Faculty Research

University of Michigan Business School

WORKING PAPER SERIES

Improving Online Product Recommendations By Including Nonrated Items

Yuanping Ying
University of Michigan Business School

Michel Wedel
University of Michigan Business School

Fred Feinberg
University of Michigan Business School

Working Paper 03-009

IMPROVING ONLINE PRODUCT RECOMMENDATIONS

BY INCLUDING NONRATED ITEMS

Yuanping Ying

Michel Wedel

Fred Feinberg

University of Michigan Business School
701 Tappan Street
Ann Arbor, MI 48109

Draft. Under review at *Journal of Marketing Research*. Please do not quote or distribute. All comments welcome.

The authors can be reached at yingyp@umich.edu, wedel@bus.umich.edu, feinf@umich.edu, or by ordinary mail at the above address. The authors wish to thank Peter Lenk for his helpful suggestions.

IMPROVING ONLINE PRODUCT RECOMMENDATIONS BY INCLUDING NONRATED ITEMS

Abstract

Product Recommendation Systems have emerged as backbones of some of the Internet economy's largest and most venerable firms. They rely on the 'power of many', using data on prior customers' preferences to generate suggestions for current customers, typically in real time. A key feature of the data used in recommendation systems is the exceptionally large proportion of missing values: most customers don't rate most items. Statistical methods underlying such systems have only recently been studied rigorously. To date, all models have assumed the missing rating data to be missing completely at random (MCAR), tacitly presuming that there is either no pattern to the missing data, or that any such patterns cannot be exploited to improve ratings quality. We formulate two models to investigate whether this is the case. For the EachMovie data widely used by prior studies, we find that missing data is strongly *non-ignorable*, and that ratings can be improved substantially by explicitly modeling them, in a model component that is distinct from that for the ratings themselves, yet integrated with it. Indeed, we find that, in a "new movies / new people" holdout sample, errors in predicted ratings can be reduced by as much as 15% simply by carefully modeling missing data. The models substantially improve over previously proposed recommendation models. Because our models also account for the ordinal nature of ratings data, consumer heterogeneity, and interactions between movie genres and subject descriptors such as age and gender, they offer a rich portrait not only of which items are rated well, but which are rated at all, and how these interact. We find, for example, that Classic movies, while seldom rated at all, are apparently rated highly, with an opposite pattern for Action movies, while these patterns differ by gender. We discuss implications for Marketers based on the model results, and suggest how the models can be implemented within existing recommendation systems.

IMPROVING ONLINE PRODUCT RECOMMENDATIONS BY INCLUDING NONRATED ITEMS

Introduction

Aiming to be "the Earth's most customer-centric company", Amazon.com is engaged in an ongoing effort to refine its online personalization capabilities. Chief among these is the ability to offer product recommendations based on individuals' prior behavior at their website. Using product recommendation software developed by Net Perceptions, Amazon.com compares the target customer's browsing and purchasing profile with those of other customers and uses other customers' product evaluations to suggest what the target customer may like. Its personalization capability, turning customer knowledge into product recommendations and ultimately into purchases, is key to the company's customer retention strategy.

Amazon is hardly alone in this endeavor. CDNow uses such a system for recommending recordings; Blockbuster makes personalized video recommendations to its customers; Netscape incorporates its "What's Related?" capability to evaluate commonalities in web site visiting behavior; and so on. An industry has emerged to offer turnkey solutions to businesses hoping to match customers with the 'right' product, leveraging the histories and product evaluations of prior customers. Key players in the industry include Alexa, WebTrends, LikeMinds, Triplehop, Epipany and NetPerceptions; as a case in point, the last's web site asks specifically "What if you knew the best products for each customer?"

This paper concerns the development of such so-called *Product Recommendation Systems*. Making personalized product recommendations has become an important goal for marketers, not only those operating through the newer electronic channels, but also those through traditional ones, for several reasons. First, by recognizing customer heterogeneity and capitalizing on

similarities in preference, Product Recommendation Systems enable companies to cater to individual preferences. Second, they allow marketers to recommend products to a target customer using information from other customers with similar product preferences, prior purchase histories and/or demographic profiles. Recommendation systems thereby enable marketers to act as word-of-mouth agents and virtually expand customers' social networks. Third, accurate product recommendations save customers search costs during the purchase decision process. This is important if customers face a large number of choice alternatives, particularly so if the product's attributes are hard to evaluate before consumption or not very helpful in anticipating the product's consumption utility. Examples include hedonic products such as music, movies, books, restaurants, tourism, and others for which there is a substantial experiential component. For such hedonic products, other customers' evaluations are particularly important, and customers actively welcome product recommendations based on others' experiences. Finally, an effective product recommendation system can help with customer acquisition and retention, as well as enhance customer loyalty, and so is an important tool in customer relationship management.

In this paper we develop a statistical framework for improving product recommendations, building on previous work by Ansari, Essegaier, and Kohli (2000). We view product recommendations as predictions of a target customer's latent consumption utility: products for which the customer has a high predicted utility should be recommended. In doing so, we address three problems that have not yet been resolved by current algorithms.

The first of these problems, concerning the nature of how products come *not* to be rated, is in our view the most serious. Much of the product rating data on which recommendation systems need rely is not only missing, but missing *non-randomly*. Consumers can only evaluate products

they've experienced, so they commonly rate only a very small subset of all available items. Consequently, considered relative to the entire product catalog an e-tailer offers, the ratings history of any particular consumer is minuscule. Many current recommendation systems overtly request that customers evaluate products or provide preferences, and such information has been found to be broadly effective in generating subsequent recommendations. For instance, Art Technology Group's Dynamo server requests customers' preference input, and has established clientele such as Blockbuster, J. Crew, and Target. Mostly customers are asked to supply evaluations for only a few products from among those they are familiar with, since evaluation of the complete set of alternatives is neither feasible nor desirable.

The chief task for recommendation systems is therefore to predict a target customer's evaluations using available ratings data from that customer and numerous others. In doing so, one should rightly consider the *causes* of the missing data pattern. Previous product recommendation-systems have been based on observed ratings only, and thus tacitly presume that the missing data are valueless, and can be safely ignored. This assumption is valid only if the missing evaluations are missing completely at random (see Rubin 1976). Thus, it is assumed that customers do not provide preference information on most of the products for reasons not systematically related to variables relevant to the study. This assumption is highly restrictive and in fact runs counter to intuition. For example, some consumers don't offer ratings simply because they have no consumption experience with the products in question; and, in turn, the very reason that consumers do not purchase or consume the products in question may be that they simply don't like them. Even if consumers do have consumption experience with the products, they're not necessarily willing to provide marketers with their evaluations. Some

consumers may only vote on products that they either very much like or dislike, to champion some while giving warning about others.

In sum, there is ample reason to believe that the missing evaluations are not, in fact, missing at random. It is well known that failing to accommodate non-ignorable missing data mechanisms can lead to biased estimates (Little and Rubin 1987), and thus in this case to sub-optimal or even erroneous recommendations. One goal of the present paper, therefore, is to explicitly model the missing data mechanism, alleviating the problem of biased estimates and thereby improving recommendation quality. Although we cannot explore what the specific causes underlying the missing data pattern might be, as we will see, the data speak plainly that they cannot be ignored.

The second, separate problem concerns the ordinal nature of most product evaluation data input to recommendation algorithms. Previous approaches which treat the evaluation data as either nominal or interval-scaled thereby fail to reflect the true data generation mechanism and thus potentially offer inconsistent forecasts. By contrast, we specifically exploit the ordinal nature of the prior preference data using a parsimonious rank-order binomial formulation (Rost 1985).

The third problem is the heterogeneity inherent in consumers' preference data. Although marketing researchers have used both finite mixture and hierarchical Bayes models to capture customer heterogeneity, several studies suggest that they offer roughly equal performance in a variety of empirical settings (Andrews, Ansari, and Currim 2002; Wedel et al. 1999). For the purpose of recommendation systems, mixture models offer the advantage of identifying groups of subjects that have similar preferences for the products in question, and thus identify subject-subject matches that enable direct recommendations to be made. In addition, in the context of

modeling missing recommendations, the mixture model may serve to pick up different patterns of missing ratings related to underlying causes, such as rating liked/disliked movies. However, mixture models are restrictive in treating all customers within a segment as interchangeable. By comparison, Hierarchical Bayes models offer a more parsimonious approach to modeling individual level heterogeneity, one that is especially useful in the context of one-to-one marketing approaches (Allenby and Rossi 1999). We therefore use continuous and discrete representations of customer heterogeneity, simultaneously.

In Section 2, we provide a review of past research and approaches to making individualized recommendations. In Section 3, we lay down the theoretical framework and delineate our model. Section 4 describes the empirical tests, where we reinvestigate the benchmarking EachMovie data set that has been used in other research on recommendation systems. We then compare our approach with several popular alternatives. The paper concludes with a discussion of future research directions.

Recommendation Systems Literature Review

Whereas recommendation systems have been a popular topic of study in the computer science and machine learning literature, only recently has the marketing literature made this a core topic of study, thanks to the pioneering work of Ansari, Essegaier, and Kohli (2000). Here, we focus on statistical recommendation algorithms, but for a complete overview of the history of recommendation systems, we refer to Ansari, Essegaier, and Kohli (2000).

For ease of exposition, we formulate the general research problem encountered in making product recommendations as predicting the missing entries in Table 1, which represents customers' ratings for a given set of products. Such ratings are typically made on pre-ordained

ordinal scales. For example, Amazon and Blockbuster each ask customers to award products 1-5 "stars", and the EachMovie data analyzed in Section 4 has each film rated as 0-5 "stars".

Insert Table 1 about Here

Existing recommendation methods can be categorized into two classes: heuristic methods and model-based methods. Heuristic methods (often clustering-type algorithms) are widely used in the computer science literature, where researchers have attempted to filter out irrelevant information from that available on the Internet. The popularity of these heuristic approaches stems from their ease of implementation. But, they are often *ad hoc* and have been shown to be broadly inferior to model-based methods (Breese, Heckerman, and Kadie 1998). Model-based methods invoke a probability distribution for customers' responses and therefore explicitly hypothesize a data generation process. Model-based methods that have been used to generate product recommendations include the mixture model (Chien and George 1999), the hierarchical Bayes model (Ansari, Essegaier, and Kohli 2000), factor analysis (Canny 2002), and both Bayesian and dependency network models (Breese, Heckerman, and Kadie 1998).

Next we will review the most important heuristic method ('nearest neighbor') and the two most important model-based methods (mixture and hierarchical Bayes models). We review the nearest-neighbor methods because they are the most commonly used in commercial recommendation systems, and the mixture and hierarchical Bayes model because they represent the most effective model-based methods that have been proposed and applied in marketing to date.

Nearest-Neighbor Methods

Commercial recommendation algorithms mostly employ nearest-neighbor methods, which identify a set of customers as neighbors of the target customer and predict their missing preference as a function, typically a weighted average, of neighbors' data. Neighbors are selected based on some measure of similarity. Early work, including Resnick et al. (1994), uses the Pearson correlation between two customers' rating vectors. Other similarity measures include the cosine of the angle between two rating vectors and related measures popular in the computer science literature (O'Connor and Herlocker 1999; Sarwar et al. 2000).

Nearest-neighbor methods are easy to implement but are often based on *ad hoc* assumptions and heuristics that lack statistical foundation. For example, similarity measures such as the Pearson correlation have been shown to have undesirable properties in the presence of sparse data (Lee 1999). And, as argued before, sparseness is a key characteristic of Internet customer preference data: even if the similarity between two customers' preferences is high, they may have purchased or stated their preferences toward only a few common products out of many on offer. In addition, nearest-neighbor methods do not consider customer characteristics or product attributes. Such attribute data on products in different categories may be used to better predict a target customer's liking for a particular product. Not including attribute data may yield recommendations that do not match the customer's preference well. Recommendations made by nearest-neighbor methods are thus not very reliable, as found by Breese, Heckerman, and Kadie (1998).

Mixture Models

The mixture model is a generic candidate for finding groups of customers with similar preferences. Chien and George (1999) were the first to propose a powerful recommendation system based on a Bayesian mixture model. Customers are assumed to come from S latent segments with mixing proportions $\{\pi_1, \ldots, \pi_S\}$. Conditional on the latent segment membership, customer i's rating for product j, Y_{ij}, follows a product-specific multinomial distribution, $p(Y_{ij} \mid s)$. Hence, the unconditional distribution of Y_{ij} is given by

$$p(Y_{ij}) = \sum_{s=1}^{S} \pi_s p(Y_{ij} \mid s) \tag{1}$$

Chien and George (1999)'s approach can be readily implemented, and they show it to outperform nearest-neighbor methods on the EachMovie data. But, despite the appeal of the mixture model to determine recommendation groups, it nevertheless fails to accommodate some critical features of the EachMovie customer preference data, several of which characterize other recommendation data sets as well.

Because much subsequent research has taken its cue from Chien and George's pioneering work, it is instructive to discuss its limitations, and how they may be overcome. First, although the customer preference data are ordinal, Chien and George (1999) treat it as nominal. While this allows for a great deal of flexibility – for example, in accommodating different 'patterns' of response – it cannot capitalize on the ordinal information intrinsic to the true data generation process and, as a practical concern, is not very parsimonious. Second, they do not include explanatory variables to help elucidate why customers may like or dislike a product, and thus both improve recommendations and clarify drivers of choice or satisfaction. Consequently their model also cannot predict customers' preference toward *new* products, about which detailed

attribute information is all one has to go on. Third, the mixture approach assumes that all customers in a particular segment have the same preference structure; while intuitively appealing, this can prove restrictive in practice. Fourth, they estimate a discrete heterogeneity distribution for each product, which requires a large number of parameters. Although estimation of a large quantity of parameters does not present any problems in a Bayesian context, it does require substantial computational time and may reduce holdout predictive validity, which is critical in making recommendations. Finally, as in other current recommendation systems, and most importantly from the perspective of the present paper, Chien and George (1999) assume that data are missing completely at random, and thus do not posit a mechanism for the (non-ignorable) missing data. As we will show, doing so pays great dividends in terms of predictive accuracy, and in gaining a proper substantive understanding of consumer preference data.

Hierarchical Bayes Model

Ansari, Essegaier, and Kohli (2000) provide the first model-based product recommendation system in Marketing. They propose an effective hierarchical Bayes model to predict customers' ratings for products, allowing for both fixed effects and random effects of customer characteristics and of product attributes. Specifically, let X_i denote customer i's characteristics and Z_j denote product j's attributes. Then customer i's rating for product j, Y_{ij}, is modeled as:

$$\begin{aligned} Y_{ij} &= \beta^X X_i + \beta^Z Z_j + CH_i + PH_j + e_{ij} \\ CH_i &= \lambda_{i1} + \lambda_{i2} Z_j \\ PH_j &= \gamma_{i1} + \gamma_{i2} X_i \end{aligned} \tag{2}$$

where $e_{ij} \sim N(0, \sigma^2)$, $\lambda \sim N(0, \Lambda)$, and $\gamma \sim N(0, \Gamma)$. Heterogeneity is captured by two model components: CH_i for Customers and PH_j for Products. Their model is also tested on the EachMovie data.

Although Ansari, Essegaier, and Kohli (2000)'s approach is elegant, there are a few caveats. Rather than mimicking the ordinal nature of customer preference data, customers' discrete product ratings are treated as interval scales. Then, by using a normal distribution of unobserved customer heterogeneity, the approach does not naturally identify recommendation groups, as is done in Chien and George's (1999) mixture model. Further, and most importantly for this study, they do not consider the non-ignorable data missing data in the preference rankings.

In summary, although the mixture and HB approaches have proven to be powerful model-based approaches to generate product recommendations, none of these has accommodated the three key features of customer preference data that we focus on in this study: non-ignorable missing data, the ordinal nature of ratings scales, and complete account of heterogeneity. In the next section, we present a unified approach which captures all these features, and may thus provide an improved platform for generating product recommendations.

The Recommendation Model

In line with prior research, we propose a product recommendation algorithm based on customer preference data. We assume customers' ratings for products across different product categories are elicited using an ordinal scale with K categories.

Denote the observed rating as Y_{ij}, where $Y_{ij} = k$ if customer i rates product j as k, with $k = 1,\ldots, K$. Customers differ in the number of products they rate, i.e., $j = 1,\ldots, J_i$, resulting in an unbalanced data set. The model admits of two covariate sets: information on customer characteristics X_i and on product attributes Z_j. Our objective is to predict: (1) how an existing customer would have rated an existing product that s/he has not rated; (2) an existing customer's

rating for a new product; (3) a new customer's rating for an existing product; (4) and a new customer's rating for a new product (see Ansari, Essegaier, and Kohli 2000).

Ordinal Data Model

To utilize the ordinal nature of Y_{ij}, we model the probability that customer i will rate product j at scale point k as

$$p\left(Y_{ij}=k\right)=\binom{K-1}{k-1}\theta_{ij}^{k-1}\left(1-\theta_{ij}\right)^{K-k} . \tag{3}$$

The above representation was first proposed by Rost (1985) to model ratings data. It is much more parsimonious than an unconstrained or rank-order multinomial specification because it requires only one parameter θ_{ij} for all K ratings for a customer/product combination. Besides, such parsimony is achieved without sacrifice of logical consistency. The reason that marketers use more than two categories when eliciting customers' product evaluations is to obtain more detailed information. The ultimate goal of analyzing customers' product evaluations data is not to quantify customer's tendency to choose *each* response category, but to quantify their *global* tendency to like the products. The single parameter θ_{ij} acts as a location parameter for modeling customer i's global tendency of liking product j and therefore has considerable conceptual appeal (Rost 1985).

Missing Data Mechanism

Missing product ratings are ubiquitous in data collected for recommendation systems. Previous research on the topic implicitly assumes that ratings are missing completely at random

(MCAR[1]). In reality, as noted previously, customers may not rate certain products for numerous reasons: lack of awareness or consumption experience; unwillingness to share ratings with marketers; a tendency to rate only strongly liked or disliked products; among others. Hence, the MCAR assumption is unlikely to provide a faithful description of the process by which the missing data are generated, and we thus relax it by explicitly incorporating a missing data mechanism in our model.

We introduce a missing data indicator matrix $\boldsymbol{M}$, where $M_{ij} = 1$ if customer i rates product j, $M_{ij} = 0$ otherwise. We account for the non-ignorable missing data mechanism by formulating a model[2] for the joint distribution of Y_{ij} and M_{ij}. Specifically,

$$p\left(Y_{ij}, M_{ij}\right) = p\left(Y_{ij} \mid M_{ij}\right) p\left(M_{ij}\right) . \tag{4}$$

Here, M_{ij} follows a Bernoulli distribution. In other words, there is certain probability, δ_{ij}, that customer i will rate product j:

$$p\left(M_{ij} = 1\right) = \delta_{ij} . \tag{5}$$

We call δ_{ij} the *selection* probability. Although no assumptions need be made about temporal ordering, one might well interpret this formulation as being consistent with a two-stage ratings provision process: first, customers decide whether to provide a product rating or not, according to (5); if the decision is affirmative, then a rating is provided, where the conditional distribution of Y_{ij} is given by (3).

---

[1] Here we use MCAR in the definition provided by Rubin (1976).

[2] Our specification of the joint distribution of Y_{ij} and M_{ij} is slightly different from selection models (Heckman 1979). Selection models specify $p(Y_{ij}, M_{ij}) = p(Y_{ij})p(M_{ij} \mid Y_{ij})$, which is simply an alternative way of factoring the joint distribution of Y_{ij} and M_{ij}. When the data are missing completely at random, i.e., Y_{ij} and M_{ij} are independent, the two specifications are equivalent. We posit the specification in (4) because we believe it more accurately reflects customers' latent decision process.

Heterogeneity

It is crucial to consider heterogeneity across customers. As detailed in the literature review above, there are two dominant ways to model customer heterogeneity: through a discrete process, or though a normal distribution, giving rise to finite mixture or hierarchical normal specifications of heterogeneity, respectively. Wedel et al. (1999) provide a discussion of both representations; Andrews, Ansari, and Currim (2002) presents extensive simulation studies comparing them. Each approach has strong conceptual appeal when applied to modeling recommendation systems. The mixture model approach first used by Chien and George (1999) enables one to find recommendation classes, while the hierarchical normal specification allows for the estimation of individual level parameters that are a powerful basis for one-to-one recommendations (Ansari, Essegaier, and Kohli 2000). Therefore, we will include both forms of heterogeneity in our model. In addition, we not only model heterogeneity in the ratings data in this manner, but account for heterogeneity in the selection process generating the missing product data as well.

We assume S latent segments of customers, with mixing proportions $\{\pi_1, \ldots, \pi_S\}$. Conditional on his/her latent segment membership s, customer i decides whether to rate product j and, if s/he does decide to rate, the rating follows the distribution specified in (3).

Thus, the conditional (on segment membership, s) joint distribution of Y_{ij} and M_{ij} is

$$
\begin{aligned}
p(Y_{ij}, M_{ij} \mid s) &= p(Y_{ij} \mid M_{ij}, s)\, p(M_{ij} \mid s) \\
&= \left[\binom{K-1}{Y_{ij}-1} \theta_{ijs}^{Y_{ij}-1} \left(1-\theta_{ijs}^{K-Y_{ij}}\right) \delta_{ijs} \right]^{M_{ij}} \left(1-\delta_{ijs}\right)^{1-M_{ij}}
\end{aligned} \tag{7}
$$

The *un*conditional joint distribution of Y_{ij} and M_{ij} is therefore simply a summation over segments:

(8) $$p(Y_{ij}, M_{ij}) = \sum_{s=1}^{S} \pi_s p(Y_{ij}, M_{ij} \mid s)$$

The continuous heterogeneity approach, using a hierarchical specification, will be described next, along with a discussion of how the effects of explanatory variables are incorporated.

Explanatory Variables

We take into consideration the possible effects of customer characteristics X_i and product attributes Z_j in two different ways. The reason for this division is that, while the direct form of the effect of product characteristics on the ratings is unambiguous, customer characteristics may be hypothesized to exert their influence on the ratings for the movies in two distinct ways. The first possible way is indirect, where the customer characteristics modify the effects of the product characteristics; for example, a particular type of customer may prefer, say, action movies. Put slightly differently, there may be certain *interaction* effects between customer and product characteristics. The second possible mechanism involves a *direct* effect of customer characteristics on ratings. The first results in a hierarchical setup given by

(9) $$\begin{pmatrix} \text{logit}(\theta_{ijs}) \\ \text{logit}(\delta_{ijs}) \end{pmatrix} = \alpha_{is}^{Z} Z_j$$
$$\alpha_{is}^{Z} \sim N(\beta_s X_i, \Sigma_s)$$

The second is a non-hierarchical formulation:

(10) $$\begin{pmatrix} \text{logit}(\theta_{ijs}) \\ \text{logit}(\delta_{ijs}) \end{pmatrix} = \alpha_{is}^{X} X_i + \alpha_{is}^{Z} Z_j$$
$$\begin{pmatrix} \alpha_{is}^{X} \\ \alpha_{is}^{Z} \end{pmatrix} \sim N(\mu_s, \Phi_s)$$

Note that both of these formulations include a continuous distribution of the effect parameters – α_{is}^{Z} or $\{\alpha_{is}^{X}, \alpha_{is}^{Z}\}$ – over customers, accounting for within-segment preference

heterogeneity. In these formulations we assume that Φ_s and Σ_s are diagonal[3]. Equations (9) and (10) differ in that, in (9), product characteristics directly affect product selection and preference ratings, the effects of these product characteristics being heterogeneous, and partially explained by customer characteristics. This is a fully hierarchical model setup, which has been used successfully in a number of marketing applications (Rossi, McCulloch, and Allenby 1995; Boatwright, McCulloch, and Rossi 1999). In (10), by contrast, the effects of customer characteristics are not hierarchical, and it is assumed that both product and customer characteristics affect the selection probabilities and product ratings directly and similarly. Although the hierarchical specification has received greater attention in prior Marketing literature, because we have no prior theory that favors one form over the other, we investigate them both empirically using the EachMovie data.

Estimation

We estimate the model using Markov Chain Monte Carlo (MCMC) Methods. We specify the following priors: the mixing proportions $\{\pi_1, \ldots, \pi_s\}$ follow a Dirichlet distribution; for the hierarchical setup in (9), β_s follows a segment-specific normal distribution; for the covariance matrix Σ_s, we estimate its inverse, the precision matrix, and each diagonal element is given a Gamma prior; for the non-hierarchical setup in (10), similarly μ_s is given a normal prior and the diagonal elements of the precision matrix Φ_s^{-1} are Gamma. The models are estimated by recursively drawing from the posterior distributions of the parameters. We use a burn-in of at least 5000 draws, and 20000 target draws. Convergence is assessed by inspecting the plots of

[3] Specifying off-diagonal elements for these matrices resulted in convergence problems in the Gibbs sampler used to estimate the models. The off-diagonal entries represent correlations which, in our set-up, are in any case accounted for by mixture-model heterogeneity, which induces a correspondence between sets of parameters for the missing data model and the ratings model.

draws against the iterations, as well as several convergence statistics such as the Gelman-Rubin convergence statistic (Brooks and Gelman 1998). Model results are summarized through the posterior means and standard deviations of the parameters.

Empirical Analysis

The EachMovie Data

Compaq Equipment Corporation provided the data[4] for our analysis. Data were collected through a movie recommendation system called EachMovie. Because much previous research has used the same data source – including Ansari, Essegaier, and Kohli (2000), Breese, Heckerman, and Kadie (1998), Chien and George (1999), and Hofmann and Puzicha (1999) – the EachMovie data provides an excellent benchmark to gauge relative model performance.

The dataset includes 72916 customers' numerical ratings for 1628 different movies, collected between March 1996 and September 1997. The ratings were originally on a zero-to-five "star"-scale, which we recode as 1 - 6. The dataset also includes information on customers' age and gender, as well as the movies' genre (comedy, thriller, etc.); note that movies can fall into multiple genres. As might be expected, any particular customer was unlikely to rate even a moderate proportion of all available movies, resulting in an enormous number of missing values in the 72916 by 1628 matrix implied by Table 1.

For model calibration, we draw a random sample of 1873 customers out of those who provided complete demographic information. To ensure stability in model estimation, we choose forty movies rated by at least some people. Still, our calibration sample is quite sparse, with 89.7% missing values. Analogous to Ansari, Essegaier, and Kohli (2000), three holdout samples

[4] The data are publicly available at http://www.research.compaq.com/SRC/eachmovie/.

are constructed for various prediction purposes. Table 2 provides descriptive statistics of the calibration sample and three holdout samples.

Insert Table 2 about Here

Models

We estimate the two models given by (9) and by (10), respectively referred to as Model 1 and Model 2, that have a hierarchical and a non-hierarchical specification of the customer characteristics effects respectively. We also estimate a restricted version of the model in (9), which we will refer to as Model 3, in order to explore the impact of accounting for a non-ignorable missing data mechanism. In the restricted model, the missing ratings are assumed to be missing completely at random (MCAR). Therefore, no parameters are estimated for the selection process.

Results

We compute the approximate log marginal density of the data to select the optimal number of segments for all three models. For Models 1 and 2, the two-segment solutions are unambiguously best (Model 1 (hierarchical): S = 1, –45264; S = 2, –44923; S = 3, –45173. Model 2 (non-hierarchical): S = 1, –45405; S = 2, –44955; S = 3, –45872. As for Model 3, although the log marginal density improves as the number of segments increases (S = 1, –22479; S = 2, –22399; S = 3, –22359), the chains for three segments display lack of convergence. Hence, we explore two-segment versions of all three models. Furthermore, the log marginal density for Model 1 is superior to that of Model 2 across all numbers of segments, indicating that

the hierarchical specification fits these data better than the non-hierarchical one. This further seems to suggest that the effect of the consumer characteristics is to modify subjects' genre preferences, in terms of both selecting and rating. We will see this interpretation at least partially borne out in comparing estimated effects for Models 1 and 2, below.

In Table 3, we report the mixing proportions and the fixed effects of genre for Model 1 with two segments. Table 3 indicates that there are two sizable segments that can be delineated for purposes of movie recommendations; it is comforting to note that these segments are quite similar to those picked up by Model 2 (in Table 4; discussed separately below). The richness of the hierarchical specification allows us to explore several types of differences systematically: (1) across genres; (2) by demographic groups (interactions); (3) between segments; (4) between selection and rating; and (5) across models. Throughout, recall that "selection" refers to *whether* a movie is rated and "rating" refers to *how* it is rated.

Insert Tables 3 and 4 About Here

Differences across Genres. There are strong and obvious differences across genres in terms of how well each is liked (rated) overall. Holding aside differences between Segments and consumer types, it is quite clear that Horror movies aren't well liked by the panelists, nor are action movies. By contrast, Drama and Classic movies are quite well received on average. Were one to formulate a model for ratings alone, these may well be the final conclusions. However, note that although Classic movies are rated highly, they are *selected* less frequently than any other genre type, by a wide margin. We further elaborate on such differences below.

Differences by Demographic Groups. Because the hierarchical specification allows genre preferences to differ by demographics – by age (standardized) and gender (contrast coded) – we can explore how these affect both selection and ratings; several such interactions appear at the bottom of Table 3. In some cases, such effects are exceptionally strong. In Segment 1, for example, each standard deviation in age (approximately 12 years, as per Table 1) translates into .5 on the latent propensity scale for Rating, a value higher than many of the 'main effects' for the genres themselves. Gender effects are also particularly salient: in Segment 1, Action, Animation and Classic are each at least 2.4 higher for men than for women; these effects are stronger, in fact, than nearly *all* the main effects for genre, in either Segment. Such pronounced differences in preference between the genders are picked up only in the hierarchical specification. Although cross-age and -gender *selection* differences seem more modest, they can nonetheless be dramatic, as seen for Classic movies in Segment 1.

Differences between Segments. It would be fair, overall, to claim that Segment 1, the smaller of the two, comprises 'negativists' and 'extremists': in terms of their ratings, they tend to hold harsher and more diametric views than their counterparts in Segment 2. For example, while neither Segment cares much for Action or Horror films, the views of Segment 1 are far stronger in each case. And, with the exception of Drama (and non-significantly, Romance), Segment 1 likes each genre type less than Segment 2. This may be exacerbated by the large differences in how men and women in Segment 1 view several of the movie types (e.g., Action, Animation, Classic), as discussed above. There are also large differences in movie selection between the segments, but the pattern is more complex than for the ratings. Segment 1 tends to select (i.e., to offer ratings for) Action, Animation, Horror and Thriller films more than Segment 2, while the reverse is true for the Comedy, Drama and Family

genres. That this pattern is quite different from that for Ratings underscores both the complexity of consumer behavior in such hedonic categories and the ability of the proposed model formulation to capture it.

Differences between Selection and Rating behavior. Even without considering segment or demographic differences, those between *whether* a film is rated, and *how* it is rated, are stark. This is perhaps clearest for Classic films: they are apparently not very popular, judging from how often they are selected for rating, with exceptionally negative coefficients for each segment. However, for those who choose (and subsequently rate) them, they are very well liked indeed, overall as well as *any* of the genres, with the possible exception of Drama. Simply put, the genre selected *least* is rated *best*. While this sort of pattern is common for luxury goods in many categories, we see no way to have anticipated it for movies, particularly since Classics are widely available, and are typically less costly. By contrast, Action movies are very widely selected in Segment 1, but receive very low ratings overall. Just focusing for examples on Segment 1, we see other complex patterns: Men select Animation less, but like it more, than women, yet they both select *and* rate Classic films more positively. The clearly pronounced pattern of differences between selection and rating behavior isn't easily captured with a single label, but were ratings and selection behavior broadly concordant, one needn't model them separately, and an account of ratings alone – as offered by prior models – would suffice.

Differences between Models. Relative measures of fit speak plainly as to the better performance of the proposed models, but tell only part of the story. Given that Models 1 and 2 posit heterogeneity, and each settle on two segments, one might question whether there is any 'tangible reality' to them, or whether they are merely technically convenient. A glance

at the estimated coefficients for the two models makes it clear that, despite their differences, they are each picking up broadly identifiable characteristics of two latent segments. For example, the patterns of selection coefficients are nearly identical. While there are some differences in ratings coefficients between the two models, we might well remember that it was for rating behavior that we saw the greatest differences across demographic groups: the 'interaction' effects, accounted for only by the hierarchical model, apparently capturewhat Model 2 attributes to main effects of the Genres. Let us consider Animated movies as an example: Model 1 and Model 2 estimate nearly identical selection coefficients for the segments (despite small differences in segment sizes). However, Model 2 shows positive coefficients for Ratings in both segments – .997 and .807, respectively – and so would conclude that Animation is very well-liked in both. Model 1, by contrast, reveals a more complex portrait: holding aside Gender, Animation has a slightly *negative* rating for Segment 1(compared to the baseline across all ratings); yet it is for this segment that Men rate Animation 2.8 units higher than Women. Thus, by not considering how demographics such as Gender *mediate* the effects of the Genres, Model 2 paints a different portrait of consumer rating behavior. We consider (the hierarchical) Model 1 superior in terms of fit as well as interpretation.

Holdout Recommendation Validity

In order to compare model performance, we use the hierarchical Bayes model in Ansari, Essegaier, and Kohli (2000) and the mixture model in Chien and George (1999) – henceforth AEK and CG – as benchmarks. A somewhat modified version of AEK's original model must be estimated, in that the experts' ratings they used as explanatory variables are not available for all

movies in our data. Hence, for comparison purposes, explanatory variables include Age and Gender for customers, and the nine Genres for movies.

We report the holdout recommendations for Models 1 to 3 in Table 5, as well as those for CG (with two segments) and AEK. Two statistics are calculated for cross-model comparisons – Root Mean Square Error (RMSE) and Mean Absolute Deviation (MAD). We note that the AEK model predicts ratings on a continuous scale; we use these 'raw' values rather than convert to the discrete ordinal scale of the data, which would involve arbitrary truncation. We compute the average predicted ratings for Models 1 to 3 and for CG as Σ(rating probability $\times$ discrete rating). For Models 1 and 2, which account for the missing data mechanism, the average predicted ratings are multiplied by the selection probabilities.

We compute RMSE and MAD for the predictions of both observed and missing ratings. However, the fact that some of the models don't account for missing data greatly affects their holdout performance, given the sheer quantity of missing data. Therefore, we augment Model 3, AEK and CG to account for missing data in a simple manner: the ratings predicted by the three models are multiplied by the non-missing rate of the calibration sample. Doing so accommodates the missing data by allowing each customer the same constant probability of rating a specific movie. This simple adjustment improves the holdout performance of these models and enables a "fairer" comparison with the proposed models. Because MAD treats deviations on any point of the scale identically – as opposed to penalizing larger deviations by squaring them, like RMSE – we discuss model results largely in terms of this measure; Table 5 contains all comparisons using both MAD and RMSE.

Insert Table 5 About Here

In their standard forms, the models that do not specify a missing data mechanism are outperformed dramatically by Models 1 and 2, which do: RMSEs and MADs are over twice as large for AEK, CG and Model 3. This is unsurprising, given the relative prevalence of missing data. Even when we accommodate missing data in these three models, Models 1 and 2, which do about equally well, are superior across the board. Tellingly, Model 3, which lacks a formal missing data mechanism, is edged out by both AEK and CG. The pattern of results – with AEK and CG performing better the Model 3 but substantially less well than Models 1 and 2 – strongly suggest that the missing data are non-ignorable, and providing a model-based account of them aids greatly in prediction. Notice that the degree of dominance of Models 1 and 2 over the others increases steadily with the amount of "new" data accounted for: the difference is smallest for "Existing People / Existing Movies", larger for "New People / Existing Movies", larger still for "New People / Existing Movies", and largest for "New People / New Movies". Although somewhat speculative, we believe that the ability of Models 1 and 2 to anticipate missing data patterns helps them do nearly as well with "new" people and movies as with existing ones; the best MAD value (0.701 for the hierarchical Model 2 for "Existing People / Existing Movies") is only slightly better than the worst one (0.719 for the non-hierarchical Model 1 for "New People / New Movies"). The models of Ansari, Essegaier, and Kohli (2000) and Chien and George (1999) therefore suffer more by comparison, as they cannot capitalize on predicted patterns in the missing data, based on the Genres and subject-specific covariates Age and Gender.

Taken in total, these results demonstrate the importance of accounting for non-ignorable missing data in generating product recommendations. When other popular models are 'augmented' with a naïve missing data mechanism, their holdout performance is still substantially poorer than the proposed models. While the raw differences in such measures as

MAD may appear modest, one might well consider three points. First, *proportional* improvement using the hierarchical model over the CG and AEK models was at least as large as 10% in all conditions, greater in the "New" conditions, and as large as 15% relative to the AEK model for "New People / New Movies". Second, given the fact that the benchmark models are already very sophisticated, such an improvement in holdout performance is quite satisfactory. Finally, recommendation systems are linchpins in online businesses proffering billions in recommended merchandise, so that even small performance increases may have major effects on sales revenue. The observed improvements, by that standard, warrant a good deal of attention to the procedures proposed here from online retailers.

Discussion and Conclusion

The empirical results above demonstrate that three modeling constructs – a non-ignorable missing data mechanism, a parsimonious account of the ordinal nature of ratings data, and using both a finite mixture and a continuous heterogeneity distribution – can substantially improve the accuracy of making product recommendations. But such statistical comparisons deal only with performance metrics, not quantities of direct interest to practitioners. The suggested approach provides additional benefits that previous recommendation systems are unable to offer. Chief amongst these is the ability of our models to capture customers' selection behavior, that is, the very decision to provide an evaluation for a specific product in the first place. We believe this feature to be of great potential use, since company recommendation databases such as the EachMovie system include an enormous amount of missing data, as customers rate only a very small sample of available products.

Simply ignoring the missing data and basing recommendations on observed data alone may yield suboptimal or outright inaccurate recommendations. Selection behavior may be not entirely consistent with the customer's rating behavior, as can be seen from the model estimates in Table 3. For example, people in Segment 1 are more likely to rate (i.e., select) Action movies, but they tend to give low ratings to them; they have exactly the opposite pattern with Classic movies, which are selected (to be rated) very infrequently, but apparently liked a great deal. Although the proposed models do not pin down the exact causes of the 'inconsistency' – if indeed that's what it is – between selection and rating behavior, it may be worthwhile for marketers to further investigate the underlying reasons and act upon them.

The knowledge provided by our model can help fashion personalized requests for more ratings input from customers, and consequently further improve product recommendations. In other words, our model can not only make more accurate recommendations, but also be used to help gather more information from customers and thus refine the recommendation system in a directed manner. Specifically, the estimated selection model components may serve such a purpose: for a new customer – for whom a recommendation was made based on a high predicted utility for the movie in question – if that customer subsequently rates another product, an evaluation may be requested for the recommended products, and/or other products with a high predicted probability of being selected by the customer. In such a way, rather than depending on the ratings provided haphazardly by customers (as picked up by the selection model), the recommendation system may feed itself with relevant data to improve its own performance. For example, in our application the probability of selecting Classic movies for rating is low, while these movies were rated highly. This may result in too many positive recommendations for these

movies. Directed requests for ratings of Classic movies may therefore improve the quality of the recommendation system.

One drawback to building a more complex statistical model for making recommendations, shared by our model and those of Ansari, Essegaier, and Kohli (2000) and Chien and George (1999), is that calibration may be computationally intensive and time consuming. However, calibration of the model need be done only at certain time intervals and for samples of customers. Moreover, once the model has been calibrated, the recommendations themselves can be made very quickly, in real time, since the prediction equations are all closed form. Such a feature would be critical in an online system, where pages are served up on demand and multiple times for each customer visit, as Amazon.com currently implements. Nonetheless, the model would need to be recalibrated at regular time intervals to reflect evolution in customers' selection and rating of movies. All this is feasible with current technology, so that the model we have developed can be readily implemented for existing recommendation systems.

There are several ways, nevertheless, to extend the current study. For instance, we model customers' rating behavior as if each customer perceives and uses the scale similarly; consideration of scale usage heterogeneity (Rossi, Gilula, and Allenby 2001) may further improve model fit and the recommendation quality. Due to data restrictions and parsimony, we have not modeled a number of features of interest to online retailers: cross-category behavior, visit timing, or indeed observed browsing / purchase patterns or volume. Doing so, as such data become more freely available, may further enhance our understanding of customer behavior and recommendation quality.

Table 1
SAMPLE CUSTOMER PREFERENCE DATA

Customers	Products						
	P1	P2	P3	P4	P5	P6	…
C1	5	NA	1	3	NA	4	…
C2	NA	NA	6	2	3	NA	…
C3	4	2	NA	NA	NA	6	…
C4	1	5	3	NA	6	3	…
C5	6	NA	6	NA	2	NA	…
C6	NA	NA	2	5	4	NA	…
…	…	…	…	…	…	…	…

Table 2
SUMMARY OF SAMPLES

Samples	*# People*	*# Movies*	*Mean Age*	*STD Age*	*% Female*	*Missing Rate*
Existing People/Existing Movies*	1853	40	32.4	11.8	22.2%	89.7%
Existing People/New Movies	1606	40	32.4	12.6	19.8%	88.5%
New People/Existing Movies	1733	40	31.9	11.9	23.0%	89.7%
New People/New Movies	1498	40	31.8	12.1	21.8%	88.3%

* Calibration sample

Table 3
ESTIMATES OF GENRE EFFECTS
ON SELECTION AND RATING PROBABILITIES: MODEL1

Variables	*Segment 1*		*Segment 2*	
Mixing Proportions	0.368		0.632	
	(0.021)		(0.021)	
	Selection	*Rating*	*Selection*	*Rating*
Action	0.459	-2.192	-0.491	-0.556
	(0.087)	(0.266)	(0.066)	(0.110)
Animation	1.100	-0.198	0.374	0.733
	(0.071)	(0.505)	(0.059)	(0.115)
Classic	-6.742	0.431	-4.579	0.718
	(1.648)	(0.567)	(0.637)	(0.154)
Comedy	-0.223	-1.045	0.576	-0.102
	(0.040)	(0.239)	(0.048)	(0.039)
Drama	-0.687	1.213	0.730	0.286
	(0.049)	(0.242)	(0.028)	(0.051)
Family	0.108	-0.147	0.687	-0.198
	(0.061)	(0.377)	(0.038)	(0.069)
Horror	-0.308	-3.678	-1.365	-1.009
	(0.075)	(0.797)	(0.200)	(0.486)
Romance	0.389	0.028	0.362	-0.029
	(0.054)	(0.279)	(0.073)	(0.054)
Thriller	0.570	-0.320	-0.420	0.225
	(0.073)	(0.266)	(0.152)	(0.117)
Age*	-0.169	-0.346	-0.060	0.140
	(0.041)	(0.187)	(0.046)	(0.054)
Gender**	0.137	-1.108	-0.332	0.001
	(0.113)	(0.459)	(0.103)	(0.122)
Intercept	-2.485	1.339	-2.850	0.916
	(0.055)	(0.231)	(0.048)	(0.060)
Comedy*Age	-0.097	0.506	0.009	-0.190
	(0.054)	(0.191)	(0.033)	(0.044)
Action*Gender	0.474	3.110	-0.068	0.112
	(0.141)	(0.526)	(0.123)	(0.224)
Animation*Gender	-0.350	2.799	-0.178	0.111
	(0.113)	(1.029)	(0.084)	(0.226)
Thriller*Gender	-0.200	0.996	-0.467	-0.464
	(0.177)	(0.496)	(0.224)	(0.219)
Classic*Gender	2.678	2.417	-0.053	-0.864
	(1.971)	(1.221)	(0.528)	(0.321)

* Age is standardized

** Female is coded as -1/2; male as 1/2

*** Standard deviation of parameter estimates is in parentheses

Table 4
ESTIMATES OF GENRE AND DEMOGRAPHIC EFFECTS ON SELECTION AND RATING PROBABILITIES: MODEL 2

Variables	*Segment 1*		*Segment 2*	
Mixing Proportions	0.431		0.569	
	(0.025)		(0.025)	
	Selection	*Rating*	*Selection*	*Rating*
Action	0.485	-1.032	-0.529	-0.506
	(0.061)	(0.166)	(0.085)	(0.103)
Animation	1.061	0.977	0.268	0.807
	(0.057)	(0.210)	(0.064)	(0.111)
Classic	-6.000	1.445	-4.611	0.431
	(1.784)	(0.352)	(0.693)	(0.145)
Comedy	-0.254	-0.602	0.725	-0.095
	(0.032)	(0.155)	(0.035)	(0.044)
Drama	-0.645	0.968	0.955	0.323
	(0.065)	(0.150)	(0.027)	(0.049)
Family	-0.004	-0.756	0.830	-0.189
	(0.048)	(0.213)	(0.027)	(0.068)
Horror	-0.698	-3.056	-1.169	-0.902
	(0.129)	(0.884)	(0.234)	(0.385)
Romance	0.313	-0.011	0.352	-0.136
	(0.047)	(0.179)	(0.070)	(0.052)
Thriller	0.442	0.148	-0.670	0.039
	(0.072)	(0.141)	(0.202)	(0.063)
Age*	-0.099	0.107	-0.135	0.075
	(0.032)	(0.060)	(0.031)	(0.028)
Gender**	-0.015	0.079	-0.102	-0.097
	(0.073)	(0.025)	(0.073)	(0.073)
Intercept	-2.416	0.905	-2.972	0.944
	(0.034)	(0.141)	(0.048)	(0.056)

* Age is standardized

** Female is coded as -1/2; male as 1/2

*** Standard deviation of parameter estimates is in parentheses

Table 5
IN-SAMPLE AND HOLDOUT RECOMMENDATION RESULTS [1]

	Existing People Existing Movies		*Existing People New Movies*		*New People Existing Movies*		*New People New Movies*	
	RMSE	**MAD**	**RMSE**	**MAD**	**RMSE**	**MAD**	**RMSE**	**MAD**
Model 1*	1.354	0.702	**1.425**	0.706	1.361	0.707	**1.446**	0.719
Model 2**	**1.351**	**0.701**	1.432	**0.703**	**1.359**	**0.705**	1.453	**0.715**
With Bernoulli Missing Data Mechanism								
Model 3***	1.379	0.798	1.436	0.833	1.389	0.804	1.454	0.842
Ansari et al. (2000)	1.379	0.780	1.441	0.817	1.389	0.784	1.459	0.826
Chien & George (1999)	1.369	0.758	---	---	1.380	0.764	---	---
No Missing Data Mechanism								
Model 3	4.265	4.060	4.241	4.032	4.273	4.074	4.234	4.021
Ansari et al. (2000)	4.028	3.880	4.006	3.848	4.023	3.874	3.997	3.838
Chien & George (1999)	3.846	3.654	---	---	3.860	3.669	---	---

* Model 1: the hierarchical model in Equation (9)

** Model 2: the non-hierarchical model in Equation (10)

*** Model 3: the restricted version of Model 1 with MCAR assumption

[1] Minimum value across all methods in each column indicated in boldface type

References

Allenby, Greg M. and Peter E. Rossi (1999), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, 57-78.

Andrews, Rick L., Asim Ansari, and Imran S. Currim (2002), "Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery," *Journal of Marketing Research*, 39 (February), 87-98.

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000), "Internet Recommendation Systems," *Journal of Marketing Research*, 37 (August), 363-75.

Boatwright, Peter, Robert McCulloch and Peter Rossi (1999), "Account-Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model," *Journal of the American Statistical Association* , 94, 1063-1073.

Breese, Jack, David Heckerman, and Carl Kadie (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, WI: Morgan Kaufmann Publisher.

Brooks, S. P. and Andrew Gelman (1998), "Alternative methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, 7, 434-55.

Canny, John (2002), "Collaborative Filtering with Privacy", in *IEEE Symposium on Security and Privacy*, Oakland, CA, May 2002, 45-57.

Chien, Yung-Hsin and Edward I. George (1999), "A Bayesian Model for Collaborative Filtering," Working Paper, University of Texas at Austin.

Heckman, James J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47 (1), 153-61.

Hofmann, Thomas and Jan Puzicha (1999), "Latent Class Models for Collaborative Filtering", in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 688-693.

Lee, Lillian (1999), "Measures of Distributional Similarity," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 25-32.

Little, Roderick J. A. and Donald B. Rubin (1987), Statistical Analysis with Missing Data. New York: John Wiley & Sons.

O'Connor, Mark and Jon Herlocker (1999), "Clustering Items for Collaborative Filtering," Working Paper, University of Minnesota.

Resnick, P., N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl (1994), "Grouplens: An Open Architecture for Collaborative Filtering of Netnews," in *the ACM 1994 Computer Supported Cooperative Work Conference*. New York: ACM.

Rossi, Peter E., Zvi Gilula and Greg M. Allenby (2001), "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, 96, 20-31.

Rossi, Peter E., Robert E. McCulloch and Greg M. Allenby (1995), "Hierarchical Modeling of Consumer Heterogeneity: An Application to Target Marketing," *Case Studies in Bayesian Statistics*, 323-349, New York: Springer-Verlag.

Rost, Jurgen (1985), "A Latent Class Model for Rating Data," *Psychometrika*, 50 (1), 37-49.

Rubin, Donald B. (1976), "Inference and Missing Data (with discussion)", *Biometrika*, 63 (3), 581-92.

Sarwar, Badrul M., George Karypis, Joseph A. Konstan, and John T. Riedl (2000), "Analysis of Recommendation Algorithms for E-Commerce," Working Paper, GroupLens Research Group/Army HPC Research Center, University of Minnesota.

Wedel, Michel, Wagner A. Kamakura, Albert C. Bemmaor, J. Chiang, Terry Elrod, R. Johnson, Peter J. Lenk, Scott A. Neslin, and C. S. Poulsen (1999), "Discrete and Continuous Representation of Heterogeneity," *Marketing Letters*, 10 (3), 217-30.