Predictive Profiles for Transaction Data using Finite Mixture Models

Technical Report No. 01–67,
Information and Computer Science Department,
University of California, Irvine

Igor V. Cadez[1], Padhraic Smyth[1], Edward Ip[2] and Heikki Mannila[3]

[1] Department of Information and Computer Science,
University of California, Irvine, CA 92697, U.S.A.
[2] Marshall School of Business
University of Southern California
Los Angeles, CA 90089, USA
[3] Nokia Research
Helsinki, Finland
icadez@ics.uci.edu, smyth@ics.uci.edu,
eddie.ip@marshall.usc.edu, heikki.mannila@nokia.com

Abstract

Massive transaction data sets are routinely recorded in a variety of applications including telecommunications, retail commerce, and Web site management. In this paper we address the problem of inferring models from such transaction data in the form of predictive profiles of individual behavior. We describe a generative mixture model that accounts for population heterogeneity in transaction generation. An approximate Bayesian framework is used for parameter estimation that combines an individual's specific history with more general population patterns. The proposed model is shown to consistently outperform non-mixture and non-Bayesian techniques in predicting out-of-sample individual behavior on two large real-world transaction data sets.

1 Introduction

Transaction data sets consist of records of pairs of individuals and events, e.g., items purchased (market basket data), telephone calls made (call records), requests to consumer help desks (call center data), and Web pages visited (from Web logs). Such data sets are increasingly common across a variety of applications in telecommunications, retail commerce, and Web site management. The analysis of such data has attracted increasing interest in recent years. For example, early research on association rule algorithms to efficiently search for correlations among items in retail transaction data (Agrawal, Imielenski & Swami, 1993) has led to a significant amount of subsequent work in data mining on a variety of related methods (e.g., see Han & Kamber, 2000). Similarly, collaborative filtering algorithms (Resnick et al., 1994; Heckerman et al., 2000) have been used to infer which items an individual may rate highly, given information about other items that the individual has already purchased or rated.

In this paper we are interested in a somewhat different problem, that of automatically inferring *predictive profiles* for individuals from historical transaction data, where a predictive profile is considered to be a model of an individual's transaction behavior. More specifically, it is a probability model that describes which items an individual is likely to purchase (or visit) in the future. The problem of inferring predictive profiles can be viewed as fundamental to the analysis of such data. Predictive profiles support many different types of analysis that a data-owner might wish to carry out on transaction data, e.g., visualizing and understanding customer behavior, forecasting of individual behavior, determining the life-time value of a customer, change detection, cross-selling and personalization, fraud detection, and so forth.

Figure 1 shows a set of transactions for five different individuals where rows correspond to market baskets (transactions) and columns correspond to categories of items (store departments in this example). The data set from which these examples are taken involves over 500,000 transactions from 200,000 customers over a two-year period in a set of retail stores. The heterogeneity of purchasing behavior, in terms of items purchased and number of transactions, is clear even from this simple plot. Our goal is to investigate parsimonious and accurate models for each individual's purchasing behavior given such data and we will refer to such models at the individual level as *predictive profiles*.

We propose a combination of mixture models and Bayesian estimation methods for learning predictive profiles. The mixture model is used to address heterogeneity: different individuals purchase different combinations of products on different visits to the store. The Bayesian framework is used to address the fact that we have varying amounts of data for different individuals. For an individual with very few transactions (e.g., only one) we can augment ("shrink" in Bayesian terminology) our predictive profile for that individual with (towards) a general population profile. On the other hand, for an individual with many transactions, the predictive model can be much more data-driven and individualized. Our goal is an accurate and computationally efficient modeling framework that smoothly adapts a profile to each individual based on both their own historical data as well as general population patterns. The techniques proposed can also be viewed as providing a model-based Bayesian alternative to other non-probabilistic approaches such as association rules that are widely used in data mining of transaction data.

The paper begins in section 2 by defining the general problem of profiling and discussing the

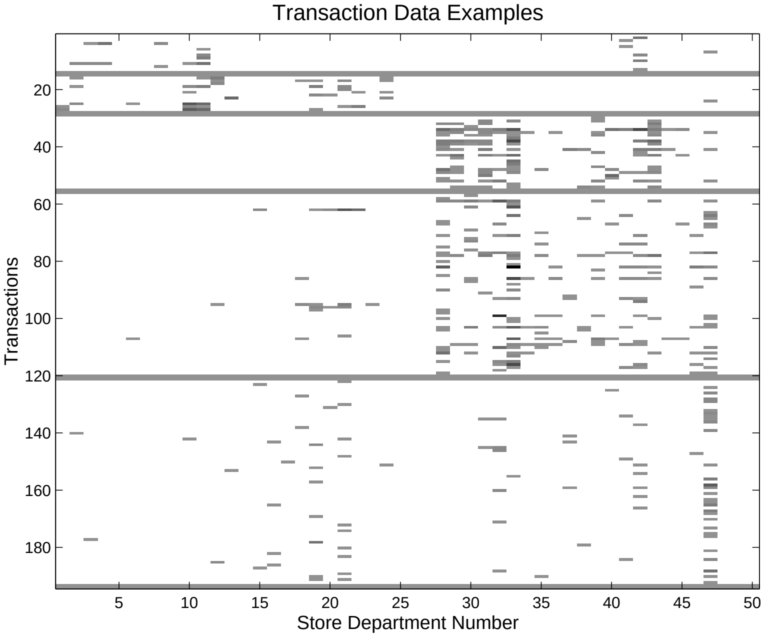

Figure 1: Examples of transactions for several individuals. The rows correspond to market baskets (or transactions) and the columns correspond to particular categories of items. The darker the pixel, the more items were purchased (white means zero). The solid horizontal gray lines indicate the boundaries between transactions of different individuals.

spectrum between sparse individual-specific information and broadly supported global patterns. In section 3 we define and motivate a mixture model framework for modeling transaction behavior at the individual level, and section 4 provides details on parameter estimation. Section 5 introduces the two real-world transaction data sets used in the paper and section 6 discusses experimental results where we compare various models in terms of their out-of-sample predictive performance and demonstrate the scalability of the algorithms to large data sets. Section 7 provides a general assessment of the modeling framework proposed in the paper and section 8 discusses related work. Conclusions are outlined in section 9.

2 Notation and Problem Definition

In this section we introduce some basic notation and provide a high–level overview of the problems involved in modeling transaction data.

2.1 General Notation

Consider an observed transaction data set $D = \{D_1, \ldots, D_N\}$ generated by N individuals, where D_i is the observed data on the ith customer, $1 \leq i \leq N$. Each individual data set D_i consists of a set of one or more transactions for that customer, i.e., $D_i = \{\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ij}, \ldots, \mathbf{y}_{in_i}\}$, where n_i is the total number of transactions observed for customer i, and $\mathbf{y}_{ij}$ is the jth transaction for customer i, $1 \leq j \leq n_i$. This notation could represent a customer buying products, an individual visiting Web pages, etc. For concreteness we will focus on examples from retail data, but the general approach is applicable to more general forms of transaction data.

An individual transaction $\mathbf{y}_{ij}$ consists of a description of the set of products that were purchased at the same time by customer i. For the purposes of the experiments described below, each individual transaction $\mathbf{y}_{ij}$ is represented as a set of C counts, $\mathbf{y}_{ij} = \{n_{ij1}, \ldots n_{ijc}, \ldots, n_{ijC}\}$, where for $1 \leq c \leq C$ the count n_{ijc} indicates how many items of type c are in transaction ij, $1 \leq c \leq C$. One can straightforwardly generalize this representation to include (for example) the price for each product, but here we focus just on the number of items (the counts). For the purposes of this paper we will ignore any information about the time or sequential order in which items are purchased or in which pages are visited within a particular transaction $\mathbf{y}_{ij}$, but the approach could be generalized to account for sequential order or timing information if available (e.g., using mixtures of Markov chains, Cadez et al., 2000).

We are assuming above that each transaction is "tagged" with a unique identifier for each individual. Examples of such identification schemes can include frequent shopper cards, driver's licenses or credit cards for retail purchasing, and login or cookie identifiers for Web visits. There are of course various practical problems associated with such identification, such as data entry errors, missing identifiers, fraudulent or deliberately disguised IDs, multiple individuals using a single ID, ambiguity in identification on the Web, and so forth. Nonetheless, in an increasing number of transaction data applications reliable identification is possible. In the rest of the paper we will assume that this identification problem is not an issue and assume that either the identification process is inherently reliable or that there are relatively accurate techniques to discern identity.

In fact for the two real-world transaction data sets that we use to illustrate our techniques, the identification process is considered to be quite reliable.

2.2 The Predictive Profiling Problem

We define a predictive profile for individual i as a model that predicts the distribution of items in future transactions $\mathbf{y}_{ij}$ from that individual, e.g., a multinomial model over items purchased. The problem then is to infer predictive profiles for each of N individuals, given historical transaction data D. In the work described in this paper our predictive models are conditioned on the total number of items n_{ij} purchased by each individual per transaction, i.e., in scoring our models on future data we want to know *which* items an individual bought given that they purchased some total number. One could also incorporate a predictive component that forecasts for each individual the total number of items purchased per transaction and the *rate* at which transactions are generated. For example, a Poisson model for store visits (transactions) is commonly used in the marketing literature (Schmittlein, Morrison & Colombo, 1987; Wedel & Kamakura, 1998). The Poisson rate could be allowed to vary per individual, to incorporate seasonality, and so forth. In the marketing literature the rate model (how often purchases are made) and the choice model (which items are purchased) are often decoupled and developed separately. In this paper we focus on the choice component. However, given that we are using a probabilistic framework, we could in principle broaden the scope of the model by coupling our choice model with any appropriate rate model, and so forth.

Data sparsity is a distinctive feature of many transaction data sets. For example, for many transaction data sets (including the particular data set corresponding to Figure 1), a histogram of "number of transactions" peaks at 1 (i.e., more customers have a single transaction than any other number of transactions) and then decreases exponentially quickly. Thus, for many customers there are very few transactions on which to base a profile, while for others there are large numbers of transactions. The challenge is to systematically leverage all such information in generating individual-level predictions.

Assume for example that we model each individual via a simple multinomial probability model to indicate which items are chosen, namely, a vector of probabilities p_c, one for each of the C categories of items, with $\sum_{c=1}^{C} p_c = 1$. A very simple approach would be to estimate this multinomial via a maximum likelihood "histogram" estimated from raw counts for that individual. This is certainly individual-specific. However, it suffers from at least two significant problems. First, for individuals with very small amounts of data (such as those with only one item in one transaction) the profiles will be extremely noisy and unreliable. Secondly, even for individuals with significant amounts of data, the "raw counts" do not contain any notion of generalization: if an individual did not purchase a specific item in the past the profile probability for that item is zero, i.e., the model predicts that the individual will never purchase it.

We can see the lack of generalization in the example of Figure 2. For example, this individual did not make any purchases in department 14 and consequently a maximum likelihood model would put a zero probability in that cell. However, this individual did in fact make a purchase in that department in the future (lower plot, Figure 2c).

An obvious solution is to try to smooth the histogram. A simple approach is to use a maximum

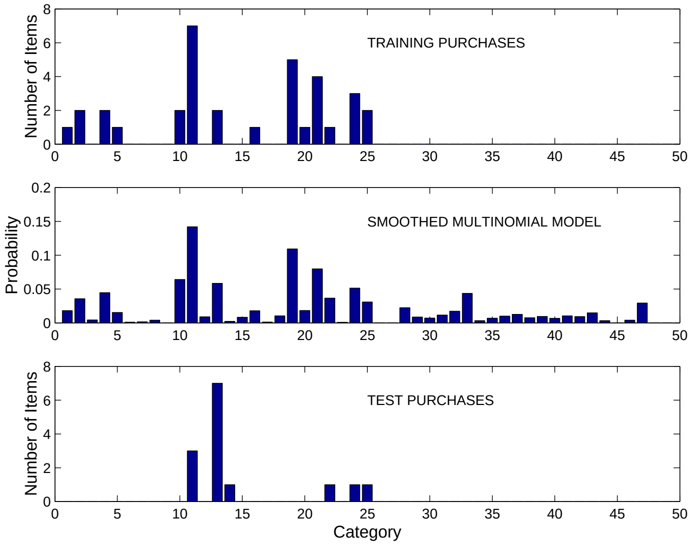

Figure 2: An example of (a) a particular individual's historical data, (b) a smoothed histogram model, and (c) future purchases for the same individual. The x-axis corresponds to items categorized at the department level (as in Figure 1) and the y-axis to either categories purchased or multinomial probabilities of purchasing categories of items.

a posteriori (MAP) estimate of the multinomial model. For example if we use a Dirichlet prior, where the mean corresponds to the normalized histogram of items purchased by all individuals and the equivalent sample size is set to 5 items, we get an MAP estimate of the multinomial as shown in Figure 2b. The "MAP profiles" will be proportional to a linear combination of the maximum likelihood histogram (from the observed data D_i) and the mean of the Dirichlet prior. If we have no historical data at all for an individual, their profile will correspond to the mean of the prior, the "population histogram." If we have substantially more than 5 items, the profile will tend to look much more similar to the maximum likelihood estimate, but will be smoothed to have small non-zero probabilities for items that were not purchased.

Although this is likely to be a better predictor than the simple maximum likelihood histogram, there is still a problem with the MAP profile model in Figure 2b. Note that the individual only purchased items in departments numbered from 1 to 25, and did not make any purchases in departments above 25, both in the historical data used to build the model and in the future data. Nonetheless, the MAP predictive profile "over-generalizes" to the extent that it places considerable probability mass in departments between 26 and 50. This is because a significant fraction of overall population purchases come from those departments. It turns out that in this data set the departments numbered 25 and lower are primarily men's clothes, and those above 25 are primarily women's clothes. The data in Figure 2 is likely to be from a male shopper who only shops in men's departments. By using a very broad prior based on the whole population (both male and female) we have in effect over-generalized and are making predictions about departments between 26 and 50 that are unlikely to be true.

To deal with these problems we investigate (a) a mixture model to handle heterogeneity of purchasing behavior (e.g., different male and female customer behaviors) and (b) an approximate Bayesian scheme for parameter estimation, allowing the combination of sparse information about an individual with global population patterns.

3 Mixture Models for Transaction Data

In this section we first describe a global mixture model for transactions and then describe how individual-specific weights are derived in the context of this model.

3.1 The Global Mixture Model

We propose a simple generative mixture model for an individual's purchasing behavior, namely that a randomly selected transaction $\mathbf{y}_{ij}$ from individual i during store visit j is generated by one of K components in a K-component mixture model, i.e.,

$$p(\mathbf{y}_{ij}) = \sum_{k=1}^{K} \alpha_k P_k(\mathbf{y}_{ij}), \tag{1}$$

where $\sum_k \alpha_k = 1$. The k-th mixture component P_k, $1 \leq k \leq K$, is a specific model for generating the counts in a basket and we can think of each of the K models as "basis functions" describing prototype transactions. For example, one might have a mixture component that acts as a prototype for suit-buying behavior, where the expected counts for items such as suits, ties, shirts, etc., given this component, would be relatively higher than for the other items.

There are several modeling choices for the component transaction models for generating item counts. In this paper we choose a particularly simple memoryless multinomial model that operates as follows. Conditioned on n_{ij}, the total number of items in the basket, each of the individual items is selected in a memoryless fashion by n_{ij} draws from a multinomial distribution $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kC})$ on the C possible items. Thus, the overall mixture model can be written as

$$p(\mathbf{y}_{ij}) = \sum_{k=1}^{K} \alpha_k \prod_{c=1}^{C} \theta_{kc}^{n_{ijc}}, \tag{2}$$

where $\mathbf{y}_{ij} = \{n_{ij1}, \ldots n_{ijc}, \ldots, n_{ijC}\}$. Note that this probability model is *not* a multinomial model, i.e., the mixture has richer probabilistic semantics than a simple multinomial.

Other component models are possible. For example, one could model the data as coming from C conditionally independent random variables (given the component label), each taking non-negative integer values. The multinomial model in Equation 2 constrains the number of purchases of any item to follow a geometric distribution (per cluster), but the conditional independence model allows any form of distribution to be modeled in principle. Such a conditional independence model could allow (for example) the modeling of the purchase of specific numbers of specific items, per component, in a manner that the multinomial multiple trials model cannot achieve. However, it can also require more parameters than the multinomial model to achieve this degree of flexibility. McCallum and

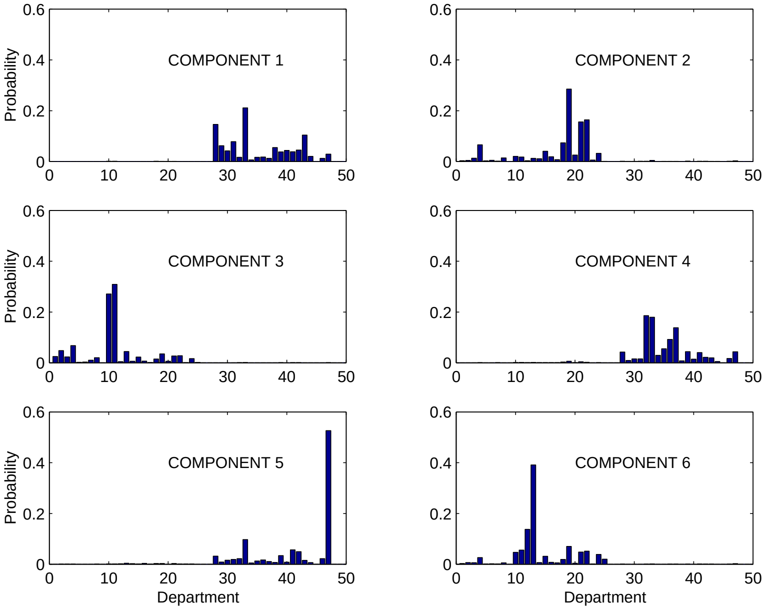

Figure 3: An example of 6 "basis" mixture components fit to retail transaction data.

Nigam (1998) compared the use of multinomials and conditional independence models as class models for term vectors in document classification using naive Bayes. The conditional independence model provided more accurate classification, but the difference in accuracies between the two models was not particularly significant. In this paper we focus on the multinomial model since it is the simpler of the two component models to work with.

The use of multinomial mixture models for "count" data is of course not new. Similar models have been widely used in applied statistics for many years to model relatively low-dimensional contingency table data (e.g., Lazarsfeld & Henry, 1968). More recently, multinomial mixture models have become popular in machine learning for applications such as document classification and clustering (e.g., McCallum, 1999). The novelty of the work presented in this paper lies in the application of such models to high-dimensional massive transaction data sets, and their extension to account for individual heterogeneity as described in the next section.

Figure 3 shows an example of $K = 6$ such basis mixture components that have been estimated from the large retail transaction data of Figure 1 (more details on estimation will be discussed below). Each window shows a particular set of multinomial probabilities that models a specific type of transaction. The components show a striking bimodal pattern in that the multinomial models involve sets of departments that are either above or below department 25, but there is very little probability mass that "crosses over." The models are capturing the fact that departments numbered lower than 25 correspond to men's clothing and those above 25 correspond to women's clothing.

We can see further evidence of this bimodality in the data itself in Figure 1 noting that some individuals purchase items from "both sides" depending on the transaction. However, typically this "cross-over" does not occur within the same basket of items. This might suggest (for example) a husband and wife who are each using the same shopping card or ID, each with their own shopping patterns. Our mixture approach can directly capture such heterogeneity.

3.2 Individual-Specific Weights

We now assume that for each individual i there exists a set of K individual-specific weights, denoted by $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})$, where $\sum_k \alpha_{ik} = 1$. Thus, the model for each individual is a mixture model where the weights are specific to individual i:

$$\begin{aligned} p(\mathbf{y}_{ij}) &= \sum_{k=1}^{K} \alpha_{ik} P_k(\mathbf{y}_{ij}) \\ &= \sum_{k=1}^{K} \alpha_{ik} \prod_{c=1}^{C} \theta_{kc}^{n_{ijc}}, \end{aligned} \tag{3}$$

where θ_{kc} is the probability that the cth item is purchased given component k and n_{ijc} is the number of items of category c purchased by individual i, during transaction j.

The weight α_{ik} represents the probability that when individual i enters the store, his or her transactions will be generated by component k. In other words, the α_{ik}'s govern individual i's propensity to engage in "shopping behavior" k (again, there are numerous possible generalizations such as making the α_{ik}'s dependent over time, that we will not discuss here). The α_{ik}'s in effect define *individual i's predictive profile*, relative to the K component models.

This idea of individual-specific weights (or profiles) is a key component of our proposed approach. The mixture component models P_k are fixed and shared across all individuals, providing a mechanism for borrowing of strength across individual data. The individual weights α_{ik} are in principle allowed to freely vary for each individual within a K-dimensional simplex. In effect the K weights can be thought of as basis coefficients that represent the location of individual i within the space spanned by the K basis functions (the component $\boldsymbol{\theta}_k$ multinomials). This approach is quite similar in spirit to the recent work of Hofmann (1999) on "aspect models" for text documents, where he proposes a general mixture model framework that represents documents as existing within a K-dimensional simplex of multinomial component models, somewhat similar to a probabilistic principal component analysis of binary data.

3.3 The Full Data Likelihood

The unknown parameters Θ in our full model consist of both the parameters of the K component multinomials, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, and the N vectors of individual-specific profile weights $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}$. Assuming that each of the N individuals behave independently given the model, the full likelihood of the data can be written as

$$p(D|\boldsymbol{\Theta}) = \prod_{i=1}^{N} p(D_i|\boldsymbol{\Theta}), \tag{4}$$

Furthermore, assuming that each transaction (or basket) for each individual is generated independently of the other transactions for that individual, conditioned on the model and the parameters, we have

$$p(D_i|\boldsymbol{\Theta}) = \prod_{j=1}^{n_i} p(\mathbf{y}_{ij}|\boldsymbol{\Theta}), \tag{5}$$

where $p(\mathbf{y}_{ij}|\boldsymbol{\Theta})$ was defined above (absent the explicit dependence on $\boldsymbol{\Theta}$). While both the global and the individual-specific models depend on the conditional independence assumption of baskets to simplify computation, the model based on individual-specific weights represents an important improvement over the global model. In the global model, baskets from different individuals are not distinguished and are regarded as exchangeable under the conditional independence assumption. This is far from realistic given that multiple baskets generated by a single individual tend to be somewhat similar. On the other hand, the individual-specific weights capture the correlation among multiple baskets that belong to the same individual i, because all of them are generated, memorylessly, by the common set of α_{ik} under the individual-specific model.

Although the assumption of conditionally independent transactions for each individual is viable, there might be temporal or sequential dependencies between certain purchases that could still invalidate the model. Nonetheless, for simplicity, we model transactions as being memoryless in this paper since it allows us to capture general purchasing behavior to first-order using a relatively simple setup. The idea of conditional independence given individual-specific parameters has long been used in the statistics and psychometric literature to model within-individual repeated measurements that are clustered (e.g., Lord, 1980). This type of memoryless assumption is also widely used in the marketing literature for modeling consumer brand choice behavior and has been found to be both useful and accurate in various studies (see Wedel & Kamakura, 1998, for a general review).

4 Estimation of the Model Parameters

We use an MAP approach to estimate the unknown parameters $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}$. We will refer to the parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ for the K component multinomials as global structural parameters, and the weight vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$ as individual profiles. The MAP optimization problem is defined as

$$\Theta^{MAP} = \arg\max_{\Theta} \left\{ P(D|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) \prod_{k=1}^{K} P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_{i=1}^{N} P(\boldsymbol{\alpha}_i|\boldsymbol{\xi}) \right\}, \tag{6}$$

where $P(\boldsymbol{\theta}_k|\boldsymbol{\gamma})$ and $P(\boldsymbol{\alpha}_i|\boldsymbol{\xi})$ are independent Dirichlet priors on the component multinomial parameters and weight vectors, with parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ respectively. The parameter set $\boldsymbol{\gamma}$ consists of a C-dimensional mean vector $\boldsymbol{\gamma}_{mean}$ plus an equivalent sample size $\boldsymbol{\gamma}_{ess}$, and $\boldsymbol{\xi}$ is a K-dimensional mean vector $\boldsymbol{\xi}_{mean}$ plus an equivalent sample size $\boldsymbol{\xi}_{ess}$. Both $\boldsymbol{\gamma}_{mean}$ and $\boldsymbol{\xi}_{mean}$ are defined as probability vectors that sum to 1 with $\boldsymbol{\gamma}_{ess}$ and $\boldsymbol{\xi}_{ess}$ providing scaling for each. The same prior parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ are shared (respectively) by all K multinomial component models $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ and by all N weight vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$.

In a fully Bayesian framework, to locate the mode of the expression in equation 6, we would define hyperpriors on $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ and then integrate $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ out. Since there is no closed form solution for this integration given a mixture model of this form, this would require the use of numerical approximation techniques (such as Markov chain Monte Carlo sampling). Such sampling is not computationally tractable given the sizes of the transaction data sets that we analyze (e.g., 1.9 million transactions for one of the data sets considered in section 6). Similarly, a conventional empirical Bayes approach, where point estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ would first be determined by integrating out all other parameters, is also not practical for the same computational reasons.

Nonetheless, it is important to use sensible priors $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ to regularize the estimation of the parameters Θ in equation 6 above. This is particularly true for the individual profiles $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$ since they can be severely underconstrained by the data (for example, for individuals for whom only a small number of transactions are available), necessitating some form of regularization. With this in mind we use an initialization step to determine a reasonable data-driven value for the mean of the weight prior $\boldsymbol{\xi}$.

Specifically, we pool the transactions from all N individuals (ignoring the individual origins of the transactions) and constrain all of the weights for each individual to be identical, i.e., $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$. This is the "standard" mixture model approach, treating each transaction as being conditionally independent given a mixture of multinomials with a global set of weights. We run EM using noninformative priors to generate parameter estimates $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K$ for the global component models, and to generate an estimate $\hat{\boldsymbol{\alpha}}$ for the "global weights," under this model. The resulting mixture model will be referred to as the global mixture model in the remainder of the paper, since it can be used to provide a predictive profile where each individual has the same weight vector $\hat{\boldsymbol{\alpha}}$.

The parameter estimates produced by this initialization procedure, $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K$ and $\hat{\boldsymbol{\alpha}}$, are then used to provide initial parameter guesses for a full EM procedure that (locally) maximizes equation 6 over all parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ and over all N weight vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$. In this full EM procedure the transactions are grouped by individual, enabling individual weight estimation. The mean of the Dirichlet weight prior $\boldsymbol{\xi}_{mean}$ is set to the global weight vector $\hat{\boldsymbol{\alpha}}$ obtained from the earlier constrained estimation and the equivalent sample size $\boldsymbol{\xi}_{ess}$ is set to 5 baskets purchased. Experiments with various sample sizes between 0.1 and 10 indicated that the parameter estimates and/or predictions were not particularly sensitive to the exact value chosen for the equivalent sample size.

Both the initialization and full EM steps are summarized in table form in Table 1 (more complete details on the derivation of the EM procedure are provided in the Appendix). To initialize parameters during the initialization procedure we use a uniform weight vector for the weights $\boldsymbol{\alpha}$. For the multinomial components $\boldsymbol{\theta}_k$ we sample from a Dirichlet centered at the marginal probabilities of individual items $\boldsymbol{\theta}_{marginal}$ with an equivalent sample size $2C$ where C is the number of items. This "marginal sampling technique" was shown by Meila and Heckerman (1998) to be a useful heuristic for initializing component parameters for EM estimation of mixture models. For the initialization procedure we select the highest MAP solution from 20 random starts. For the final full EM procedure we only perform a single run. In effect our heuristic is that the initialization procedure focuses attention on relevant parts of the global parameter space for the full EM procedure.

Although the global structural parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ are allowed to vary freely during the final full EM procedure, we have found empirically that the final estimated components are quite close to the initial values provided by the initialization procedure. Thus, in effect, one could view the initialization procedure as estimating the global structure of the model, and the final full EM procedure as determining how best to represent each individual within the K-dimensional simplex of mixture weights relative to the global model. In section 6.6 we show that this "two-stage" approach

Table 1: Specification of the EM procedure.

	Initialization Procedure	**Full EM Procedure**
Parameters Estimated	$\hat{}\,, \widehat{}_{k}$	$\hat{}_{i}, \widehat{}_{k}$
Priors (Weights)	$\sim Dir\,(\ _{mean},\ _{ess})$ $_{mean} = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right), \quad _{ess} = 1$	$_{i} \sim Dir\,(\ _{mean},\ _{ess})$ $_{mean} = \hat{}\,, \quad _{ess} = 5$
Priors (Components)	$_{k} \sim Dir\,(\ _{mean},\ _{ess})$ $_{mean} = \left(\frac{1}{C}, \ldots, \frac{1}{C}\right), \quad _{ess} = 10^{-5}$	$_{k} \sim Dir\,(\ _{mean},\ _{ess})$ $_{mean} = \hat{}\,, \left(\frac{1}{C}, \ldots, \frac{1}{C}\right), \quad _{ess} = 10^{-5}$
Initialization (Weights)	$= \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$	$_{i} = \hat{}$
Initilialization (Components)	$_{k} \sim Dir\left(\ _{marginal}, 2C\right)$	$_{k} = \hat{\theta}_{k}$
Number of Runs	Best of 20 random starts	Single start
Convergence	Change in relative logL $< 10^{-4}$	Change in relative logL $< 10^{-4}$

(initialization plus full EM) produces comparable (indeed more accurate) results when compared to an alternative single-stage method where the prior is treated as an additional parameter. We will also show that the two-stage approach has certain computational advantages. For example, the initialization procedure can be run on a subsample of the full data set and obtain virtually identical results in a fraction of the computation time required by using the full data.

Although our use of an initialization phase to determine a data-driven prior is very much in the spirit of empirical Bayes, we are not performing the full integration over parameter space that would be traditionally carried out in an empirical Bayes methodology (for the computational reasons mentioned earlier). Nonetheless, the estimation scheme we use can loosely be viewed as a form of MAP approximation to a traditional empirical Bayes approach.

To illustrate the application of these ideas, recall that in Figure 2 we looked at "non-mixture" histogram models as predictive profiles of an individual. In Figure 4 we now plot for the same individual the predictive profile, obtained using both the global weight and the individual weight methods described above, where the global weights are those obtained from the initialization procedure based on a global model. One can see that the global weight profile based on $\boldsymbol{\alpha}$ reflects broad population-based purchasing patterns and is not representative of this individual. The individual-weight profile based on $\boldsymbol{\alpha}_i$ appears to be a much better representation of this individual's behavior and indeed it does provide the best predictive score of all the models on the test data in Figure 2. Note that the individual weight profile in Figure 4 "borrows strength" from the purchases of other similar customers, i.e., it allows for small but non-zero probabilities of the individual making purchases in departments even if he or she has not purchased there in the past (such as departments 6 through 9). This particular individual's weights, the α_{ik}'s, are $(0.00, 0.47, 0.38, 0.00, 0.00, 0.15)$ corresponding to the 6 component models shown in Figure 3. The most weight is placed on components 2, 3 and 6 (components that correspond to men's clothing departments), which agrees with

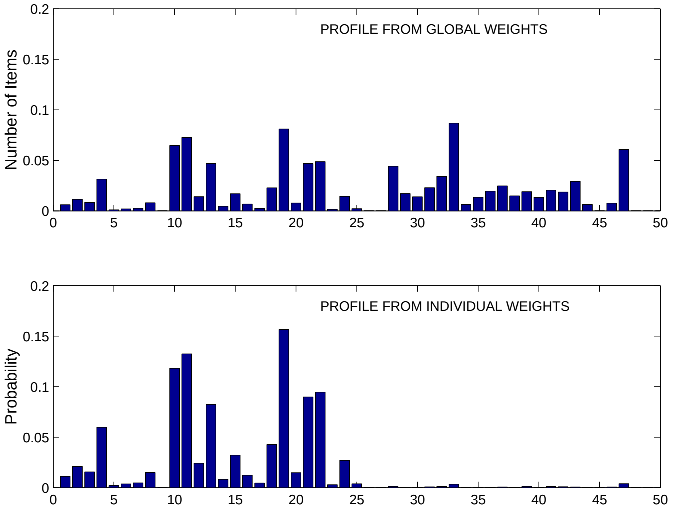

Figure 4: Inferred predictive profiles from (a) global weights, and (b) individual-specific weights, for the individual whose data was shown in Figure 2.

our intuition given the individual's training data (all purchases were in such departments).

5 Data Sets

We use two large real-world retail transaction data sets—a retail clothing data set, and a retail drugstore data set—in our experiments. When a basket is paid for, each item in the basket is recorded, along with the ID of the individual making the purchase, and the time and day of the purchase. For both data sets the items being purchased are categorized according to a product hierarchy. At the top of the hierarchy are broad categories such as departments within the store—at the bottom of this hierarchy are very specific product descriptions. For example, in the retail clothing data set below, there are 53 departments at the top level and approximately 50,000 specific item descriptions at the bottom level of the hierarchy. The categories at each level of this hierarchy are mutually exclusive and exhaustive. In this paper we focus on making predictions at the top two levels of this product hierarchy, and refer to them as level 1 and 2. For the retail clothing chain, the names of individual categories at each level have been replaced by numbers in this paper due to the proprietary nature of the data.

The first data set consists of purchases over a two-year period at a set of stores for a retail clothing chain in the United States. There are approximately 1.2 million items purchased during 500,000 separate transactions by approximately 200,000 different individuals. At level 1 in the product hierarchy items are categorized into 53 departments and at level 2 there are 409 categories.

The second data set contains transaction records collected during the period of 1996–1999 from a national drugstore retail chain in Japan. The data set consists of approximately 15.6 million items purchased during 2.5 million separate transactions by about 300,000 individuals over approximately 1,000 stores across Japan. The department level (level 1) of the product hierarchy comprises 21 categories (such as medical devices, baby products, etc.,) and the second level contains 151 more detailed categories.

6 Experimental Results

In this section we investigate the predictive performance of our proposed individual-specific mixture model and compare it to the alternative global mixture model and a non-mixture MAP histogram model.

6.1 Model Evaluation using Predictive LogP Scores

We separate each data set into two time periods. We train our mixture and weight models on the first period and evaluate our models in terms of their ability to predict transactions that occur in the subsequent out-of-sample test period.

To evaluate the predictive power of each model, we calculate the log-probability ("logp scores") of the transactions as predicted by each model. Higher logp scores mean that the model assigned higher probability to events that actually occurred. The log probability of a specific transaction from individual i, $\mathbf{y}_{ij} = (n_{ij1}, \ldots, n_{ijC})$ under mixture model parameters $\mathbf{\Theta}$, is defined as

$$\log p(\mathbf{y}_{ij}|\mathbf{\Theta}) = \log \sum_{k=1}^{K} \alpha_{ik} \prod_{c=1}^{C} \theta_{kc}^{n_{ijc}} \tag{7}$$

and the total logp score for a full data set D is the sum of the quantities, $\log p(\mathbf{y}_{ij}|\mathbf{\Theta})$, over all individuals i and their transactions j. The α's and θ's are parameter estimates, obtained using the methods described earlier in the paper. Note that the total negative logp score over a set of transactions, divided by the total number of items, can be interpreted as a predictive entropy term in bits (for log base 2). The lower this entropy term, the less uncertainty in our predictions (bounded below by zero of course, corresponding to perfect predictions).

We evaluate the prediction accuracy of each model for each data set at both the level 1 and level 2 aggregation levels of the product hierarchy. For the retail clothing data set we also evaluate the predictions of models built and evaluated on customers for whom there are 10 or more transactions in the training data, and models built and evaluated on customers for whom there are only 2 or more transactions in the training data. The "10 or more" group are of interest since they are the more frequent customers, while the "2 or more" group reflects a broader set of customers for whom there is a relatively small amount of historical data available for building predictive profiles. For the retail drugstore data set we evaluate the model only on the "10 or more" transactions group in the interests of time, since this data set is sufficiently large that it can take days to fit our models to the full data set over all values of K of interest.

6.2 Experimental Results on Retail Clothing Data

We constructed two different pairs of training and test sets where in each case the training data consists of the first 70% of transactions (chronologically) and the test data consists of the remaining

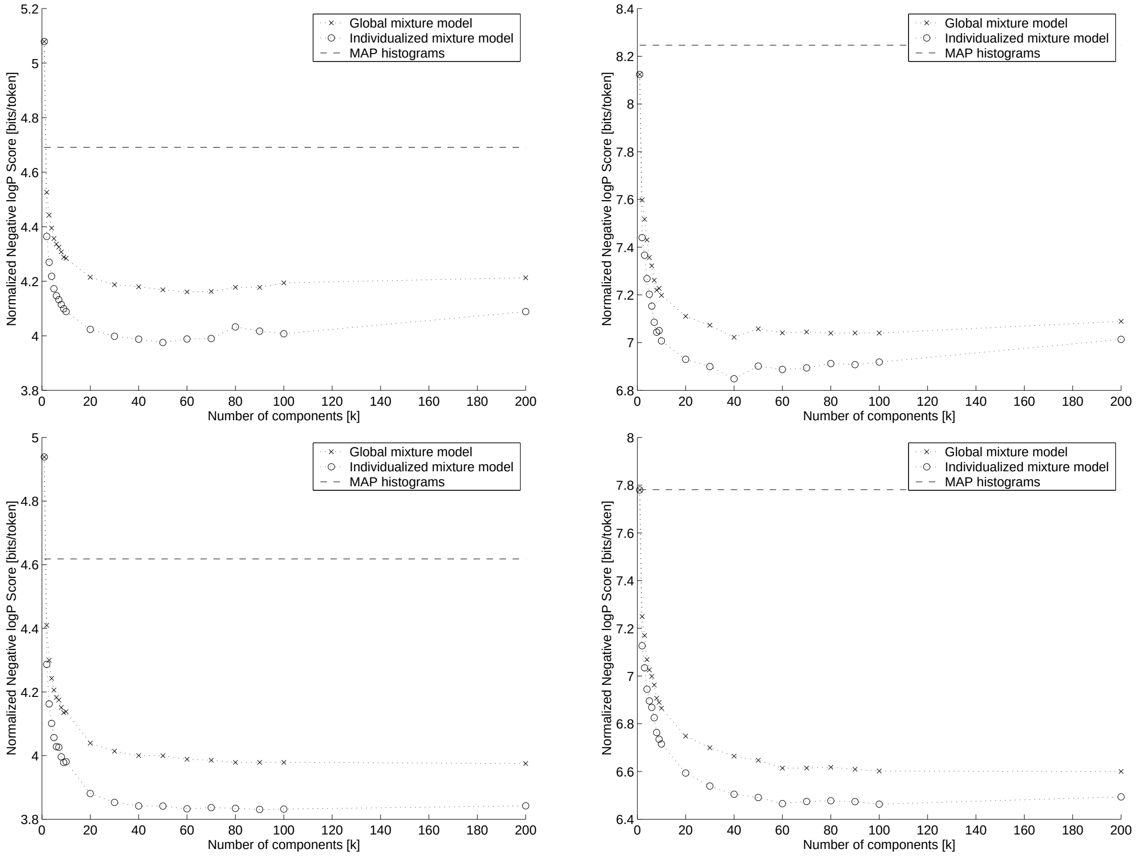

Figure 5: Retail Clothing Data: plot of the negative log probability scores per item (predictive entropy) on out-of-sample transactions as a function of K, the number of mixture components, using (a) individuals with at least 10 transactions in the training data with predictions at level 1 with 53 categories (upper left) and at level 2 with 409 categories (upper right) and (b) individuals with at least 2 transactions in the training data with predictions at level 1 with 53 categories (lower left) and at level 2 with 409 categories (lower right).

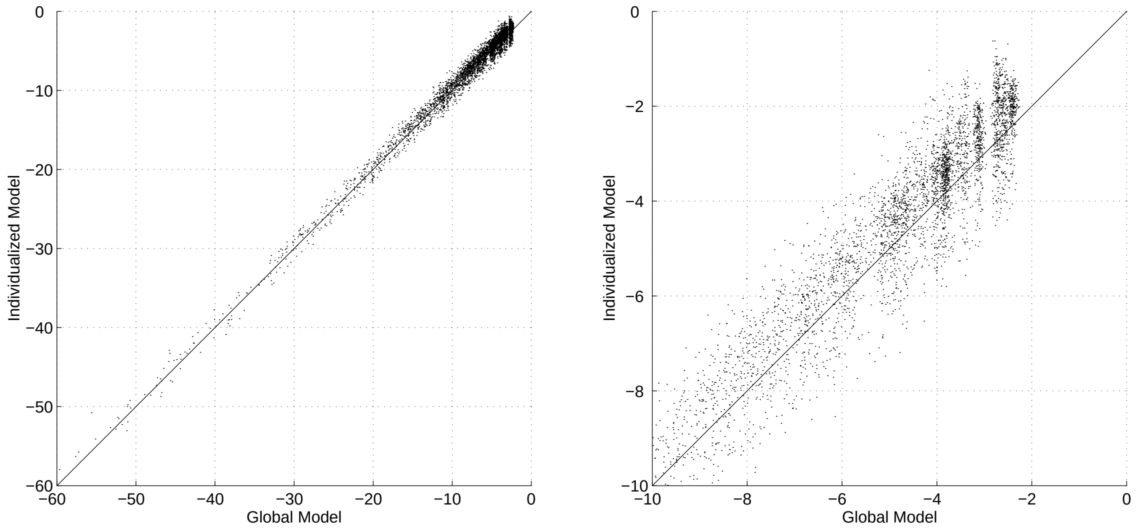

Figure 6: Retail Clothing Data: Scatter plots of the log probability scores for 5000 randomly selected out-of-sample transactions, for individuals with at least 10 transactions in the training data, $K = 20$ models, plotting log probability scores for the individual-specific model versus log probability scores for the global weights model. Left: all of the data, Right: close up detail of a fraction of the data.

30%. The first training/test pair consists of individuals who have 10 or more transactions during the training period. For this set the training data contains data on 2,941 individuals, 49,388 transactions, and 135,543 separate items purchased. The test data consists of 2,471 individuals (since some individuals in the training data did not make any purchases during the test period), 15,638 transactions, and 42,577 items purchased.

The second training/test pair consists of individuals who have 2 or more transactions during the training period. The training data has 56,054 individuals, 217,327 transactions, and 587,439 items. The test data consists of 20,123 individuals, 54,490 transactions, and 148,828 items.

Figure 5 compares the out-of-sample predictive entropy scores as a function of the number of mixture components K for (a) the mixture-based global weight model (where all individuals are assigned the same marginal mixture weights), (b) the mixture-based individual-weights model and (c) the non-mixture MAP histogram method (as a baseline). The two mixture-based approaches can be seen to outperform the non-mixture MAP histogram approach over almost all values of K. The mixture-based individual weights method consistently provides the most accurate predictions, providing a 12% decrease in predictive entropy compared to the MAP histogram method, and a roughly 1 to 3% decrease compared to non-individualized global mixture weights.

The two mixture models generally provide improved predictions as K is increased from $K = 1$ to $K = 40$ components, after which the curves tend to flatten out and eventually show some evidence of overfitting (higher entropy) as K increases.

Not surprisingly, predictions at level 1 with 53 categories (the two plots on the left) have much lower entropy than predictions at level 2 with 409 categories (the two plots on the right).

Somewhat less intuitive is the observed difference between the top two plots and the bottom two plots where the models were trained and tested on individuals with 10 or more, and 2 or more, transactions respectively. Predictions on the “2 or more” group are systematically better (lower entropy) than for the “10 or more” group. One might in fact have expected the opposite to occur, i.e., that the predictions would be better for individuals for whom more historical data is available. A possible reason that this does not happen is that the “2 or more” group contains a significant subset of individuals that are more predictable on average than the typical “10 or more group.” In other words infrequent shoppers show less variability in their purchases. For example, an individual might buy only very specific items from the store but do most of his or her shopping elsewhere. Further investigation of this difference between the two groups provided some empirical support (not shown here) that the “2 or more” dataset may contain a subset of more predictable individuals, for example, individuals who shop relatively rarely but always purchase the same items.

Figure 6 shows a more detailed scatter plot comparison of the difference between individual weight predictions and global weight predictions. Here K is fixed at 20 components and each data point represents the log-likelihood for a particular transaction (a basket) in the test data set. Generally speaking the individualized weights provide better predictions on many of the transactions, but not on all (some of the predictions are below the diagonal). While most transactions are grouped in the upper right corner of each plot, there are also some transactions where both models provide very low-likelihood scores (lower left corner of each plot), corresponding to either large baskets (multiplying many probabilities together), or low-probability events relative to the model, or a combination of both. The detailed plot on the right shows an interesting effect of the model, namely that the highest possible likelihood score of the model is bounded above by the nature of the underlying mixture model. More specifically, the highest possible probability predictions for each transaction are constrained by both the maximum component weight being used in the predictive profile for the individual who generated the transaction, and the maximum probability value in any component. This is reflected in the “hard cut-off” in logp scores that is evident in the upper-right corner of figure 6.

The upper plot of Figure 7 shows a set of bar charts for the estimated multinomial mixture components at level 1, for the “10 or more” group, with $K = 20$. In the lower part of the figure, we plot the same set of clusters after using multidimensional scaling (MDS) to determine a two-dimensional representation for each of the clusters. The MDS solution was obtained by minimizing the sum of squares of the distances in the two-dimensional solution relative to a proximity matrix of L1 (absolute) pairwise distances between the multinomial histograms.

The individual component multinomials (top plot) illustrate that most of the mixture components are “tuned” to a relatively small set of categories. A few components such as 9, 13, 17, and 19, have a single dominant category. Almost all components, with the exception of component 12, involve items in men’s categories or in women’s categories, but not both (recall that categories numbered below 25 are primarily men’s clothes and those above 25 are primarily women’s clothes). This is further reflected in the MDS plot where the cluster of components on the left represents men’s categories, the cluster on the right represents women’s categories, and component 12 is in the middle. Due to the proprietary nature of the data we cannot provide full details on the names of the individual categories. However, the components are quite intuitive: for example, the two largest probabilities for one component correspond to dress shirts and neckwear, while the three

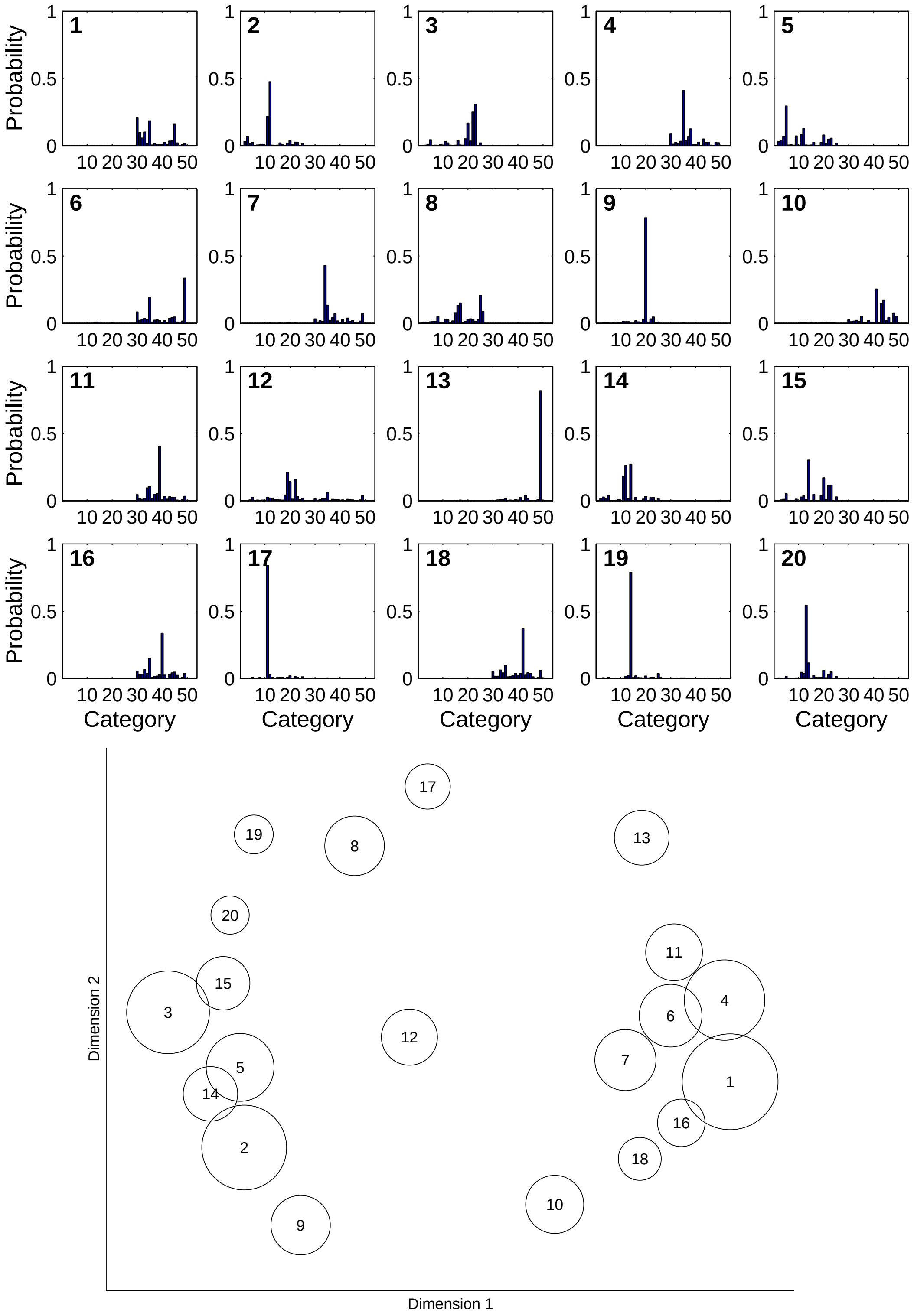

Figure 7: Retail Clothing Data: (a) $K = 20$ multinomial mixture components fitted to the “10 or more” group, (b) a 2-dimensional MDS plot of the same 20 components. The area of each circle is proportional to the mixture weight for that component.

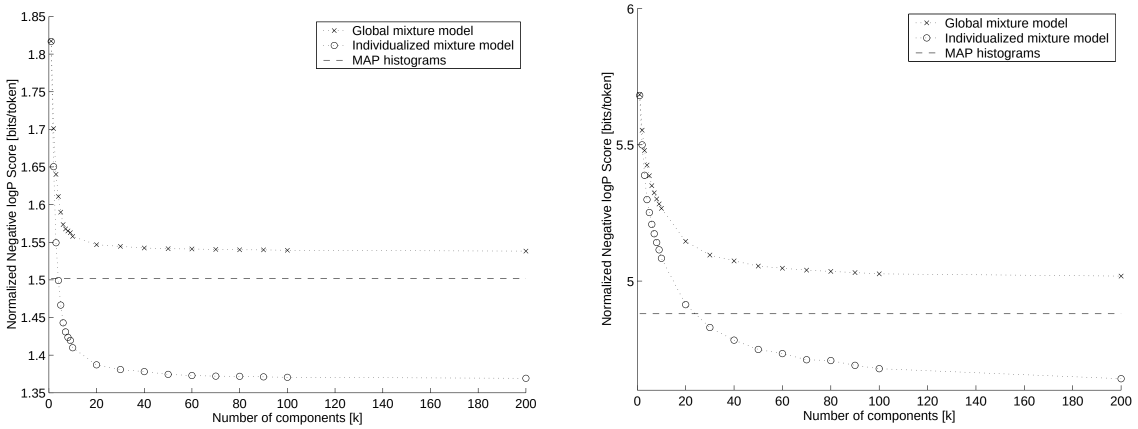

Figure 8: Retail Drugstore Data: plot of the negative log probability scores per item (predictive entropy) as a function of K, the number of mixture components, using individuals with at least 10 transactions in the training data with predictions at the level of 21 department categories (left) and at the second level of 151 categories (right).

largest probabilities for another component correspond to sport shirts, slacks, and active wear.

6.3 Experimental Results on the Retail Drugstore Data

For this data set the training data period was defined as the first 50% of transactions (chronologically) and the test data period consisted of the remaining 50%. As with the retail clothing data, individual-level models were constructed for individuals with 10 or more transactions during the training period. This resulted in 78,197 individuals in the training data, with 1.90 million transactions, and 6.01 million items. The test data contained 69,549 individuals, with 1.51 million transactions, and 4.78 million items. For computational reasons we focused on these frequent customers (10 or more transactions) since forecasting their behavior is likely to be of primary interest in a practical retail setting and fitting individuals with 2 or more transactions (for example) would have taken an inordinate amount of time to carry out over a range of values of K for a data set of this size.

Even restricting attention to the "10 or more" data results in a very large data set of 6 million items. To make the model training more computationally tractable we ran the initialization procedure (to find the global multinomial components and mixture weights used to initialize the second stage of the EM) using only a 10% random sample of transactions in the full data set. We then used all of each individual's transaction data in the full EM procedure to determine the individual weights. All of the results below for the drugstore data were obtained in this manner. We expect that the global component structure will be relatively robust to such sampling and we provide specific evidence of this later in Section 6.6.

Figure 8 compares the predictive entropy scores on the test data, as a function of the number of mixture components K. Again, the models evaluated are the mixture-based global weights (where all individuals are assigned the same marginal mixture weights), the mixture-based individual

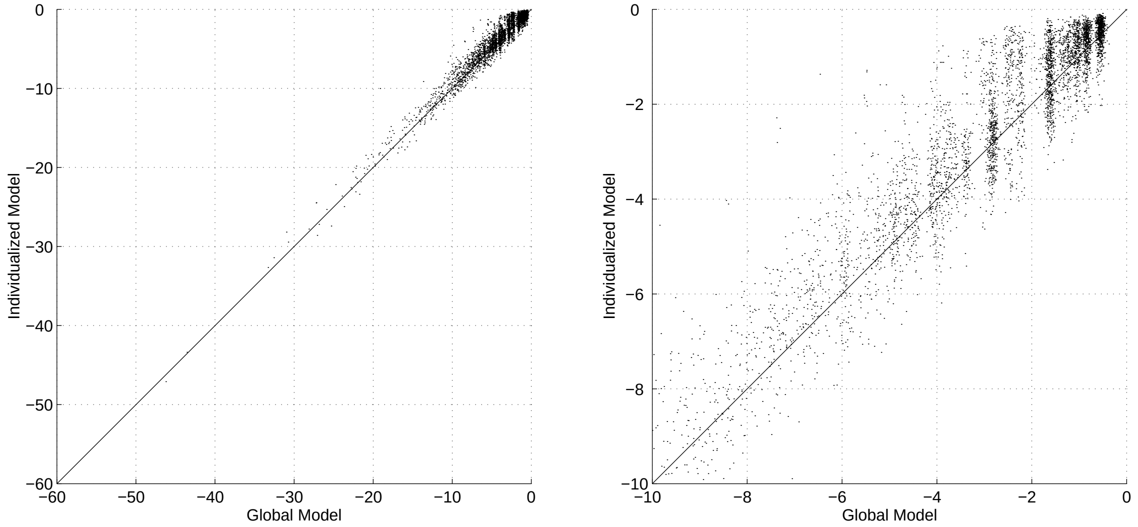

Figure 9: Retail Drugstore Data: scatter plots of the log probability scores on for 5000 randomly selected out-of-sample transactions, for individuals with at least 10 transactions in the training data, $K = 20$ models, plotting log probability scores for the individual-specific model versus log probability scores for the global weights model. Left: all of the data, Right: close up detail of a fraction of the data.

weights, and the baseline non-mixture MAP histogram method.

For this data the non-mixture MAP histogram is more competitive than for the clothing data set. For both levels of aggregration (left and right plots) the MAP histogram is actually more accurate than the most accurate of the global weights mixture model. The individual weights mixture model is still the most accurate model of all models for large values of K, on both level 1 and level 2 data. It is interesting that on this data the mixture models do not appear to overfit as a function of K, and the out-of-sample predictive entropy score is still decreasing up to $K = 200$. This may be due to the fact that this data set is much larger than the retail clothing data set (in Figure 5, where overfitting is evident) which had 135,543 transactions for training compared to 1.9 million for training here.

The logp scores for the drugstore data set are systematically lower than those obtained for the clothing data set in Figure 5. This difference can largely be explained by the fact that clothing purchases generally have higher entropy across categories than drugstore purchases, e.g., approximately 5 bits for clothing versus about 2 bits for drugstore for level 1 categories. If we divide the logp scores in Figure 5 and Figure 8 by the relevant entropies, the resulting normalized scores for the best of the individual-specific models fall in the range of 75 to 90% of the full entropy, across both data sets and levels. Thus, the main differences in predictability across the two data sets appears to be explained by an overall difference in entropy across categories in the two data sets.

Figure 9 shows scatter plots of individual weight predictions versus global weight predictions for transactions in the test data set. As with the clothing data, the predictions using individualized weights can be seen to be systematically more accurate for specific transactions (although not universally more accurate).

Figure 10 shows bar charts for the estimated multinomial mixture components at level 2, for the "10 or more" group, with $K = 20$. We see again that each of the multinomial components tend to be "tuned" to a relatively small set of categories. In the lower part of the figure, we again show a two-dimensional MDS plot generated in the same manner as for the clothing data components earlier in Figure 7.

6.4 Interpretation of Multinomial Components

Table 2 describes the estimated component models at level 2 for the drugstore retail data. We only list high-frequency items (or categories) that have lift ratios significantly greater than 1. A lift ratio is defined as the probability of an item being purchased given the component compared to the probability of an item being purchased in general. The components reveal some interesting structure in the data. First, the components are intuitive—similar items tend to appear together in the same component. For example, the first cluster can be characterized as generic medicine, clusters 2 and 9 as feminine personal care products, cluster 3 as household products, and cluster 4 as baby products. This suggests that baskets are actually less heterogeneous than one might expect. One explanation is that in Japan, unlike their US counterparts, retail drugstores are usually smaller in size and often located in neighborhoods that are within walking distances of customer's homes. Shoppers pay short and frequent visits, each time simply buying targeted items from a few categories. Thus, a typical basket tends to concentrate on similar items.

The multinomial mixture model detects dependencies among items that might not otherwise be apparent. For example, for component 1 (vitamins, stomach medicine, etc.), there are about 195,000 baskets in the training data that are most likely to have come from this component according to the model. Individual high-lift items shown in Table 2 (such as vitamins and stomach medicine) are about 4 to 7 times more likely to be present in this set of baskets than in the rest of the baskets. Pairs of such items are even more likely to be present. For example, one is 18 times more likely to find both vitamins and stomach medicine in a random basket from component 1, than in a random basket chosen from the remainder of the baskets. It is interesting to note that such dependencies would not be uncovered by association rules. For example, the conditional probability (or "confidence" in association rule terminology) of stomach medicine given vitamins is only about 0.05 across all baskets. This is rather low and would be well below the typical "confidence thresholds" used in association rule analyses.

In fact many such conditional probabilities relating pairs of items are quite low in this data set. This is due to the confounding effect of the relatively large percentage of baskets that only have a single item in them (35% overall). This "single item effect" leads to a negative correlation between virtually all pairs of items in the data (since purchasing an item lowers the probability of any additional item being purchased). In turn this makes dependency detection rather difficult if based directly on conditional probabilities.

The mixture model approach tends to construct multinomial components that group together items that co-occur with each other, even if such pairings are not common. Component 19 provides an interesting example. It consists of 31,000 baskets and suggests an association between alchohol,

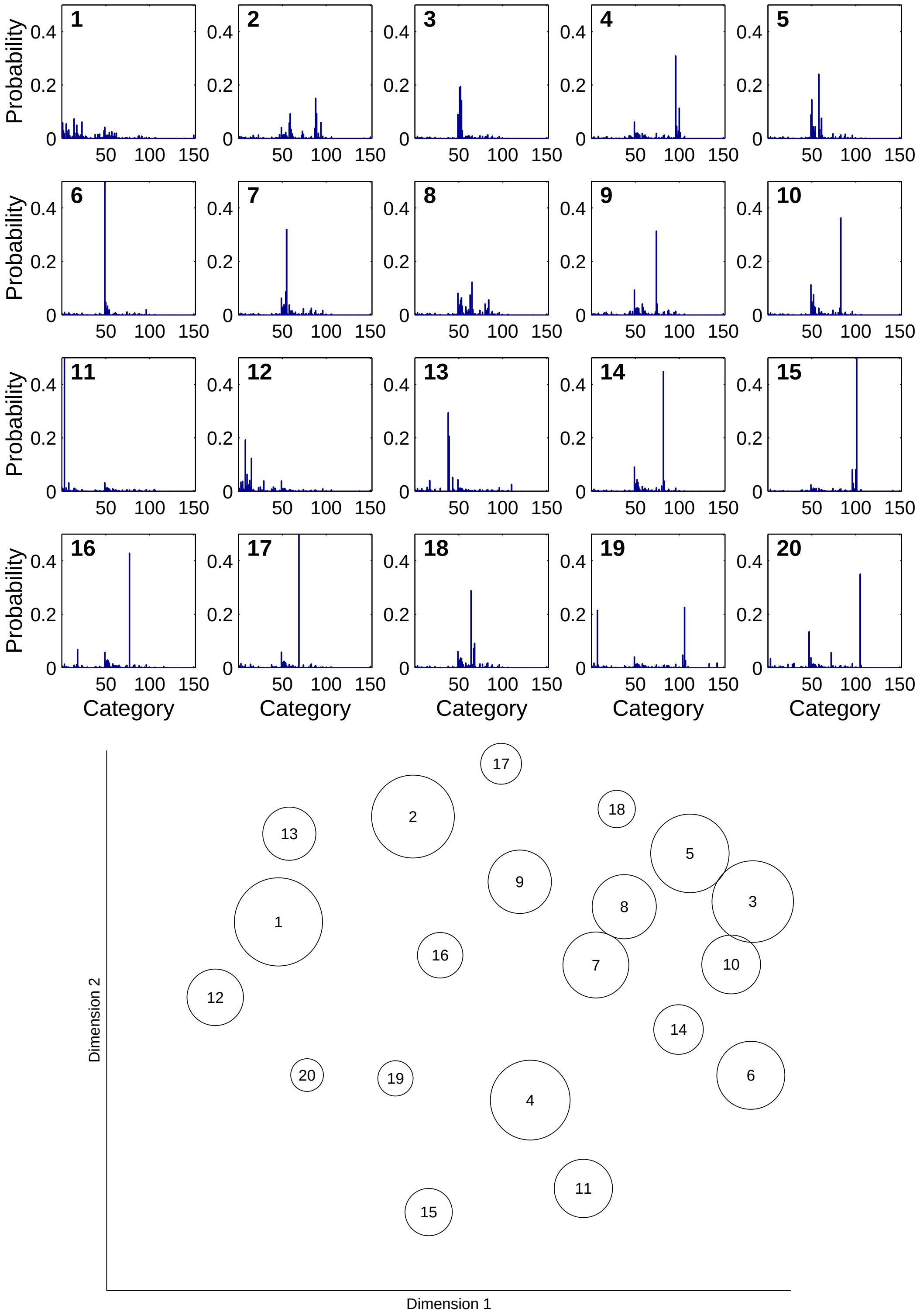

Figure 10: Retail Drugstore Data: (a) $K = 20$ multinomial mixture components fitted to the "10 or more" group, (b) a 2-dimensional MDS plot of the same 20 components. The area of each circle is proportional to the mixture weight for that component.

Table 2: High lift items and associated probabilities for the $K = 20$ components (clusters) in the drugstore data, level 2.

Cluster k=1, Weight=0.102 (194760 Baskets)

Item	P(Item|k)	P(Item)	Lift
External Pain Killer	0.0744	0.0110	6.8
Eye Drops	0.0627	0.0111	5.6
Vitamins	0.0588	0.0088	6.7
Stomach and Intestinal Medicine	0.0553	0.0090	6.2
Skin Medicine	0.0503	0.0111	4.5

Cluster k=2, Weight=0.090 (171222 Baskets)

Item	P(Item|k)	P(Item)	Lift
Basic Cosmetics	0.1513	0.0224	6.8
Makeup Cosmetics	0.0940	0.0103	9.1
Hair Care Products	0.0929	0.0190	4.9
Cosmetics Products	0.0603	0.0086	7.0

Cluster k=3, Weight=0.087 (166125 Baskets)

Item	P(Item|k)	P(Item)	Lift
Kitchen Cleaner	0.1956	0.0468	4.2
Fabric Softener	0.1910	0.0392	4.9
House Cleaner	0.1425	0.0354	4.0
Laundry Detergent	0.0803	0.0389	2.1

Cluster k=4, Weight=0.084 (159132 Baskets)

Item	P(Item|k)	P(Item)	Lift
Baby Diaper	0.3097	0.0360	8.6
Milk Powder	0.1144	0.0130	8.8
Lactation and Weaning Products	0.0467	0.0050	9.4
Baby Skin Care Products	0.0288	0.0044	6.6

Cluster k=5, Weight=0.081 (154382 Baskets)

Item	P(Item|k)	P(Item)	Lift
Shampoo and Conditioner	0.2413	0.0406	5.9
Laundry Detergent	0.1461	0.0389	3.8
Soap	0.0762	0.0174	4.4

Cluster k=6, Weight=0.061 (115911 Baskets)

Item	P(Item|k)	P(Item)	Lift
Paper Products	0.6183	0.0929	6.7

Cluster k=7, Weight=0.057 (108696 Baskets)

Item	P(Item|k)	P(Item)	Lift
Toothbrush	0.3200	0.0327	9.8
Toothpaste	0.0873	0.0282	3.1

Cluster k=8, Weight=0.054 (102972 Baskets)

Item	P(Item|k)	P(Item)	Lift
Air Freshener	0.1235	0.0142	8.7
Deodorizer	0.0756	0.0089	8.5
Table Seasoning	0.0572	0.0055	10.4

Cluster k=9, Weight=0.053 (100269 Baskets)

Item	P(Item|k)	P(Item)	Lift
Sanitary Napkins	0.3142	0.0286	11.0
Tampons	0.0415	0.0029	14.1

Cluster k=10, Weight=0.045 (86276 Baskets)

Item	P(Item|k)	P(Item)	Lift
Food Wrappers	0.3641	0.0345	10.6

Cluster k=11, Weight=0.045 (85330 Baskets)

Item	P(Item|k)	P(Item)	Lift
Bottled Strengthening Drinks	0.6075	0.0363	16.8
Cold Medicine	0.0325	0.0147	2.2

Cluster k=12, Weight=0.042 (79862 Baskets)

Item	P(Item|k)	P(Item)	Lift
Cold Medicine	0.1930	0.0147	13.1
Internal Pain Killer	0.1244	0.0087	14.3
Cough Medicine	0.0643	0.0036	17.8
Throat Drops	0.0417	0.0027	15.7
Regulated Medicines	0.0399	0.0034	11.7

Cluster k=13, Weight=0.037 (70780 Baskets)

Item	P(Item|k)	P(Item)	Lift
Light Medical Treatment Products	0.2945	0.0156	18.9
Nursing Care Suppl.	0.2069	0.0096	21.5
Bandages	0.0526	0.0079	6.6
Skin Medicine	0.0408	0.0111	3.7

Cluster k=14, Weight=0.033 (61848 Baskets)

Item	P(Item|k)	P(Item)	Lift
Cleaning Tools	0.4491	0.0319	14.1

Cluster k=15, Weight=0.029 (56060 Baskets)

Item	P(Item|k)	P(Item)	Lift
Baby Food	0.5105	0.0241	21.2
Baby Diaper	0.0822	0.0360	2.3
Milk Powder	0.0818	0.0130	6.3
Lactation and Weaning Products	0.0301	0.0050	6.1

Cluster k=16, Weight=0.027 (51802 Baskets)

Item	P(Item|k)	P(Item)	Lift
Insecticides	0.4291	0.0163	26.3
Anti-itch Cream for mosquito-bites	0.0690	0.0063	11.0

Cluster k=17, Weight=0.022 (42083 Baskets)

Item	P(Item|k)	P(Item)	Lift
Body Warmers	0.5204	0.0170	30.7

Cluster k=18, Weight=0.018 (34849 Baskets)

Item	P(Item|k)	P(Item)	Lift
Mothballs	0.2895	0.0093	31.2
Kitchen Gloves	0.0917	0.0047	19.5
Dehumidifiers	0.0729	0.0027	26.7

Cluster k=19, Weight=0.017 (31477 Baskets)

Item	P(Item|k)	P(Item)	Lift
Snacks	0.2271	0.0077	29.6
Constipation Medic.	0.2157	0.0070	31.0
Alcoholic Drinks	0.0484	0.0008	58.6
Processed Foods	0.0276	0.0012	22.6

Cluster k=20, Weight=0.014 (26907 Baskets)

Item	P(Item|k)	P(Item)	Lift
Beverages	0.3511	0.0075	46.7
Contact Lens Cleans.	0.1357	0.0066	20.5
Shaving Products	0.0576	0.0045	12.8

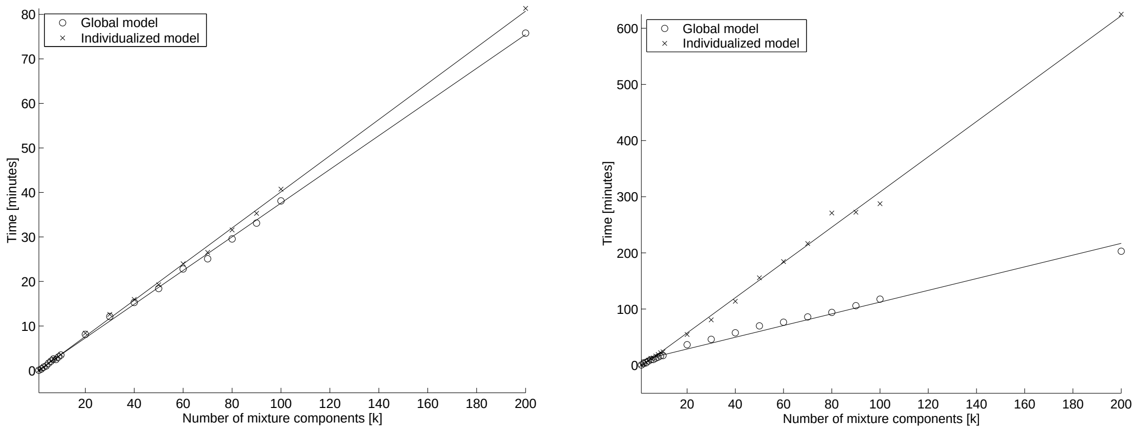

Figure 11: Plot of the CPU time to fit the global and individual weight mixture models, as a function of model complexity (number of components K), with a linear fit superposed on the time measurements (left: retail clothing data, right: retail drugstore data).

junk food (snack foods and processed foods), and constipation medicine[1]. The conditional probabilities relating these pairs of these items are again quite low due to the single-item basket effect. For example, the conditional probability of constipation medicine given processed food is only 0.03. Nonetheless, one is 140 times more likely to find both of these items in a randomly chosen basket that belongs to this component, than in a random basket not in the component, i.e., they co-occur much more frequently among baskets from this component.

It also turns out that only about 1500 baskets in total contain alcohol out of the 1.9 million baskets in the training data, and about 1460 of these baskets are most likely to belong to component 19 given the model. Thus, alcohol is very rarely purchased, but if it is purchased then the basket containing it is highly likely to be from component 19 in this model. Furthermore, if an individual has purchased alcohol, the empirical evidence suggests that they are 20 times more likely to have purchased processed foods, than if they did not purchase alcohol. This suggests a significant dependence between the two items.

In “drilling down” to examine the composition of baskets assigned to various components we have systematically found many subtle dependencies of this nature that lend further interpretability to the item groupings in Table 2. Analyzing basket structure at the component level provides insights that might not be realized by untargeted mining across the entire sample. Thus, the latent component variables serve to focus attention on key structural dependencies that would otherwise not be noticed.

6.5 Scalability Experiments

We conducted some simple experiments to determine how the methodology scales up computationally as a function of model complexity. We recorded the CPU time for the EM algorithm as a

[1] This might serve as a possible replacement for the widely quoted (but thought to be apocryphal) association rule linking beer and diapers!

function of the number of components K in the mixture model. The experiments were carried out on a Pentium III 1 Ghz machine with 512MB RAM (no paging). The CPU times for estimation of global ($\hat{\boldsymbol{\alpha}}$) and individual weights ($\hat{\boldsymbol{\alpha}}_i$) were both recorded. The estimation methodology was the same as prescribed in table 1 in section 4. The global weights and components correspond to parameters estimated during the initialization procedure, and individual weights and components correspond to those obtained after convergence of the full EM procedure.

While the time-complexity per iteration of EM is linear in K, one might anticipate that EM would require more iterations to converge as K increases, due to the possibility of more overlap among the component models and/or local maxima in the likelihood surface. Figure 11 shows that (for these data sets at least) this does not happen in practice and that the total convergence time of EM remains roughly linear in K. Note that there is a constant multiplicative factor in computation time between the global and individual weight methods. In the individual weight method we use the parameters of the global model as the starting point and learn individualized weights while slightly adjusting the multinomial mixture components. The extra computation required by the second phase ("full EM") appears empirically to be a relatively constant fraction of the time taken by the first stage to learn the global component structure. Note that in learning the global component structure (the initialization procedure of table 1) we use 20 random starts, while for the full EM procedure we only use a single start.

Note also that for the figure on the right, the drugstore data, the global component model is trained on only a 10% sample of transactions in the full data set. Thus, in this case, the additional time to estimate the individual weight model (the vertical distance to the line above it on the graph) is relatively larger than for the figure on the left, since in the full EM procedure the full data set is being used to fit the model (10 times more data than was used to fit the global model).

6.6 Estimation Methodology: Sensitivity Experiments

In this section we investigate the effect on predictive accuracy of (a) using a random sample rather than the full data set in the initialization procedure, and (b) using a "single-phase" EM procedure without our initialization phase where we do not use the global model to initialize the individual-specific model.

6.6.1 Initialization on Random Samples versus the Full Data

We examined the sensitivity of the individual weight estimation procedure when the initialization procedure (consisting of estimation of the global mixture model) is based on a 10% random sample of the full data set, compared to using all of the data. We used the same general EM methodology as used throughout the paper (summarized in Table 1), the only difference being the amount of data used in the initialization stage.

We used the clothing data set for these experiments since it takes far less time to train than the much larger drugstore data. Figure 12 shows the out-of-sample predictive accuracy. The average accuracies from the random sample method are virtually indistinguishable from those obtained using the full data during initialization. This indicates that the global mixture components reflect large-scale structure in the data and are not particularly sensitive to the specific training data

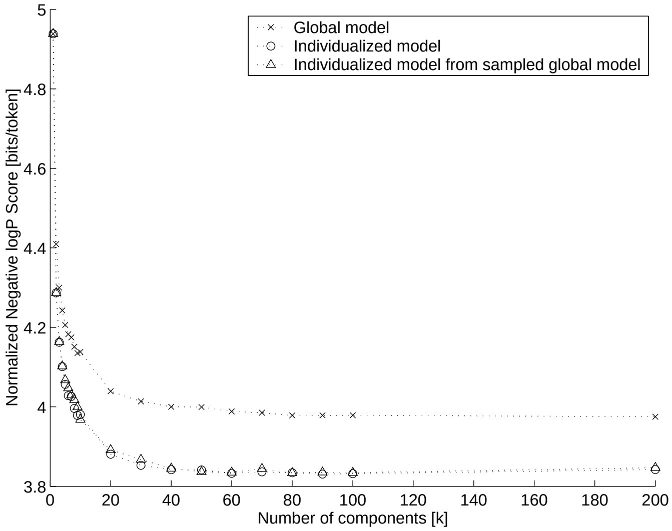

Figure 12: Predictive accuracy at level 1 on out-of-sample retail clothing data (as a function of K), using individuals with at least 10 transactions in the training data, comparing the proposed estimation procedure using all of the training data during the initialization phase with an approach that only uses 10% of the training data in the initialization phase.

sample used for estimation. This is useful from a practical viewpoint since fitting the global model during initialization can take a significant fraction of the overall time required for model estimation (e.g., see Figure 11). To put this in perspective, for 200 components for the drugstore data, the time difference in training is 3 hours for the random sampling method versus 30 hours for the full data method.

6.6.2 Single-Stage EM versus Two-Stage EM

The approximate empirical Bayes estimation procedure that we use could be replaced by a more direct "typical" single-stage EM procedure. Specifically, rather than first estimating data-driven priors $\boldsymbol{\xi}$ in the manner described earlier, we can use a "single-stage" EM approach that maximizes the objective function defined in Equation 6 treating the prior parameters $\boldsymbol{\xi}$ as additional parameters to be maximized over, and using a non-informative approach as in Table 1 for setting the multinomial component prior parameters $\boldsymbol{\gamma}$. In Figure 13 we compare both approaches in terms of predictive accuracy on the retail clothing data as a function of K. The single-stage EM procedure consists of running EM from 20 random starting values for the parameters and selecting the highest MAP solution, where both components and individualized weights are estimated. The two-stage EM procedure is the same as that described earlier in the paper in Table 1.

Our proposed procedure is seen to be just as accurate as the alternative single-stage procedure. In fact, for values of K greater than about 30, the two-stage procedure is more accurate out-of-sample than the single-stage approach.

Our two-stage approach has two distinct practical advantages over the simultaneous approach.

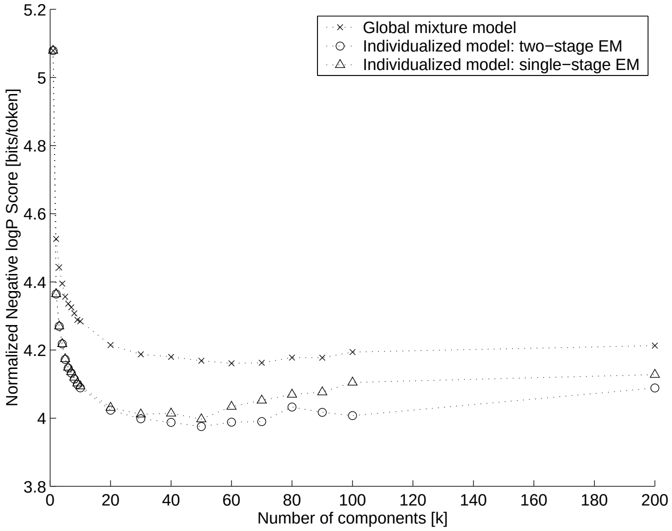

Figure 13: Predictive accuracy at level 1 on out-of-sample retail clothing data (as a function of K), using individuals with at least 10 transactions in the training data, comparing the proposed EM procedure (individualized mixture model) with a single-stage EM procedure.

First, from an estimation viewpoint, because of the size of the data set, the global structure can be precisely estimated (during initialization) on a subsample of the data with the resulting parameter estimates being relatively robust to noise in the sample (as shown earlier). Such subsampling, however, cannot be used in a single-stage EM procedure since all the individualized weights need to be updated at each EM iteration and, hence, each data point must be visited at each EM iteration.

Second, from a computational viewpoint, because the global structural component parameters are estimated in the first stage, future updating of individual weights of new customers can be highly efficient if one can afford to keep the global mixture components fixed. This would be the case, for example, if the global distribution did not change appreciably by adding new customers. In this case updating of the individual-specific model consists of updating only the individualized mixture weights. This can be performed independently of other individuals and is therefore extremely fast. This feature can be exploited to perform online updating given a stream of transactions and/or the arrival of new customers and is therefore particularly relevant to customer profiling in e-commerce environments, for example, where real-time profiling is often quite important. The single-stage methodology, on the other hand, would require rerunning estimation of both the global component parameters and the individual weights every time new customers or new transactions are added.

From these experiments we conclude that there is strong empirical evidence to suggest that the fitting of global components is relatively robust on these data sets. Hence, in practice, a reasonable and practical strategy is to first estimate the global components (perhaps on a sample of the data) and then to update individual weights as new data arrives relative to this fixed global set of mixture components. If non-stationarity is suspected (e.g., changes in products or pricing, new types of customers, etc.) the global mixture components can be periodically re-estimated (e.g., see Cadez

& Bradley, 2001).

7 Assessment of the Model

The model proposed in this paper addresses heterogeneity of individual purchasing behavior. It predicts which products an individual is likely to purchase, but not how many or when they will be purchased. Thus, there are several aspects of purchasing behavior that the proposed model does not directly address. A fully generative model would proceed for each individual by (a) generating transaction events at a certain rate over time, (b) generating the number of items for each transaction given that the event occurred, and (c) then predicting the distribution of individual categories given the total number of items. Our present model only addresses part (c) of this model.

In principle it is straightforward to add components to the present model to handle parts (b) and (c) because the generative process can be factored into conditional probabilistic components (i.e., (c) is conditioned on (a) and (b), (b) is conditioned on (a)). For example, to predict the number of items in a basket (part (b) above), a parametric (e.g., geometric) model could be applied at the individual level (different individuals purchase different numbers of items and are each allowed their own basket size model) or at the mixture component level (certain types of purchasing behavior lead to characteristic distributions in terms of basket size).

The rate at which transactions are generated (part (a) above) can also be modeled as a stationary event process (e.g., a homogeneous Poisson process). Predictable time-varying effects such as seasonality could be handled by adding a non-homogenous seasonal Poisson process (e.g., Cox & Isham, 1980) that governs the rate at which purchases are generated over a time interval. For retail purchasing such as clothing, there are often strong seasonal patterns present in the data, e.g., the purchase of warmer clothes in the fall and winter, vacation clothes in the spring and summer, and so forth. By modeling temporal aspects of the problem in this fashion, predictions from the model proposed in this paper would then be modulated by an overall "store visit rate" or "component purchase rate." In the experiments described in this paper the time interval for the test data extended over several months (7 months for the clothing data and 12 months for the drugstore data), and thus, seasonal effects are somewhat averaged out in time. In general, however, an appropriate seasonal model could allow the current predictive model to achieve more accurate forecasts out of sample.

So far we have assumed that customers are consistent in their shopping patterns. This assumption is at best a crude approximation since in practice customer behavior is often non-stationary in nature. For example, certain shoppers will make purchases at a relatively constant rate up to some time t and then no purchases are registered for the remainder of the duration for which we have observations. One might infer that such customers are in effect no longer "active" at this store. However, reliably detecting such inactivity is quite difficult given the sparsity of the data available on typical customers. Nonetheless, inference about a simple binary variable over time (to indicate active or inactive) would likely be quite useful in practice (Schmittlein, Morrison & Colombo, 1987), e.g., by adding Markov modulation to the Poisson rate process described above (e.g., see Scott, 1999). Adding a component to handle non-stationarity in customer purchase behavior would remedy a limitation of the current model—namely, results are only scored on the distribution of items purchased by any customer, but not the number, so the model is in effect not penalized if a customer purchases zero items in the future.

The mixture of multinomials transaction model could also be generalized to the mixture of conditional independence model (of which it is a special case), as discussed earlier. Furthermore, instead of assuming that each individual basket is being generated by a specific mixture component multinomial, a more realistic assumption might be that baskets are composed of mixtures of basic "behavioral components" in a manner similar to modeling documents as mixtures of basic topic components (e.g., Hoffman, 1999). For example, an individual might have a basket of items that reflects two "behaviors": purchasing of vacation clothes and purchasing of business clothes, all mixed in the same basket. We suspect that although such mixed baskets may be present in the data, that the majority of baskets are not of this form and will be more than adequately handled by the current model.

It should also be pointed out that in this paper we have only demonstrated the methodology on relatively low-dimensional problems (up to 500 categories), at least low-dimensional in terms of typical retail transaction data sets. As we descend the product hierarchy from departments all the way down to specific products (the so-called "SKU" level), there can be thousands of different items in a typical retail transaction database. It remains to be seen whether the type of probabilistic model proposed in this paper can computationally be scaled to this level of granularity. We believe that the mixture models proposed here can indeed be extended to model the full product tree, all the way down to the leaves. The sparsity of the data, and the hierarchical nature of the problem, tends to suggest that hierarchical Bayesian approaches will play a natural role here, where again it seems likely that the probabilistic models we have used will tend to match well to a hierarchical product structure. We leave further discussion of this topic to future work.

Further extensions of the model can be achieved by incorporation of customer information if available (e.g., demographics, psychographics, and other behavioral attributes such as those collected from credit card information). Such data could be used to enhance parameter estimates in the existing model, in the same way that information is borrowed across baskets to enhance individual profiles. As an example, one could incorporate a regression model into the existing empirical Bayes framework for estimating individual weights, where the variables above act as independent covariates and the individual weight for each individual is the dependent variable.

In summary, we present an interpretable and tractable approach to modeling of transaction data, where we focus on certain aspects of purchasing behavior, while other important aspects such as purchasing rates are ignored. We believe that, nonetheless, our proposed model and experimental results represent important first steps in probabilistic modeling of large-scale transaction data.

8 Related Work

Transaction data has received considerable attention from data mining researchers, going back to the pioneering work of Agrawal, Imielenski and Swami (1993) on association rules. Association rules present a different approach to transaction data analysis, searching for "directional" rules in the form of conditional probabilities for the purchase of item A given the purchase of items B and C, etc. These rules in effect represent correlations (associations) between particular sets of items. Our work here complements that of association rules in that we develop an explicit probabilistic model for the full joint distribution, rather than sets of disjoint conditional and joint probabilities (which is one view of association rules, see Pavlov, Mannila & Smyth, 2000). Indeed, one can interpret the multinomial probability mixture components as representing sets of associations among items in a manner somewhat similar to that of association rules (by representing sets of items that co-occur frequently in the data, e.g., see Figures 3, 7, and 10, and Table 2), but where the associations are constrained to form a global coherent probability model rather than being represented simply as a set of rules.

For forecasting and prediction it can be argued that the model-based approach (such as that we propose here) is a more systematic framework. As discussed earlier we can in principle integrate time-dependent factors (e.g., seasonality, non-stationarity), covariate measurements on customers (e.g., knowledge of the customer's age, educational-level) and other such information, all in a relatively systematic fashion. We note also that association rule algorithms depend fairly critically on the data being relatively sparse (e.g., Bayardo, 1998). In contrast, the model-based approach proposed here should be relatively robust with respect to the degree of sparseness of the data.

Other approaches have also been proposed for clustering and exploratory analysis of transaction data, but typically within a non-probabilistic framework (e.g., Strehl and Ghosh, 2000).

In the statistical literature, the general idea of using finite mixtures as a flexible modeling approach for discrete and categorical data has been known for many years, particularly in the social sciences under the rubric of latent class analysis (Lazarsfeld & Henry, 1968; Bartholomew & Knott, 1999). Typically these methods are applied to relatively small and low-dimensional data sets. More recently there has been a resurgence of interest in mixtures of multinomials and mixtures of conditionally independent Bernoulli models for modeling high-dimensional document-term data in text analysis (McCallum, 1999; Hoffman, 1999; Vinokourov & Girolami, to appear).

In the marketing literature there have also been numerous relatively sophisticated applications of mixture models to retail data (see Wedel and Kamakura, 1998, for a review). Typically, however, the focus here is on the problem of *brand choice*, where one develops individual and population-level models for consumer behavior in terms of choosing between a relatively small number of brands (e.g., 10) for a specific product (e.g., coffee).

In several recent applications, a hierarchical Bayes framework has been developed to include mixtures of normals for distributions of individual price sensitivities (Allenby, Arora, & Ginter, 1998), and the simultaneous modeling of quantity purchase and choice (Arora, Allenby, & Ginter, 1998). In these applications either scanner panel data or survey data were used. However, the analyses were often limited to single specific products and not market baskets across multiple categories.

The work of Breese, Heckerman and Kadie (1998) and Heckerman et al. (2000) on probabilistic model-based collaborative filtering bears some similarities in spirit to the approach described in this paper except that we focus directly on the problem of individual profiles (i.e., we have explicit models for each individual in our framework).

Our work can be viewed as being an extension of this broad family of probabilistic and statistical modeling ideas to the specific case of transaction basket data, where we explicitly deal with the problem of making inferences about specific individuals and handling multiple transactions per individual.

9 Conclusions

In this paper we investigated the use of mixture models and approximate Bayesian estimation techniques for automatically inferring individual-level profiles from transaction data records. On two large real-world retail data sets the proposed framework consistently outperformed alternative approaches in terms of the accuracy of predictions on future unseen customer purchasing behavior. Furthermore, our proposed estimation approach appears to be quite scalable to large data sets. The methodology appears to provide an interpretable and practical framework for a variety of transaction data applications including exploratory data abalysis, personalization, and forecasting.

Acknowledgements

The work of IC was supported by a Microsoft Graduate Research Fellowship. The work of PS was supported by National Science Foundation award IRI-9703120, the National Institute of Standards and Technology, the Department of Energy, IBM Research, and Microsoft Research. The authors are also grateful to Ayman Farahat and John Muller for assistance in obtaining the retail clothing data set and to Pharma, Japan for providing the drugstore data set.

APPENDIX

A EM Methodology and Scoring Details

In this Appendix we derive the EM algorithm and the corresponding update equations for the individual-specific model, taking into account prior distributions on the parameters.

A.1 Notation

We expand the notation introduced in section 2 in order to more precisely describe the models and learning tasks. According to the independence assumptions, we can write the likelihood of the parameters for a given data set D as the probability of observing the data, under the individual-specific model:

$$P(D|\mathbf{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \sum_{k=1}^{K} \alpha_{ik} P(\mathbf{y}_{ij}|\boldsymbol{\theta}_k). \tag{8}$$

In this section we concentrate on the "final full EM" estimation of the individual-specific model (Table 1) since the EM algorithm for the initialization procedure is quite straightforward and well known in the literature, namely, EM-based MAP estimation of mixture models with multinomial components.

For the individual-specific model we use priors on individual weights and work within a general maximum a posteriori (MAP) framework. The objective function we are interested in maximizing is the posterior distribution of the mixture parameters $\mathbf{\Theta}$:

$$P(\mathbf{\Theta}|D) = \frac{P(D|\mathbf{\Theta})P(\mathbf{\Theta})}{P(D)} \propto P(D|\mathbf{\Theta})P(\mathbf{\Theta}), \tag{9}$$

where the parameters $\boldsymbol{\Theta}$ consist of all the mixture model parameters:

$$\begin{aligned}\boldsymbol{\Theta} &= \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}, \\ \boldsymbol{\theta}_k &= \{\theta_{k1}, \ldots, \theta_{kC}\}, \ \sum_{c=1}^{C} \theta_{kc} = 1, \\ \boldsymbol{\alpha}_i &= \{\alpha_{i1}, \ldots, \alpha_{iK}\}, \ \sum_{k=1}^{K} \alpha_{ik} = 1. \end{aligned} \tag{10}$$

The distribution we use to describe each basket $\mathbf{y}_{ij}$, as mentioned earlier, is a multinomial:

$$P(\mathbf{y}_{ij}|\boldsymbol{\theta}_k) \propto \prod_{c=1}^{C} \theta_{kc}^{n_{ijc}}, \tag{11}$$

where the constant of proportionality depends only on basket $\mathbf{y}_{ij}$, but not on the parameters $\boldsymbol{\theta}$ and is therefore omitted. The prior term $P(\boldsymbol{\Theta})$ in Equation 9 consists of a weight-prior and a multinomial parameter-prior. This can be decomposed as:

$$P(\boldsymbol{\Theta}) = \prod_{k=1}^{K} P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_{i=1}^{N} P(\boldsymbol{\alpha}_i|\boldsymbol{\xi}), \tag{12}$$

where we use Dirichlet priors with parameters $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$:

$$\begin{aligned} P(\boldsymbol{\alpha}_i|\boldsymbol{\xi}) &\propto \prod_{k=1}^{K} \alpha_{ik}^{\xi_k}, \\ P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) &\propto \prod_{c=1}^{C} \theta_{kc}^{\gamma_c}. \end{aligned} \tag{13}$$

For individualized weights the parameters of the Dirichlet prior are proportional to the estimates of global weights $\hat{\boldsymbol{\alpha}}$ from the initialization procedure:

$$\boldsymbol{\xi} = \boldsymbol{\xi}_{ess} \hat{\boldsymbol{\alpha}}, \tag{14}$$

where the scalar $\boldsymbol{\xi}_{ess}$ represents the so-called *equivalent sample size* (ESS). The value of ESS we use for the individual-weight prior in the experiments reported in this paper is $\boldsymbol{\xi}_{ess} = 5$. This is equivalent to specifying that each individual has 5 baskets *a priori*, each of which "belongs" to the global mixture components in proportions defined by the global weights $\hat{\boldsymbol{\alpha}}$.

We use a simple "flat" prior for the multinomial parameters where $\gamma_c = 10^{-5}$ for each $c = 1, \ldots C$. This particular prior smooths multinomial estimates away from zero probabilities (that drive the loglikelihood to $-\infty$), in essence leaving the data to largely determine the non-zero values of the estimated multinomial probabilities.

A.2 The MAP Optimization Framework

To derive the necessary equations used for obtaining MAP parameters, we write the log of Equation 9 in a fully expanded form:

$$\log P(\boldsymbol{\Theta}|D) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left[\sum_{k=1}^{K}\alpha_{ik}P(\mathbf{y}_{ij}|\boldsymbol{\theta}_k)\right] + \sum_{i=1}^{N}\log P(\boldsymbol{\alpha}_i|\boldsymbol{\xi}) + \sum_{k=1}^{K}\log P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}). \qquad (15)$$

We define the *class-posterior* distribution, $P_{ij,k}$ on each basket ij, as the probability that the basket was generated by the k-th mixture component given the data. If we denote by c_{ij} the component that generated basket $\mathbf{y}_{ij}$, we can write the class-posterior distribution as:

$$\begin{aligned} P_{ij,k} &= P(c_{ij}=k|\mathbf{y}_{ij}) = \frac{P(\mathbf{y}_{ij}|c_{ij}=k)P(c_{ij}=k)}{P(\mathbf{y}_{ij})} = \frac{\alpha_{ik}P(\mathbf{y}_{ij}|\boldsymbol{\theta}_k)}{\sum_{k'=1}^{K}\alpha_{ik'}P(\mathbf{y}_{ij}|\boldsymbol{\theta}_{k'})}, \\ \bar{P}_{ij,k} &= P_{ij,k}|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}, \end{aligned} \qquad (16)$$

where the lower equation represents the class-posterior calculated for a specified set of parameters $\bar{\boldsymbol{\Theta}}$. The primary quantity in using the EM algorithm for MAP parameter estimation is the so-called Q function that represents the expected value of the log-posterior (Equation 15) over the class-posterior distribution (Equation 16) using the "current" parameters $\bar{\boldsymbol{\Theta}}$:

$$Q(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\bar{P}_{ij,k}\log\left[\alpha_{ik}P(\mathbf{y}_{ij}|\boldsymbol{\theta}_k)\right] + \sum_{i=1}^{N}\log P(\boldsymbol{\alpha}_i|\boldsymbol{\xi}) + \sum_{k=1}^{K}\log P(\boldsymbol{\theta}_k|\boldsymbol{\gamma}). \qquad (17)$$

At each EM iteration the Q function is maximized with respect to the parameters $\boldsymbol{\Theta}$ using the current parameters $\bar{\boldsymbol{\Theta}}$. At the end of each iteration, a set of new optimal parameters $\boldsymbol{\Theta}$ becomes the current parameters $\bar{\boldsymbol{\Theta}}$ for the next iteration. To calculate the optimal parameters we maximize the Q function subject to the constraints that each of the weight and multinomial parameters sum to 1. In order to perform constrained maximization, Lagrange multipliers λ (one for each parameter constraint) are introduced. The estimating equations for individualized weights are as follows:

$$\begin{aligned} \frac{\partial}{\partial\alpha_{ik}}\left[Q(\boldsymbol{\Theta}|\bar{\boldsymbol{\Theta}}) - \lambda\sum_{k'=1}^{K}\alpha_{ik'}\right]_{\alpha_{ik}=\widehat{\alpha}_{ik}} &= 0, \\ \sum_{j=1}^{n_i}\bar{P}_{ij,k}\frac{1}{\widehat{\alpha}_{ik}} + \frac{\xi_k}{\widehat{\alpha}_{ik}} - \lambda &= 0, \end{aligned} \qquad (18)$$

from which it follows that

$$\lambda\widehat{\alpha}_{ik} = \sum_{j=1}^{n_i}\bar{P}_{ij,k} + \xi_k. \qquad (19)$$

Summing Equation 19 over k we obtain an expression for the Lagrange multiplier λ:

$$\lambda = \sum_{j=1}^{n_i}\sum_{k=1}^{K}\bar{P}_{ij,k} + \sum_{k=1}^{K}\xi_k = n_i + \boldsymbol{\xi}_{ess}. \qquad (20)$$

The last equality follows from the fact that $\sum_{k=1}^{K} \bar{P}_{ij,k} = 1$ and Equation 14. Upon substituting λ into Equation 19 and solving for individualized weights, we obtain the final update equation for individualized weights:

$$\widehat{\alpha}_{ik} = \frac{1}{n_i + \boldsymbol{\xi}_{ess}} \left[\sum_{j=1}^{n_i} \bar{P}_{ij,k} + \xi_k \right]. \tag{21}$$

Similarly, we can optimize the Q function with respect to the multinomial parameters $\boldsymbol{\theta}$:

$$\begin{aligned} \frac{\partial}{\partial \theta_{kc}} \left[Q(\boldsymbol{\Theta}|\bar{\boldsymbol{\Theta}}) - \lambda \sum_{c'=1}^{C} \theta_{kc'} \right]_{\theta_{kc} = \widehat{\theta}_{kc}} &= 0, \\ \sum_{i=1}^{N} \sum_{j=1}^{n_i} \bar{P}_{ij,k} \frac{n_{ijc}}{\widehat{\theta}_{kc}} + \frac{\gamma_c}{\widehat{\theta}_{kc}} - \lambda &= 0, \end{aligned} \tag{22}$$

which yields the following update equation for the multinomial parameters $\widehat{\boldsymbol{\theta}}$:

$$\widehat{\theta}_{kc} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} \bar{P}_{ij,k} n_{ijc} + \gamma_c}{\sum_{c'=1}^{C} \left[\sum_{i=1}^{N} \sum_{j=1}^{n_i} \bar{P}_{ij,k} n_{ijc} + \gamma_{c'} \right]}. \tag{23}$$

A.3 Scoring New Baskets

Suppose user i makes a purchase in the out-of-sample period. Under the individual-specific model, a new basket $\mathbf{y}$ is scored according to the following equation:

$$P(\mathbf{y}|\widehat{\boldsymbol{\Theta}}, D) = P(\mathbf{y}|\widehat{\boldsymbol{\Theta}}) = \sum_{k=1}^{K} \widehat{\alpha}_{ik} P(\mathbf{y}|\widehat{\boldsymbol{\theta}}_k), \tag{24}$$

Note that individualized weights $\widehat{\boldsymbol{\alpha}}_i$ for individual i are used for predictive scoring. If the new basket $\mathbf{y}$ is generated by an individual that has not been observed in the training dataset D, then individualized weights $\widehat{\boldsymbol{\alpha}}_i$ do not exist for that individual. This problem is solved by noting that the individual can be considered to have been present in the training dataset D but that his or her corresponding number of transactions was 0. For such an individual the estimate of the individualized weights $\widehat{\boldsymbol{\alpha}}_i$ is independent of the model parameters and can be obtained from Equation 21 as:

$$\widehat{\alpha}_{ik} = \frac{1}{0 + \boldsymbol{\xi}_{ess}} \xi_k = \hat{\alpha_k}, \tag{25}$$

leading to the following score:

$$P(\mathbf{y}|\widehat{\boldsymbol{\Theta}}, D) = P(\mathbf{y}|\widehat{\boldsymbol{\Theta}}) = \sum_{k=1}^{K} \alpha_k P(\mathbf{y}|\widehat{\boldsymbol{\theta}}_k) = \sum_{k=1}^{K} \frac{\xi_k}{\boldsymbol{\xi}_{ess}} P(\mathbf{y}|\widehat{\boldsymbol{\theta}}_k). \tag{26}$$

Here we explicitly replace the global weights $\boldsymbol{\alpha}$ by the weight-prior $\boldsymbol{\xi}$ in order to make the notation consistent.

References

Agrawal, R., Imielenski, T., and Swami, A. (1993) Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD'98)*, New York: ACM Press, pp. 207–216.

Allenby, G., Arora, N., and Ginter, J. (1998) On the heterogeneity of demand, *Journal of Marketing Research*, 35, pp. 384–389.

Arora, N., Allenby, G., Ginter, J. (1998) A hierarchical Bayes model of primary and second demand, *Marketing Science*, 17(1), pp. 29–44.

Bayardo, R. J. (1998) Efficiently mining long patterns from databases, In *Proceedings of the 1998 ACM-SIGMOD International Conference on Management of Data (SIGMOD'98)*, New York, NY: ACM Press, pp. 85–93.

Breese J.S., Heckerman D. and Kadie C. (1998) Empirical analysis of predictive algorithms for collaborative filtering, *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann.

Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., and Wets, G. (2000) A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model, *Proceedings of the ACM Seventh International Conference on Knowledge Discovery and Data Mining,* New York: ACM Press, pp. 300–304.

Bartholomew, D. J., and Knott, M. (1999), *Latent Variable Models and Factor Analysis*, London: Arnold.

Cadez, I. V., and Bradley P. (2001), Model Based Population Tracking and Automatic Detection of Distribution Changes, to appear in *Proceedings of the NIPS 2001.*

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2000) Visualization of navigation patterns on a Web site using model-based clustering, *Proceedings of the ACM Sixth International Conference on Knowledge Discovery and Data Mining*, New York, NY: ACM Press, pp. 280–284.

Cox, D. R., and Isham, V. (1980) *Point Processes*, Chapman and Hall.

Han, J., and Kamber, M. (2000), *Data Mining: Concepts and Techniques*, San Francisco, CA: Morgan Kaufmann.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2000) Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1, pp. 49–75.

Hoffmann, T. (1999) Probabilistic latent semantic indexing, *Proceedings of the ACM SIGIR Conference 1999*, New York: ACM Press, 50–57.

Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., Duri, S.S. (2001) Personalization of supermarket product recommendations, *Data Mining and Knowledge Discovery,* to appear.

Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis*, New York: Houghton Mifflin.

Lord, F. M. (1980) *Application of Item Response Theory to Practical Testing Problems*, New Jersey: Lawrence Erlbaum Associates.

Meila, M., Heckerman, D. (1998) An experimental comparison of several clustering and initialization methods, MS-TR-98-06, Microsoft Research, Redmond, WA.

Pavlov, D., Mannila, H., and Smyth, P. (2000) Probabilistic models for query approximation with large sparse binary data sets, in *Proceedings of the 2000 Uncertainty in AI Conference*, San Francisco, CA: Morgan Kaufmann, pp. 465–472.

McCallum, A. (1999) Multi-label text classification with a mixture model trained by EM, *AAAI'99 Workshop on Text Learning.*

McCallum, A. and Nigam, (1998) A comparison of event models for Naive Bayes text classification, *AAAI-98 Workshop on Learning for Text Categorization.*

Neyman, J., and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrika, 16(1), 1-32.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, and Riedl, J. (1994) GroupLens: An open architecture for collaborative filtering of netnews. *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work,* Chapel Hill, N.C.: ACM Press, pp. 175-186.

Schmittlein, D. C. , Morrison, D. G., and Colombo, R. (1987) Counting your customers: Who are they and what will they do next? *Management Science*, 33, pp. 1–24.

Scott, S. L. (1999) Bayesian analysis of a two state Markov modulated Poisson process, *Journal of Computational and Graphical Statistics*, 8 pp. 662–670.

Strehl, A. and J. Ghosh (2000) Value-based customer grouping from large retail datasets, *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Vol. 4057*, Orlando, pp. 33–42.

Vinokourov, A and Girolami, M., A probabilistic framework for the hierarchic organisation classification of document collections, *Journal of Intelligent Information Systems*, special issue on Automated Text Categorization, to appear.

Wedel, M. and Kamakura. W. A. (1998) *Market Segmentation: Conceptual and Methodological Foundations*, Kluwer.