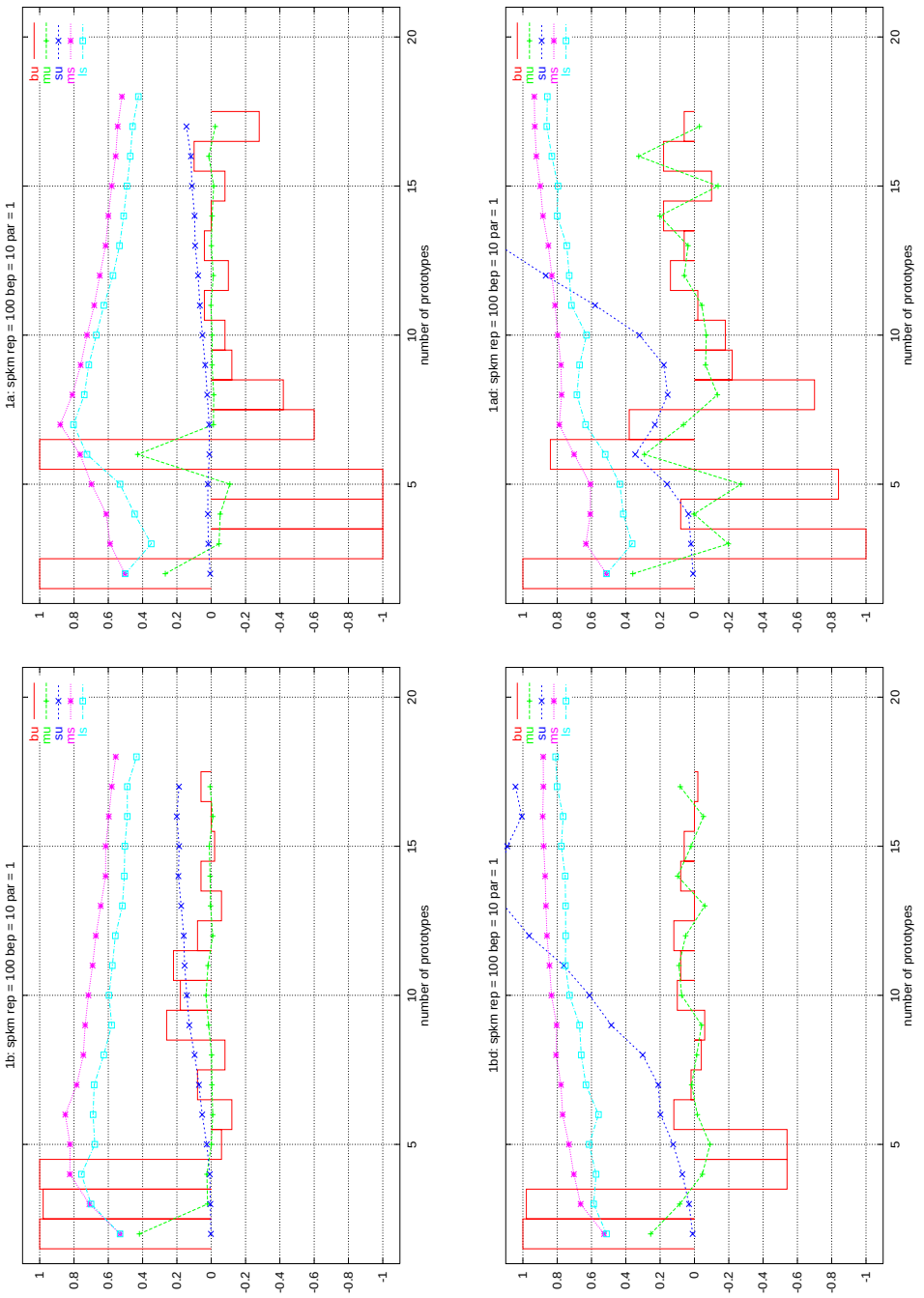
1a: spkm rep = 100 bep = 10 par = 1
1ad: spkm rep = 100 bep = 10 par = 1
1b: spkm rep = 100 bep = 10 par = 1
1bd: spkm rep = 100 bep = 10 par = 1
bu
mu
nu
su
ms
ls
number of prototypes

Figure 5:

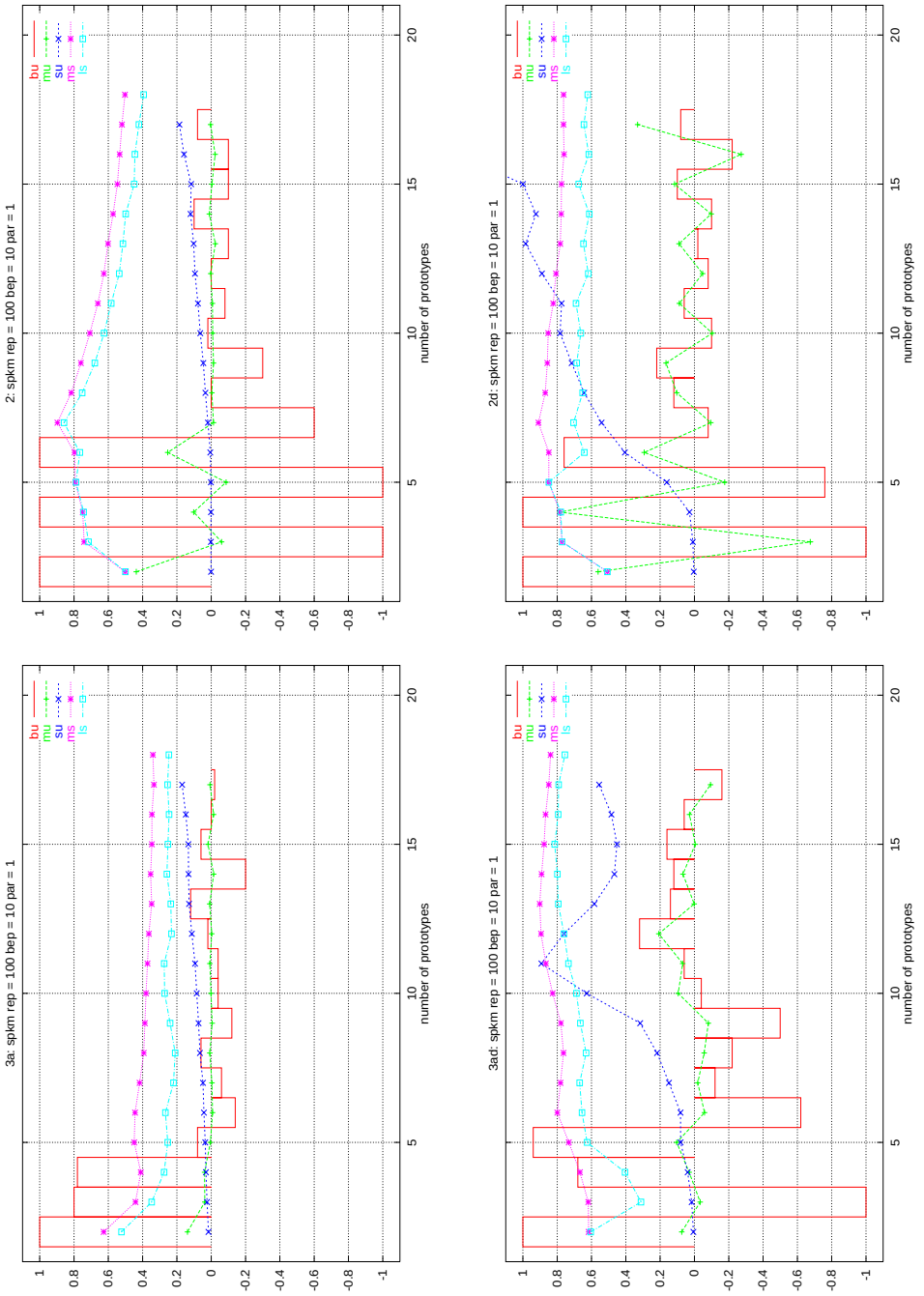
2: spkm rep = 100 bep = 10 par = 1
2d: spkm rep = 100 bep = 10 par = 1
3a: spkm rep = 100 bep = 10 par = 1
3ad: spkm rep = 100 bep = 10 par = 1
bu
mu
nu
su
ms
ls
number of prototypes

Figure 6:

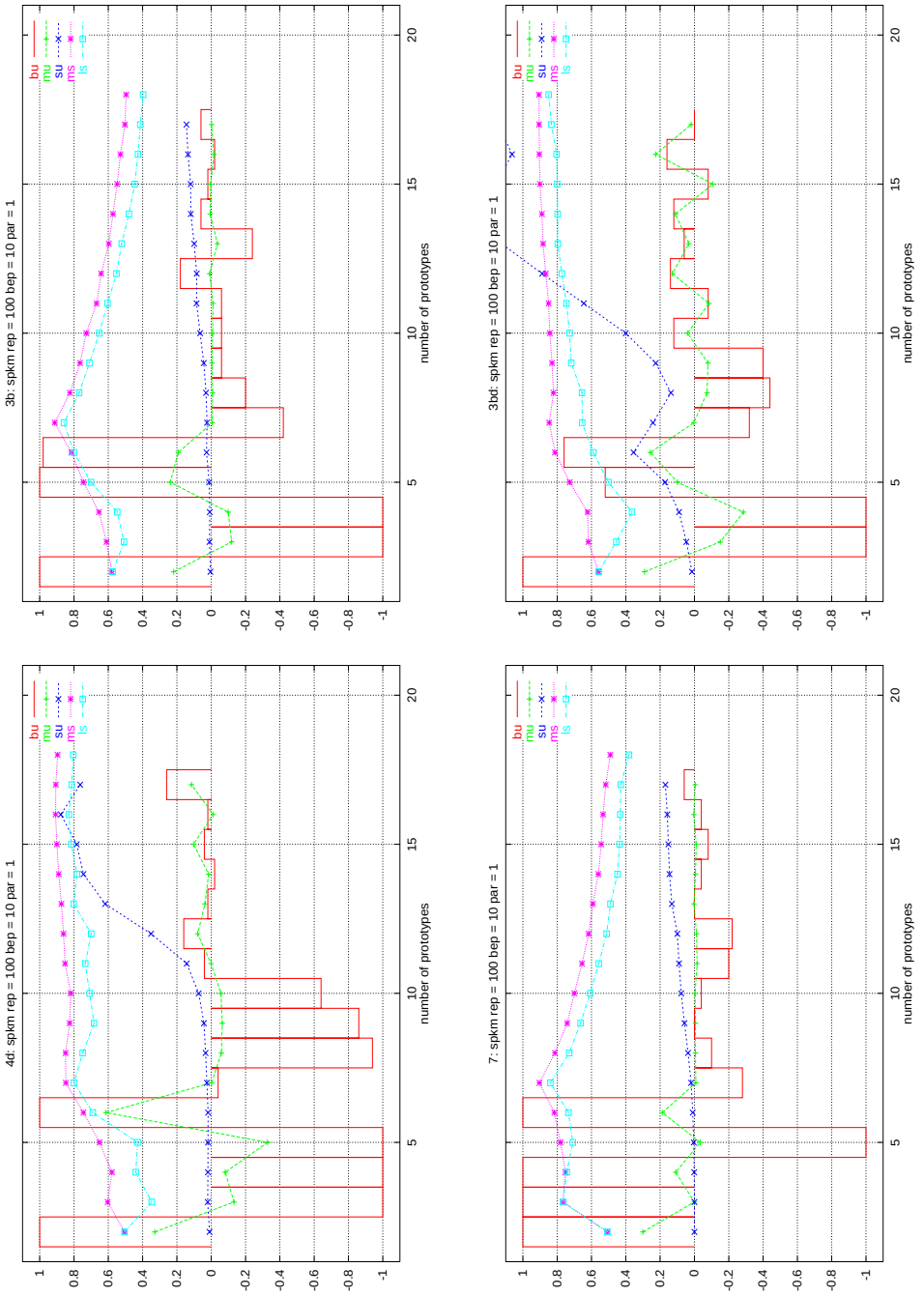
3b: spkm rep = 100 bep = 10 par = 1
3bd: spkm rep = 100 bep = 10 par = 1
4d: spkm rep = 100 bep = 10 par = 1
7: spkm rep = 100 bep = 10 par = 1
bu
mu
nu
su
ms
ls
number of prototypes

Figure 7:

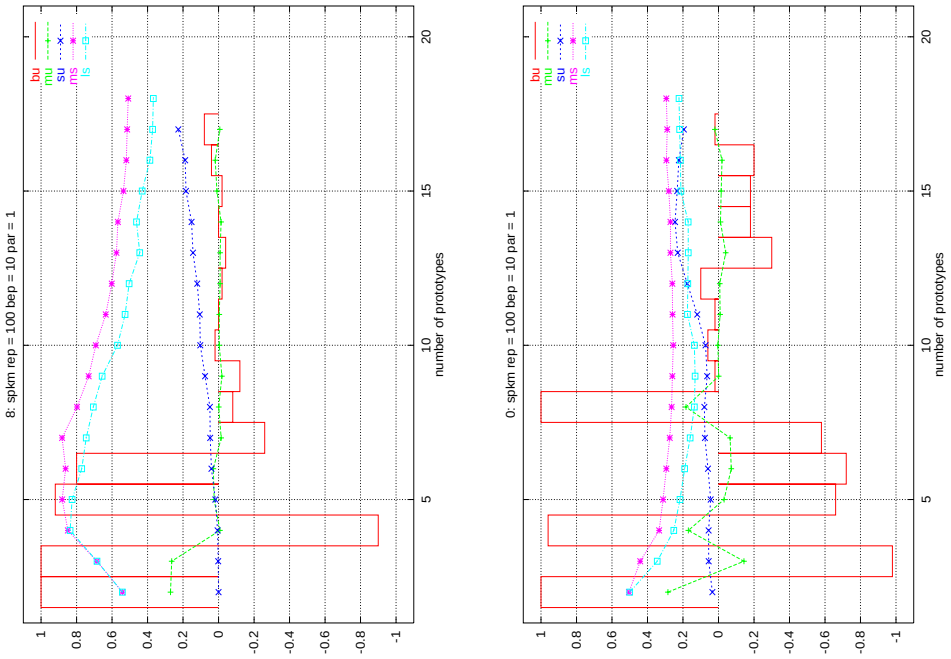
bu
mu
su
ms
ls
8: spkm rep = 100 bep = 10 par = 1
number of prototypes
0: spkm rep = 100 bep = 10 par = 1
number of prototypes

Figure 8:

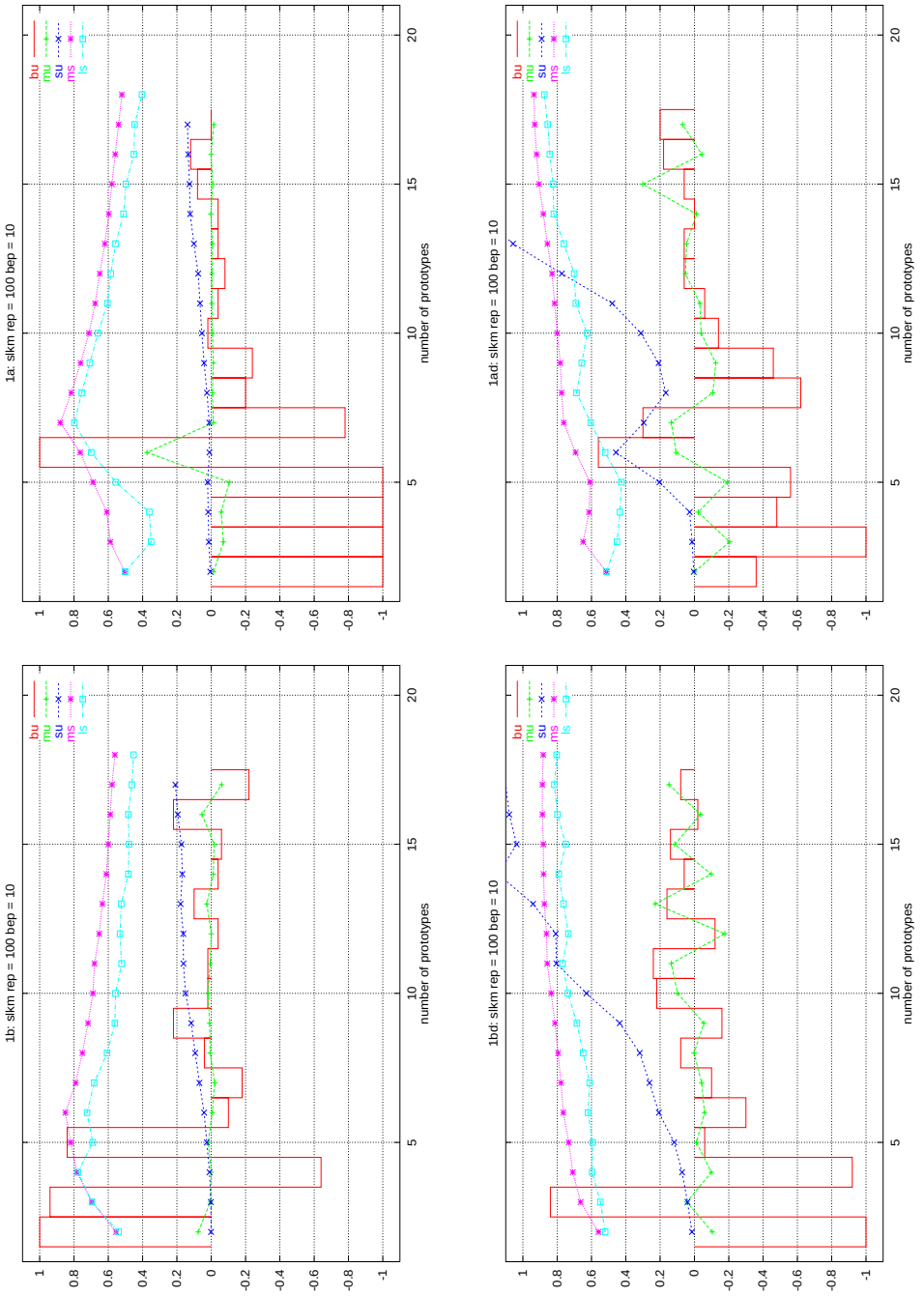
1a: skm rep = 100 bep = 10
1ad: skm rep = 100 bep = 10
1b: skm rep = 100 bep = 10
1bd: skm rep = 100 bep = 10
bu
mu
nu
su
ms
ls
number of prototypes

Figure 9:

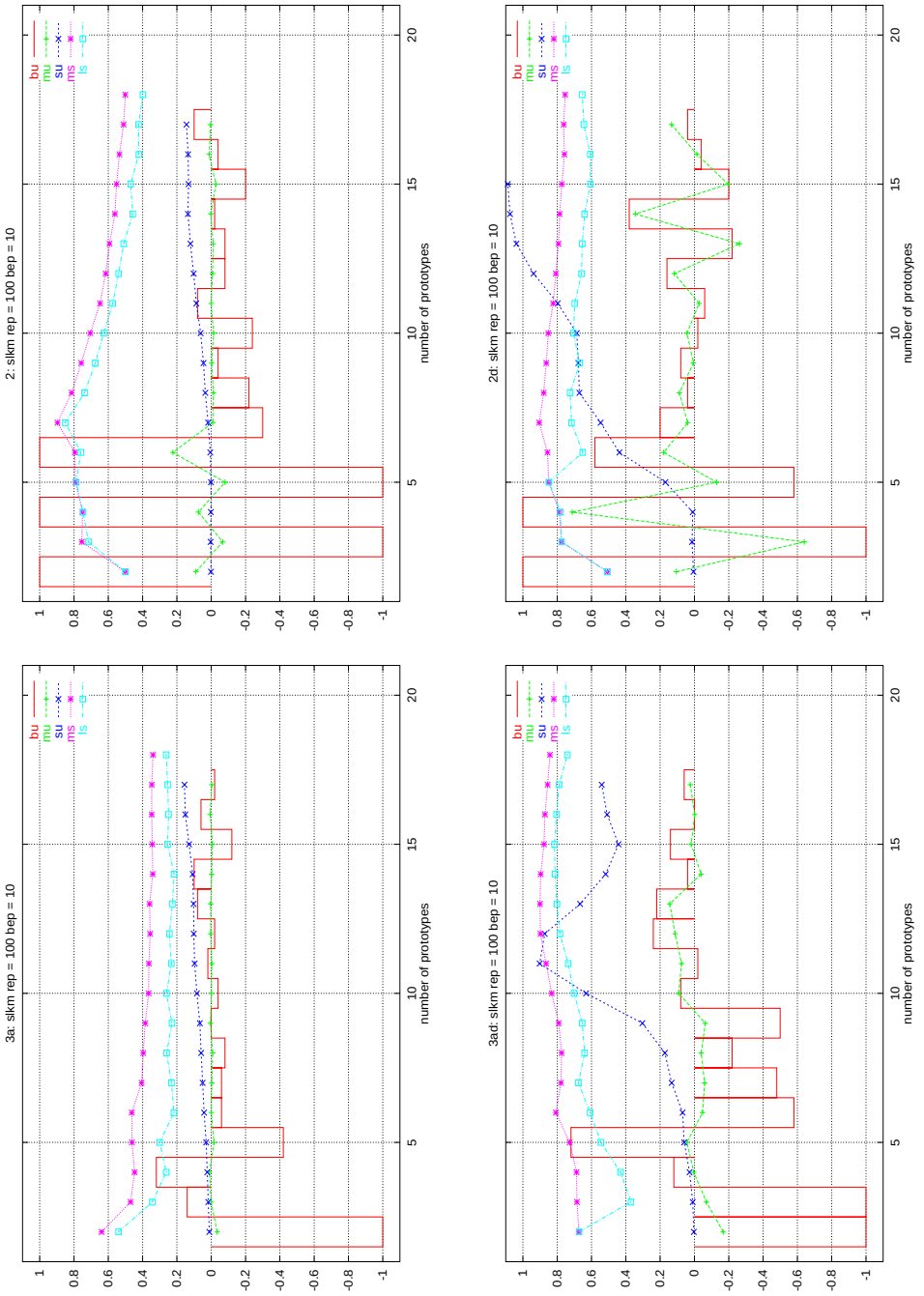
2: slkm rep = 100 bep = 10
2d: slkm rep = 100 bep = 10
3a: slkm rep = 100 bep = 10
3ad: slkm rep = 100 bep = 10
number of prototypes
bu
mu
nu
su
ms
ls

Figure 10:

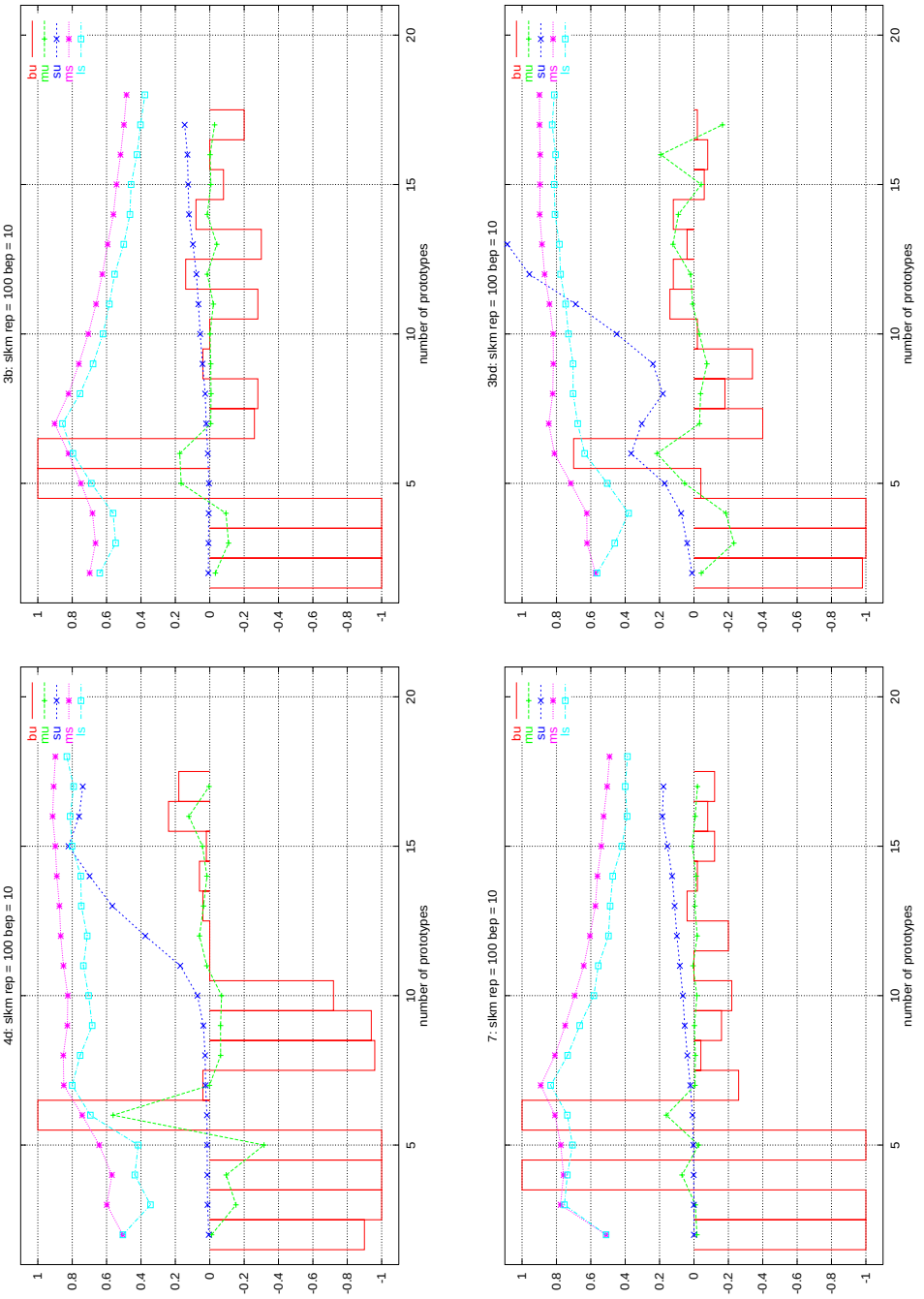
3b: slkm rep = 100 bep = 10
3bd: slkm rep = 100 bep = 10
4d: slkm rep = 100 bep = 10
7: slkm rep = 100 bep = 10
bu
mu
nu
su
ms
ls
number of prototypes

Figure 11:

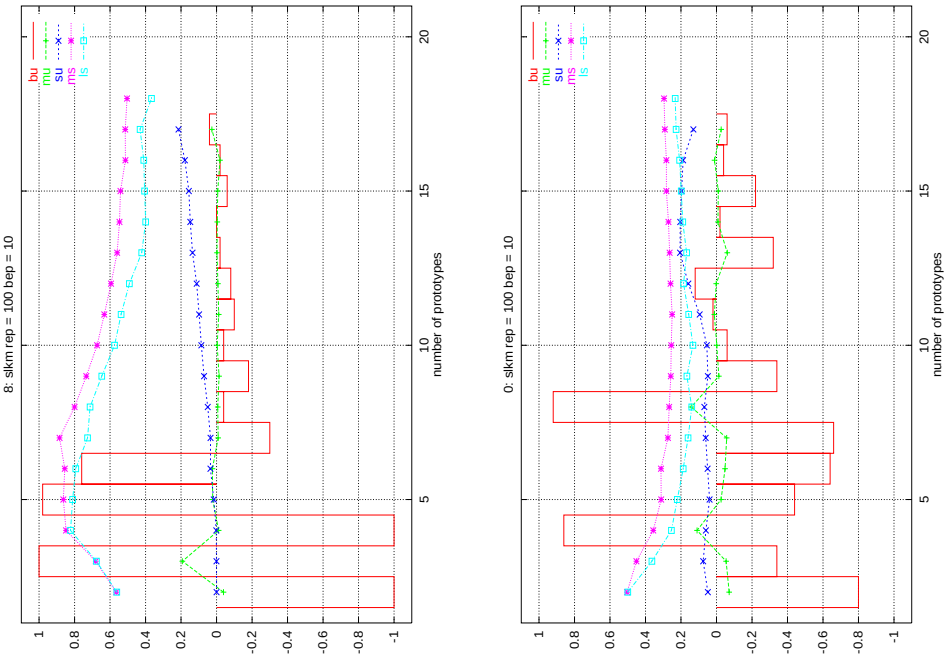
8: slkm rep = 100 bep = 10
0: slkm rep = 100 bep = 10
bu
mu
su
ms
ls
number of prototypes
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
5
10
15
20

Figure 12:

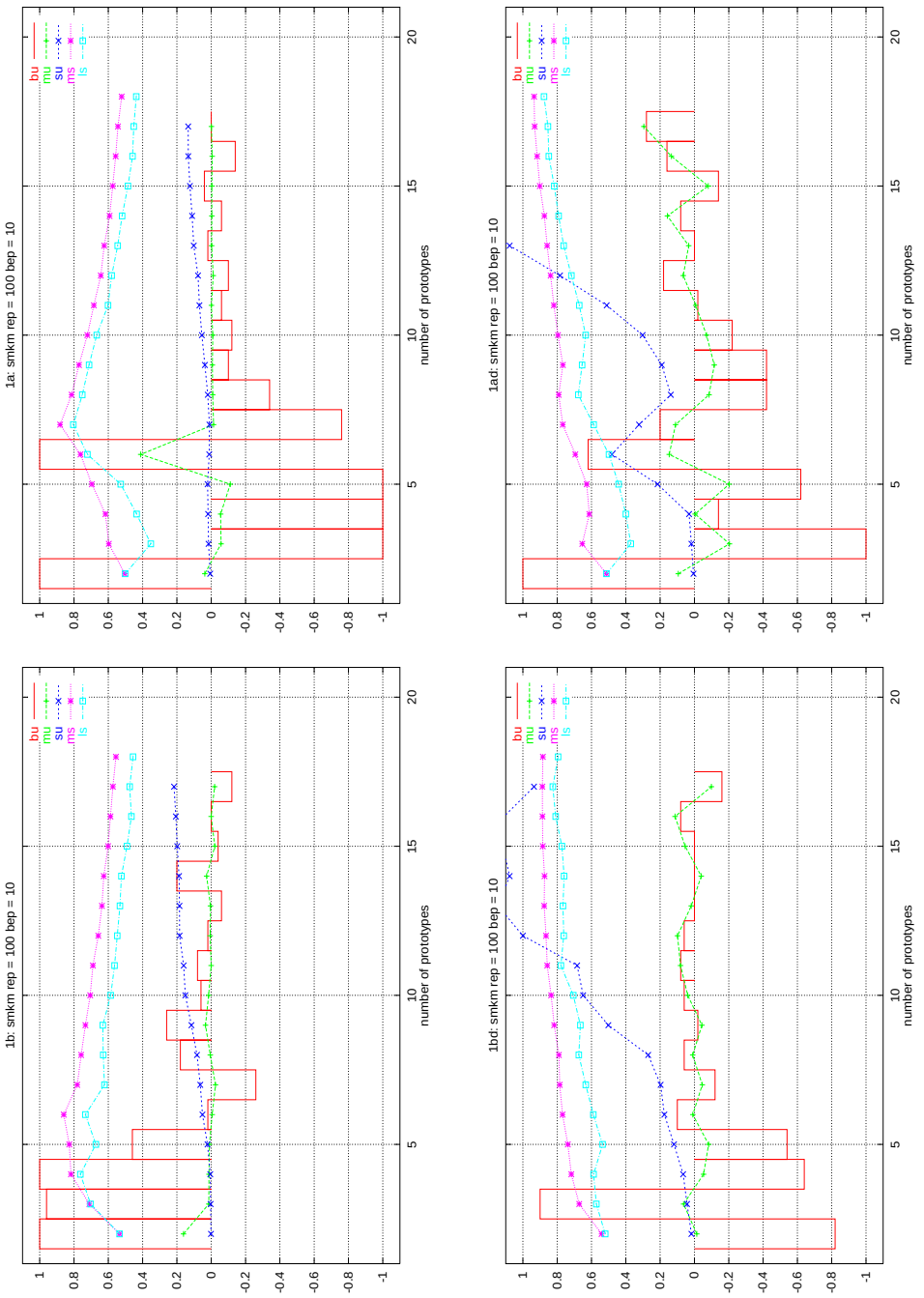
1a: smkm rep = 100 bep = 10
1ad: smkm rep = 100 bep = 10
1b: smkm rep = 100 bep = 10
1bd: smkm rep = 100 bep = 10
bu
mu
nu
su
ms
ls
number of prototypes

Figure 13:

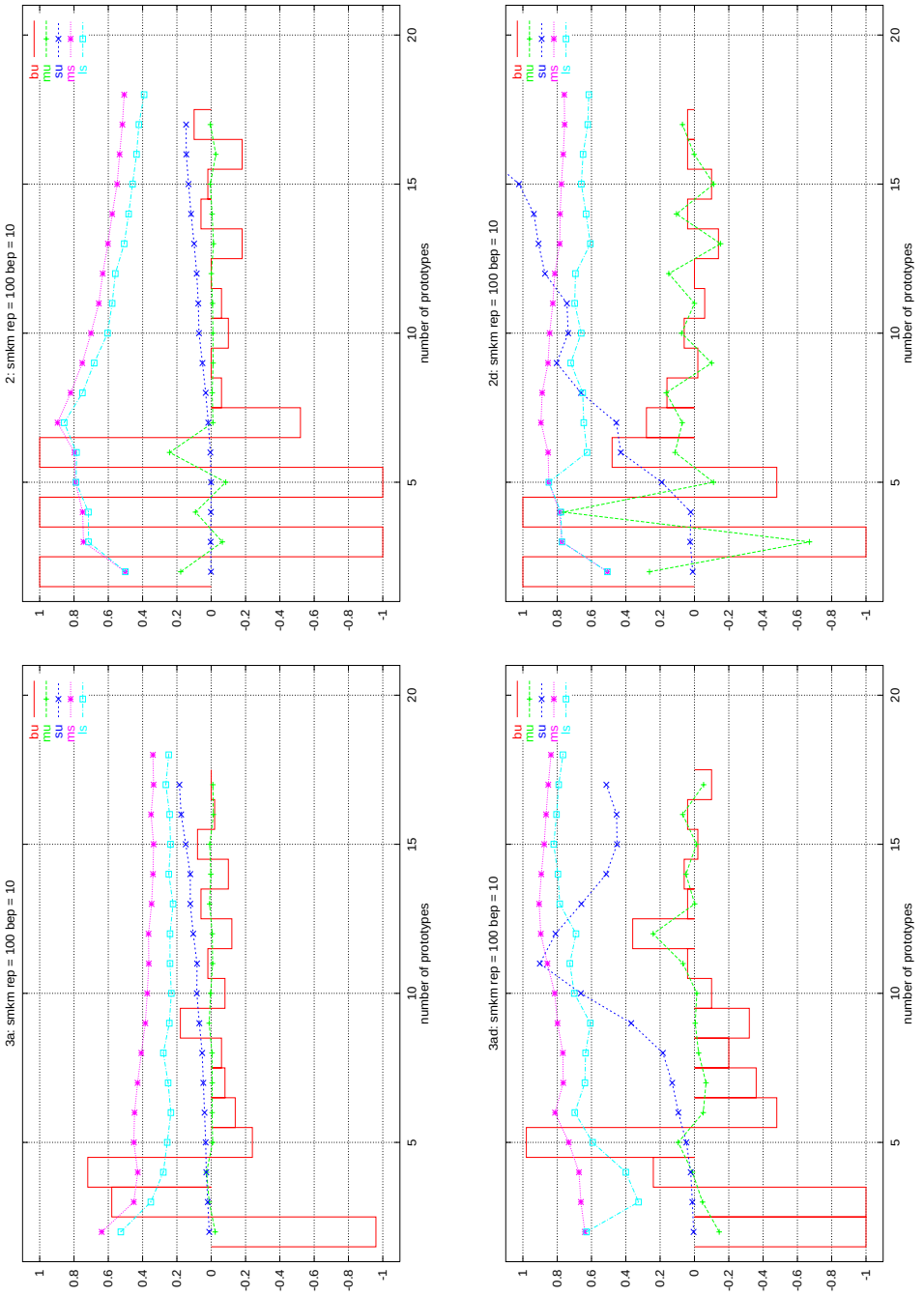
2: smkm rep = 100 bep = 10
2d: smkm rep = 100 bep = 10
3a: smkm rep = 100 bep = 10
3ad: smkm rep = 100 bep = 10
bu
mu
nu
su
ms
ls
number of prototypes

Figure 14:

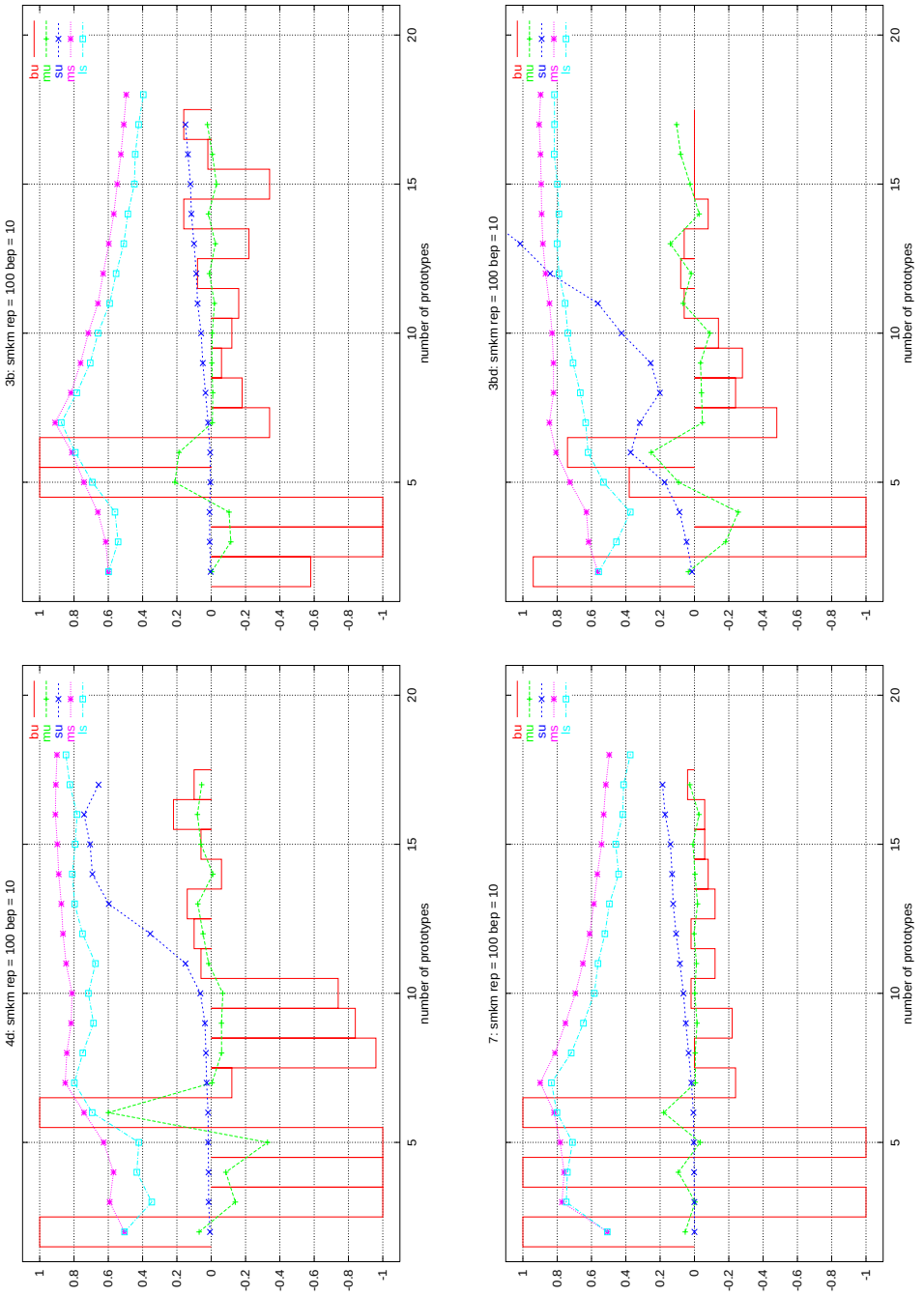
3b: smkm rep = 100 bep = 10
3bd: smkm rep = 100 bep = 10
4d: smkm rep = 100 bep = 10
7: smkm rep = 100 bep = 10
bu
mu
nu
su
ms
ls
number of prototypes

Figure 15:

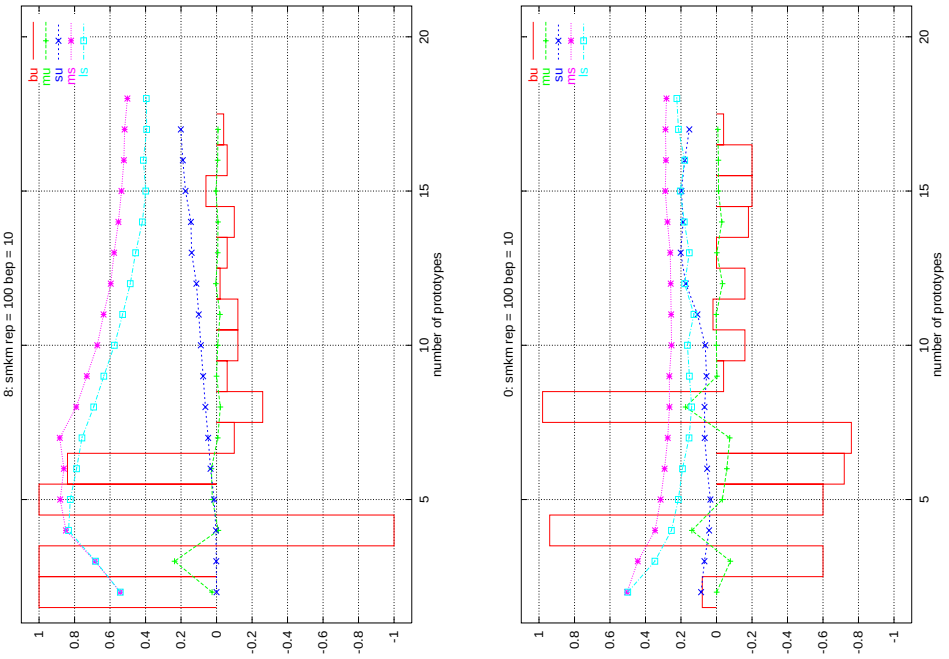
8: smkm rep = 100 bep = 10
0: smkm rep = 100 bep = 10
bu
mu
su
ms
ls
number of prototypes
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
5
10
15
20

Figure 16: