

Clustering permutations by Exponential Blurring

Mean-Shift algorithm

Le Bao and Marina Meilă

Department of Statistics

University of Washington

Seattle, WA 98195-4322

`{lebao,mmp}@stat.washington.edu`

UW Statistics Department Technical Report 524

February 14, 2008

Abstract

Suppose that a sample of people independently examine a fixed set of k items and then rank these items according to personal judgment. Whatever the nature of these items, each person produces a ranking. This paper aims at clustering people into different groups according to their preferences. We propose the exponential blurring mean-shift (EBMS) algorithm which shifts the rankings to new locations obtained by a locally weighted combination of all the data. The number of clusters does not need to be specified in advance and outliers can be detected. Our experiments show that the EBMS algorithm can be successfully applied in clustering the ranking data. The algorithm generalizes to partial orderings when only the *top-t* ranks are observed.

1 Metrics and Probability Models for Ranked Data

In this section, we introduce the notation for permutation, provide characterizations of metrics on permutations, and relate them to Mallows' model.

1.1 Metrics for Ranked Data

A full ranking of k items is simply an ordering of all these items, of the form: first choice, second choice, ... , k th choice. Any such ranking can be viewed as an element π from the permutation group S_k . Here S_k is the set of all one-to-one functions from $1, 2, \dots, k$ onto itself. In other words, we have the following convention: $\pi(i)$ is the rank given to item i ; $\pi^{-1}(i)$ is the item assigned the rank i . It will be convenient to introduce the preference matrix Q corresponding to π^{-1} , in which $Q_{ij} = 1$ if and only if j precedes i in π^{-1} . Therefore the rank of i th item is simply the sum of the i th row of the permutation matrix Q plus 1.

Now suppose there are two judges, who each rank the same k items. Let π and σ be the permutations corresponding to the two judges' rankings. Then a metric $d(\pi, \sigma)$ can be thought of as a measure of the distance between two rankings, and a proper metric satisfies: $d(\pi, \pi) = 0$; $d(\pi, \sigma) > 0$ if $\pi \neq \sigma$; $d(\pi, \sigma) = d(\sigma, \pi)$; $d(\pi, \sigma) \leq d(\pi, \tau) + d(\tau, \sigma)$ (see Diaconis 1982 for a review of various ranking metrics). In this project, we have chosen to work with Kendall's τ distance which has been studied extensively in the statistical literature.

Kendall's τ distance is defined as the number of pairs of items, (i, j) , such that $\pi(i) < \pi(j)$ and $\sigma(i) > \sigma(j)$ (Kendall 1938). It is also equivalent to the number of swaps that the bubble sort algorithm would make to place one list in the same order as the other list, and has $O(k^2)$ complexity. Here we suggest an approach of calculation based on the permutation matrix: $D_K = \sum_{i=1}^k \sum_{j=1}^k |Q_{ij}(\pi) - Q_{ij}(\sigma)|/2$. Another common distance, Spear-

man's footrule is define as $F(\pi, \sigma) = \sum_{i=1}^k |\pi(i) - \sigma(i)|$, which can also be expressed in matrix form: $F(\pi, \sigma) = \sum_{i=1}^k |\sum_{j=1}^k Q_{ij}(\pi) - \sum_{j=1}^k Q_{ij}(\sigma)|$. We can show that Kendall distance is bounded between half of the Spearman's footrule and original Spearman's footrule: $F(\pi, \sigma)/2 \leq D_K(\pi, \sigma) \leq F(\pi, \sigma)$.

For instance, the permutation matrices for $\pi^{-1} = (4, 1, 2, 3)$ and $\sigma^{-1} = (2, 1, 4, 3)$ are

$$Q(\pi) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad Q(\sigma) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \quad (1)$$

Then we have $\pi = (2, 3, 4, 1)$, $\sigma = (2, 1, 4, 3)$. The Kendall's τ distance between π^{-1} and σ^{-1} is $6 \div 2 = 3$, and the Spearman's footrule is $0 + 2 + 0 + 2 = 4$.

1.2 Probability Model for Ranked Data

A class of models over rankings is proposed for which the probability of a ranking decreases with increasing distance from a modal ordering, e.g. the Mallows model uses a concentration parameter to reflect the variability of the ranking about the modal ordering. Under the Mallows' model, the judges' rankings are assumed to be generated according to probability: $P_\theta(\pi) = \frac{\exp^{-\theta d(\pi, \pi_0)}}{\psi(\theta)}$, where π_0 is a fixed ranking and considered as the centroid, θ is the spread parameter, and $\psi(\theta)$ is the normalizing constant which does not depend on π_0 . When $\theta = 0$, the Mallows' model is the uniform distribution, and when approaches infinity P_θ becomes concentrated at the single ranking π_0 .

We say the metric d is left invariant if $d(\pi_1, \pi_2) = d(\pi_1 \cdot \pi_2^{-1}, e)$, where e denotes the identical

permutation $e = (1, 2, \dots, k)$. Therefore the distance is completely determined by the function $D(\pi) = d(\pi, e)$. It results in the following decomposition $D_K(\pi) = \sum_{j=1}^{k-1} V_j(\pi)$, where $V_j(\pi)$ is the number of items in $j+1 : k$ that are ranked before j by π , or in the matrix form $V_j = \sum_{i=j+1}^k Q_{ij}$. Fligner and Verducci (1986) proposed a $k-1$ parameters generalization of the Mallows' model: $D_\theta(\pi) = \sum_{j=1}^{k-1} \theta_j V_j(\pi)$, and $P_{\theta, \pi_0}(\pi) = \frac{\exp^{-D_\theta(\pi, \pi_0^{-1})}}{\psi(\theta)}$.

In a recent paper (Meilă and Bao 200x), we extend the stagewise ranking model to the partial ranked data and to the case of infinitely many items. A partial ranking (e.g. *top-t* ranking) refers to the situation in which there are k items ($k \leq \infty$), but each judge specifies only his first through t th choices, where $t < k$. (see Critchlow 1985 for a review of partial ranking metrics). Meilă and Bao define the infinite version of Mallows' type models (one parameter or $k-1$ parameters), give procedures to estimate its parameters and central permutation from data, and demonstrate that the *top-t* ranking model has sufficient statistics and thus an exponential family model. Suppose that we are given a set of *top-t* rankings D , then we will show the likelihood of complete data based on the sufficient statistics. Let \hat{q}_{ij} be the number of times i is observed in rank j in D ; \hat{Q}_{ij} be the number of times item j precedes item i in D ; N_j be the number of times that contain rank j and $T = \sum_j N_j$. We define $R = q1^T - Q$, and let $L(R)$ denote the sum of lower triangular part of R . We have log-likelihood for the generalized Mallows' model in eq. (2) and for single parameter Mallows model in eq. (3):

$$\ln P_{\theta, \pi_0}(D) = - \sum_j \theta_j L(R_j) - \sum_j N_j \ln \psi(\theta_j) \quad (2)$$

$$\ln P_{\theta, \pi_0}(D) = -\theta L(R) - T \ln \psi(\theta) \quad (3)$$

For any fixed centroid π_0 , the optimal θ is achieved at $\hat{\theta} = \ln(1 + T/L_{\pi_0}(R))$. The optimal π_0 is the (partial) permutation that minimizes the lower triangular part of sufficient statistic R .

2 Blurring Mean-Shift Clustering for Ranked Data

Nonparametric clustering is motivated by the fact that in many real applications the number of clusters is unknown and outliers exist. We consider an adapted version of the well known blurring mean-shift algorithm for the ranked data (Fukunaga and Hostetler 1975; Cheng 1995; Carreira-Perpinan 2006). For a dataset $D = \{x_i\}_{i=1}^n$, the Nadaraya-Watson kernel estimator is defined by

$$\hat{r}(x) = \sum_{i=1}^n \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)} x_i \quad (4)$$

where K is a kernel with bandwidth $h > 0$, and $x - x_i$ is considered as a distance between x and i th data point. In blurring mean-shift, each point x_i of the dataset actually moves to the point $\hat{r}(x_i)$ given by eq. (4). Thus, one iteration of blurring mean-shift results in a new dataset \tilde{D} which is a blurred (shrunk) version of D . Since mean-shift algorithm does not depend on parameters such as step size or number of clusters, the clustering is deterministic given the bandwidth h .

We choose the exponential Kernel $K_h(\pi) = \frac{\exp\left(-\frac{D(\pi)}{h}\right)}{\psi(h)}$ for ranked data, where $D(\pi)$ could be any proper metric for ranked data. Under the Kendall's metric it has the same form as the one parameter Mallows' model with bandwidth $\frac{1}{\theta}$: $K_\theta(\pi) = \frac{\exp^{-\theta D(\pi)}}{\psi(\theta)}$. This paper will mostly focus on the one parameter Mallows' model. The central ranking π_0 and the normalization constant $\psi(\theta)$ for Mallows type's models are often difficult to evaluate explicitly from data. However, the Nadaraya-Watson kernel estimator $\hat{r}(\pi_i) = \sum_{j=1}^n \frac{\exp^{-\theta d(\pi_i, \pi_j)}}{\sum_{i=1}^n \exp^{-\theta d(\pi_i, \pi_j)}} \pi_j$ depends on neither $\psi(\theta)$.

Some practical experience shows that blurring mean-shift merges the points into compact

clusters in a few iterations and then these clusters do not change but simply approach each other until they eventually merge at a single point (Carreira-Perpinan 2006). Therefore, to obtain a meaningful clusters, some proper stopping criterion should be proposed in advance. For the ranked data, this is not the case: at each iteration step we can round the local estimator into the nearest integer permutation and the algorithm will stop in a finite number of steps, when no ranking moves from its current position.

Moreover, because the ranked data is in a discrete set, we can also preform an accelerating process. As soon as two or more π'_i s become identical, we replace them with a single permutation and assign a weight proportional to the number of replicated permutations. The total number of iterations remains the same as for the original exponential blurring mean-shift but each iteration uses a dataset with fewer elements and is thus faster. This will be particularly effective if the number of items to be ranked is small (see 3.5 for example). Finally, we summarize the exponential blurring mean-shift algorithm for partial ranked data which is also suitable for full ranked data:

Algorithm EBMS

Input Top- t orderings $\mathcal{D} = \{(\pi_i)^{-1}\}_{i=1:n}$, with same or different lengths t_π

1. For $(\pi_i)^{-1} \in \mathcal{D}$ compute q_i, Q_i, R_i the sufficient statistics of a single data point.
2. Reduce dataset by counting only the distinct permutations to obtain reduced $\tilde{\mathcal{D}}$ and counts $n_i \geq 1$ for each ordering $(\pi_i)^{-1} \in \tilde{\mathcal{D}}$.
3. For $(\pi_i)^{-1}, (\pi_j)^{-1} \in \tilde{\mathcal{D}}$ calculate Kendall distance $d_{ij} = D_K((\pi_i)^{-1}, (\pi_j)^{-1})$
4. Set the scale θ by solving the equation

$$E_\theta[d(\tilde{\mathcal{D}})] = \frac{t \times e^{-\theta}}{1 - e^{-\theta}} - \sum_{j=1}^t \frac{j \times e^{-j\theta}}{1 - e^{-j\theta}}$$

where we set $E_\theta[d(\tilde{\mathcal{D}})]$ to be the average of pairwise distances in step 3.

5. For $\pi_i \in \tilde{\mathcal{D}}$ (*Compute weights and shift*)
 - (a) For $\pi_j \in \tilde{\mathcal{D}}$: set $\alpha_{ij} = \frac{\exp(-\theta d_{ij})}{\sum_{j'=1}^n \exp(-\theta d_{ij'})}$
 - (b) Calculate $\bar{R}_i = \sum_{\pi_j \in \tilde{\mathcal{D}}} n_j \alpha_{ij} R_j$
 - (c) Estimate σ_i^{-1} the “central” permutation that optimizes \bar{R}_i
(exactly or by heuristics)
 - (d) Set $(\pi_i)^{-1} \leftarrow \sigma_i^{-1}(1 : t)$
6. Repeat from step 4 if θ is reestimated or from step 2 if it's not until no $(\pi_i)^{-1}$ changes.

Output $\tilde{\mathcal{D}}$

Figure 1: The EBMS algorithm.

3 Experiments

In the experiment section, (1) we study effect of different bandwidths on the blurring mean-shift clustering; (2) we implement an estimation step so that the optimal bandwidth will be used at each mean-shift iteration; (3) we compare the rounding procedure in step 5 with no rounding and with the k-means clustering ; (4) we perform mean-shift clustering in the infinite situation; (5) we apply the method on a real dataset: The 1996 Hamburger Preparation Quiz (HPQ).

3.1 Experiment 1

For experiments 1-3, we generate data with 3 clusters of 150 rankings each and with spread parameters $\theta_1, \theta_2, \theta_3$ equal to 1.5, 1 and 0.7 respectively. Let l denotes the cluster index, the cluster centers are as following:

θ_l	n_l	Center Ranking
$\theta_1 = 1.5$	150	1 2 3 4 5 6 7 8 9
$\theta_2 = 1.0$	150	4 5 6 7 8 9 1 2 3
$\theta_3 = 0.7$	150	7 8 9 1 2 3 4 5 6
$\theta_4 = 0.0$	50	random outliers

Table 1: Simulation Design under Finite Mallows' Model

Let $C = \{C_1, C_2, \dots, C_L\}$ denote the real partition of the data, C' denote the clustering result, and $n_{l,l'}$ be number of data points matching between the corresponding cluster label in C and C' . The performance of classification is evaluated by the classification error (CE) distance $d_{CE} = 1 - \frac{1}{n} \max_{\delta} \sum_{l=1}^L n_{l,\delta}$, where δ is an injective mapping of $1, \dots, L$ into $1, \dots, L'$, and the maximum is taken over such mappings (Meilă 2005).

We calculate the Kendall distance between each pair of permutations and give the histogram

of these $n \times (n - 1)$ pairwise distances in Figure 2. For a full ranking of 9 items, the maximum distance is 36. The first peak in the histogram indicates that the distances between the permutations within the same cluster is around 5; the second peak indicates the distances between the permutations from different clusters. The distance among cluster centers is 18.

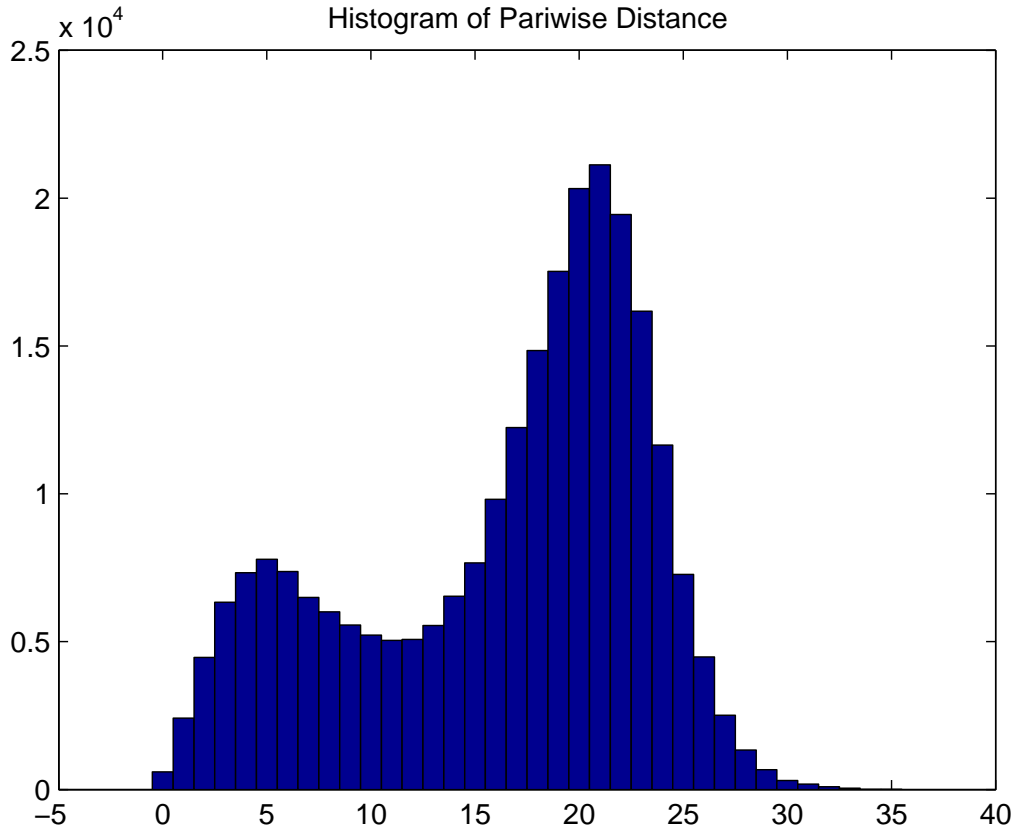


Figure 2: Histogram of Pairwise Distances

Because the Kernel bandwidth $\frac{1}{\theta}$ is deterministic for the blurring mean-shift algorithm, our first task is to compare the performance of mean-shift under different choices of θ . The number of clusters, classification errors, and number of iterations for $\theta = 0.1, 0.2, \dots, 2.0$ are shown in table 2. We find that the smallest classification error 4% is achieved at $\theta = 0.6$,

and 51 clusters are detected which is quite close to the simulation design : 3 clusters and 50 outliers. For $\theta > 1.0$, mean-shift fails in partitioning the data into clusters because only a few data points are merged together. On the other hand, for $\theta < 0.4$, it assumes an almost uniform distribution over the data, and the permutations merge too fast so that outliers are hardly detected. Note that the optimal θ is slightly smaller than θ_3 the most dispersed cluster.

θ	number of clusters	classification error	number of iterations
0.1	1	0.7000	9
0.2	5	0.1980	8
0.3	4	0.1020	7
0.4	11	0.0880	8
0.5	30	0.0560	7
0.6	51	0.0400	8
0.7	75	0.0600	12
0.8	108	0.2040	10
0.9	148	0.2980	19
1.0	238	0.4480	11
1.1	255	0.9962	19
1.2	298	0.9964	8
1.3	330	0.9968	6
1.4	345	0.9973	8
1.5	357	0.9990	7
1.6	364	0.9993	9
1.7	377	0.9994	7
1.8	395	0.9996	7
1.9	410	0.9997	4
2.0	413	0.9997	3

Table 2: The Choice of θ in Blurring Mean-Shift Clustering

3.2 Experiment 2

Instead of fixing θ value at the beginning, we consider estimating it at each iteration step. In Experiment 2, we compare the following two estimators: (a) We obtain the unique center ranking π_0 over all of the data points and then the m.l.e of θ is given by $\hat{\theta} = \ln(1+T/L_{\pi_0}(R))$

by assuming that all of the permutations coming from a single cluster, where $L_{\pi_0}(R)$ denote the sum of lower triangular part of $R = q1^T - Q$ given the cluster center π_0 . (b) Instead of assuming a unique cluster, θ can be computed from the average of pairwise distances. For one parameter Mallows' model, the m.l.e. of θ is the solution of $E_\theta[D_K] = \frac{t \times e^{-\theta}}{1 - e^{-\theta}} - \sum_{j=1}^t \frac{j \times e^{-j\theta}}{1 - e^{-j\theta}}$. We set $E_\theta[D_K]$ to be the average of pairwise distances in EBMS algorithm step 4.

The two approaches both converge in 7 steps which is the least number of iterations in Table 2 that corresponds to the classification error less than 10%. The second one by using average pairwise distances has a smaller error rate 4.2%. 32 outliers are found successfully, 5 outliers are merged into the first cluster, 6 outliers are merged into the second cluster, 7 outliers are merged into the third cluster. Moreover, 1 ranking from the first cluster and 2 rankings from the second cluster are misclassified into the third cluster.

	number of clusters	classification error	number of iterations	θ
a	27	0.0620	7	0.4833 \sim 0.5009
b	37	0.0420	7	0.5099 \sim 0.5813

Table 3: Performance of Exponential Blurring Mean-Shift Clustering with Adaptive θ

The Principal Component Plot for the original data and converged data is presented in Figure 3, where red points (circle) indicates the first cluster, blue the second cluster, green the third cluster, and yellow points are outliers. We will use method (b) to optimize θ in the following sections.

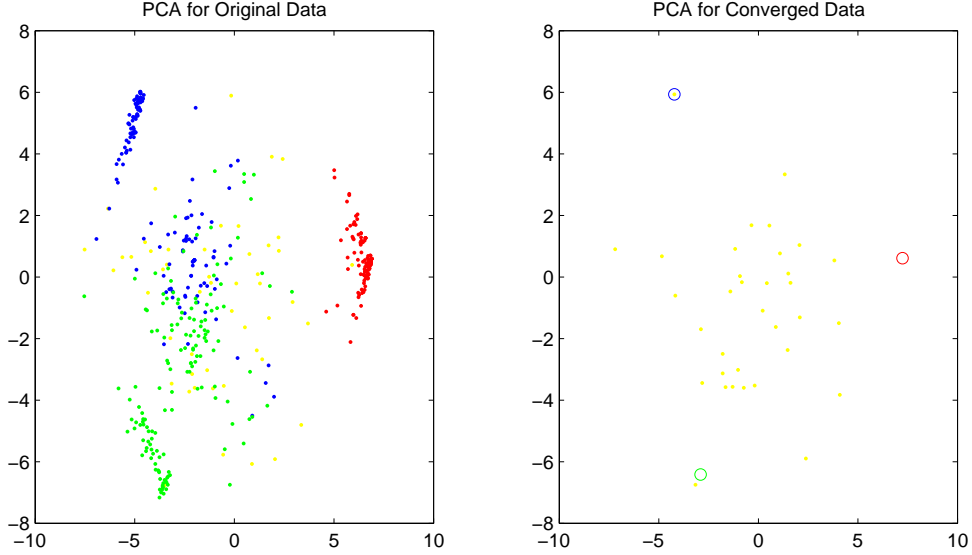


Figure 3: PCA plots for original rankings and converged rankings

3.3 Experiment 3

As we have discussed, at each blurring mean-shift iteration, we round the new data \tilde{D} into a discrete set of permutations, so that we can avoid the issues of stopping criterion and apply the accelerating process. However, we are also interested in the situation that continuous steps are taken in the convex hull of the space of ranking instead of rounding the \hat{Q}_i into permutations. In this continuous blurring mean-shift algorithm, the elements in the permutation matrix, Q_{ij} , represents the probability that j precedes i , and no acceleration will be implemented. We define the stopping criterion at the point that the average shift $\sum_{ij} |Q_{ij} - Q'_{ij}|$ is below a tolerance (say 1.00). Moreover we use the k-means method without rounding permutations at each step, and find the optimal number of clusters that gives the smallest classification error. The number of clusters (NC) and classification error (CE) of discrete mean-shift, continuous mean-shift and k-means are compared on 10 simulated data.

As a result, the k-means clustering always prefer 3 clusters and has an error rate around 10%

discrete mean-shift		smoothed mean-shift		k-means	
NC	CE	NC	CE	NC	CE
41	0.038	29	0.052	3	0.102
43	0.052	31	0.060	3	0.100
34	0.072	28	0.068	3	0.104
37	0.068	28	0.074	3	0.112
31	0.054	29	0.080	3	0.106
42	0.056	25	0.062	3	0.100
45	0.042	31	0.058	3	0.100
42	0.036	34	0.048	3	0.102
40	0.054	30	0.062	3	0.102
44	0.040	34	0.054	3	0.102

Table 4: Comparison of Clustering Methods

which is about the proportion of the outliers. The mean-shift clustering performs better than k-means in all of the cases, and the discrete version works better than continuous version. The exponential blurring mean-shift method for ranking data has the following advantages that K-means method does not have: 1. it does not depend on the initial partition; 2. it does not need to specify the number of clusters in advance; 3. it can detect the outliers; 4. The generalized Mallow’s distance is asymmetric: $D(\pi, \sigma) \neq D(\sigma, \pi)$. The mean-shift allows the asymmetric distance in the posterior probabilities; 5. K-means method treats all positions in the permutation equally, but mean-shift allows more weights in the first couple of positions.

3.4 Experiment 4

In experiment 4, we study the *top-9* rankings in the infinite items situation. We generate data with 3 clusters of 150 rankings each, having spread parameters $\theta_1, \theta_2, \theta_3$ equal to 1.5, 1.0, 0.7 respectively. The cluster centers are random permutations. In addition, each data set contains 50 outliers. We run the discrete blurring mean-shift 10 times on samples from this distribution. The scale parameter is estimated based on the average of pairwise

distances. In step 4 of the algorithm, the new ranking can be much longer than the original partial ranking. We truncate the new ranking to the length 9. Table 5 records the number of clusters and classification error in each generation. We obtain perfect classification in 8 cases and only have one misclassification in the other 2 cases.

number of clusters	classification error
54	0.002
53	0.000
53	0.000
53	0.000
53	0.000
53	0.000
53	0.000
53	0.000
54	0.002
53	0.000

Table 5: Infinite Items Situation

3.5 Experiment 5

Finally, we apply the exponential blurring mean-shift on a real data : The 1996 Hamburger Preparation Quiz (HPQ). The data was conducted by the Market Research Corporation of America as a supplement to its ongoing Menu Census Survey during March 1996 – February 1997. A part of the survey asks the responder ranks the following five types of hamburger patties in terms of their taste: rare, median rare, median, median well and well-done hamburger. The survey supplement was completed by 1,133 individuals, of which 607 provided complete responses to this question. Because only 44 different rank patterns are presented, the accelerating process could help to speed up the algorithm. We merge those 607 individuals into 6 clusters as in Table 6.

group	# of people	best	second best	third best	second worst	worst
1	191	well done	median well	median	median rare	rare
2	66	median well	well done	median	median rare	rare
3	121	median well	median	well done	median rare	rare
4	80	median	median well	median rare	well done	rare
5	79	median	median rare	median well	rare	well done
6	70	median rare	median	rare	median well	well done

Table 6: Hamburger Preference

4 Discussion

The exponential blurring mean-shift algorithm (EBMS) shifts the “points” (i.e $top-t$ orderings) to new locations obtained by a locally weighted combination of all the data. Thus, every π^{-1} is “attracted” towards its closest neighbors; as the shifting is iterated the data collapse into one or more clusters. The algorithm has a *scale parameter* θ . The scale influences the size of the local neighborhood of a $top-t$ ordering, and thereby controls the granularity of the final clustering: for small θ values (large neighborhoods), points will coalesce more and few large clusters will form; for large θ 's the orderings will cluster into small clusters and singletons. We studied different variation of this method, such as bandwidth estimation, acceleration process, discrete mean-shift, smoothed mean-shift, for finite many items and infinite many items. Our experiments show that the EBMS algorithm can be successfully applied in clustering the ranking data. In the EBMS clustering, the number of clusters does not need to be specified in advance and outliers can be detected. However the experiments are limited to a single parameter θ . We hope to overcome this limitation and to design non-parametric clustering algorithms with different θ_j parameters. The value of θ_j represents the importance of stage j to the previous stage. So, it is natural to have decreasing θ_j 's whenever the first ranks are more important than the rest.

References

- Carreira-Perpinan, M. A. (2006) “Fast Nonparametric Clustering with Gaussian Blurring Mean-Shift,” *Proceeding of the 23rd International Conference on Machine Learning*, Pittsburg.
- Cheng, Y. (1995) “Mean Shift Mode Seeking, and Clustering,” *IEEE Trans. PAMI*, 17, 790-799.
- Critchlow, D. (1985) “Metric Methods for Analyzing Partially Ranked Data,” *New York: Springer Verlag*.
- Diaconis, P. (1982) “Group Theory in Statistics,” *Institution of Mathematical Statistics Lecture Notes*.
- Fukunaga, K. , Hostetler, L. D. (1975) “The Estimation of The Gradient of a Density Function with Application in Pattern Recognition,” *IEEE Trans. Inf. Theory*, 21, 32-40.
- Kendall, M. G. (1938) “A New Measure of Rank Correlation,” *Biometrika*, 30, 81-93.
- Meilă, M. (2005) “Comparing Clusterings - An Axiomatic View,” *Proceedings of the 22nd Internatinoal Conference on Machine Learning, Bonn, Germany*.
- Meilă, M. , Bao, L. “Estimation and Clustering with Infinite Rankings,” *submitted*