ELSEVIER

Intern. J. of Research in Marketing 20 (2003) 45–65

International Journal of Research in Marketing

www.elsevier.com/locate/ijresmar

Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction

Wagner A. Kamakura[a,*], Michel Wedel[b,c], Fernando de Rosa[d], Jose Afonso Mazzon[e]

[a] *Fuqua School of Business, Duke University, Durham, NC 27708, USA*
[b] *Faculty of Economics, University of Groningen, 9700 AV Groningen, Netherlands*
[c] *University of Michigan Business School, 701 Tappan Street, Ann Arbor, MI 48109, USA*
[d] *Universidade de Brasilia, SQSW 394 Bloco 1, Apto 507, Brasilia 70673-409, DF, Brazil*
[e] *Universidade de São Paulo, Faculdade de Economia, Administração e Contabilidade, 05508-900, São Paulo, Brazil*

Received 1 August 2001; received in revised form 1 May 2002; accepted 14 May 2002

Abstract

An important aspect of the new orientation on customer relationship marketing is the use of customer transaction databases for the cross-selling of new services and products. In this study, we propose a mixed data factor analyzer that combines information from a survey with data from the customer database on service usage and transaction volume, to make probabilistic predictions of ownership of services with the service provider and with competitors. This data-augmentation tool is more flexible in dealing with the type of data that are usually present in transaction databases. We test the proposed model using survey and transaction data from a large commercial bank. We assume four different types of distributions for the data: Bernoulli for binary service usage items, rank-order binomial for satisfaction rankings, Poisson for service usage frequency, and normal for transaction volumes. We estimate the model using simulated likelihood (SML). The graphical representation of the weights produced by the model provides managers with the opportunity to quickly identify cross-selling opportunities. We exemplify this and show the predictive validity of the model on a hold-out sample of customers, where survey data on service usage with competitors is lacking. We use Gini concentration coefficients to summarize power curves of prediction, which reveals that our model outperforms a competing latent trait model on the majority of service predictions.

Keywords: Database marketing; Cross-selling; Customer relationship management

1. Introduction

As many product and service markets become saturated and highly competitive, vendors realize that the acquisition of new customers happens mostly at the expense of competitors and, at the margin, these new customers tend to be ‘‘switchers’’ who will likely switch again in response to an attractive competitive offer. This competition for new customers in mature markets leads to the phenomenon known as ‘‘churn,’’ in which each vendor becomes a revolving door of acquired and lost customers. In order to escape this vicious circle, firms are increasingly focusing on

* Corresponding author.
E-mail addresses: Kamakura@duke.edu (W.A. Kamakura), Wedel@umich.edu (M. Wedel), jamazzon@usp.br (J.A. Mazzon).

0167-8116/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0167-8116(02)00121-0

strengthening the relationships with their customers (Day, 2000). Customer relationship management (CRM) has been more than a "buzzword" in management and marketing circles. According to industry sources,[1] worldwide CRM-related investments reached $3.3 billion in 1999 and are expected to reach $10.2 billion by 2003.

One of the main CRM tools for forging stronger relationships with customers is cross-selling (Kamakura, Ramaswami, & Srivastava, 1991). The rationale for cross-selling as a strategy for reducing customer "churn" is very simple. As a customer acquires additional services or products from a vendor, the number of points where customer and vendor connect increases, leading to a higher switching cost to the customer. For example, it is easier for a customer with only a checking account to close this account than for another customer who also has automatic paycheck deposit and bill payments. Another important benefit of cross-selling, not as immediately visible as the increase in customer switching costs, is that it allows the firm to learn more about the customer's preferences and buying behavior, thereby increasing its ability to satisfy the customer's needs more effectively than competitors. For example, as a bank increases its "share-of-wallet" from a customer, it becomes more familiar with the customer's financial needs, and in a better position than competitors to develop and offer services that satisfy those needs.

On the other hand, cross-selling can also potentially weaken the firm's relationship with the customer, because frequent attempts to cross-sell can render the customer non-responsive or even motivated to switch to a competitor. In order to effectively cross-sell its products/services, the marketer must find—in commonly used jargon—the right offer for the right customer at the right time. The customer transaction database is instrumental in achieving that, because it allows the firm to learn about a customer, through its experience with other customers with similar behavioral patterns. However, usually only transaction data with the company in question are included in the database, while relevant marketing data, for example, on the use of competitive products, are lacking and need to be collected in separate surveys among a sample of customers. In addition, the development of techniques for the extraction of relevant information from the database for strategic marketing purposes, often referred to as data-mining, has lagged behind the development of tools for collecting and storing the data.

In this study, we develop a new data-augmentation tool to predict consumption of new or current products by current customers who do not use them yet. We provide a mixed data factor analyzer that is tailored to implement cross-selling based on customer transaction data and identifies the best prospects for each service. The model extends previous factor analysis procedures and enables us (1) to analyze data from a variety of different types, i.e. choices, counts, or ratings; (2) to represent the variability of those variables in a latent subspace of reduced dimensionality; and (3) to analyze data from the customer database in combination with survey data collected only on a sample from the customer database. The main purpose in applying the model is to learn from the behavioral patterns of all customers in the database and from external data gathered from a survey of a sample of customers, to identify the best prospects for the cross-selling of services, so that each customer is only offered a service she is very likely to be interested in.

The remainder of this paper is organized as follows. In Section 2, we provide a framework describing the role of cross-selling as a tool to enhance customer relationships and review relevant literature on cross-selling. Then, we explain a new mixed data factor analyzer to identify cross-selling opportunities from customer transaction databases. We show how it extends recent work on factor analysis for non-normal variables. Next, the model is calibrated on a customer transaction database from a large retail bank. We compare our model to alternative models and investigate which has better performance in evaluating ownership of financial services. Finally, we discuss other potential applications as well as limitations.

2. Cross-selling

Cross-selling pertains to efforts to increase the number of products or services that a customer uses

[1] CRM Report: "Worldwide CRM Applications Market Forecast and Analysis Summary, 2001–2005". http://www.idc.com.

within a firm. Cross-selling products and services to current customers has lower associated cost than acquiring new customers, because the firm already has some relationship with the customer. A proper implementation of cross-selling can only be achieved if there is an information infrastructure that allows managers to offer customers products and services that tap into their needs, but have not been sold to them yet.

Furthermore, we conjecture that cross-selling is effective for customer retention by increasing switching costs and enhancing customer loyalty, thus directly contributing to customer profitability and life time value. The more services a customer uses with the firm, the higher the costs of switching to other firms, which leads to loyalty and tenure. We illustrate this in Fig. 1. The graph is derived from the empirical application below and shows the number of years of being a customer versus the number of services used. Fig. 1 reveals a strong positive relationship of the number of years of being a customer and the number of services used from the bank. Although causality cannot be demonstrated, there is likely a mutually reinforcing effect. As the length of the relationship increases, customers are inclined to use more services from the bank and, when more services are used, switching costs increase, so that ending the relationship with the bank becomes less attractive. Thus, customer retention is enhanced through cross-selling as switching costs increase with multiple service relationships.

As the intensity of satisfactory interaction with the customer increases, the firm learns more about the customer's needs and wants, increasing its ability to develop customer loyalty and fend-off competitors. At the same time, the enhanced loyalty leads to increased profitability. Therefore, use of more services leads to higher profits, if the services are properly cross-sold. We illustrate this in Fig. 2, again derived from our empirical data set described below. This figure plots the profitability of a customer against the number of services s/he uses from the bank. One can see again that there is a significant positive relationship, showing that cross-selling directly generates increased profitability by enhancing the life-time value of customers.

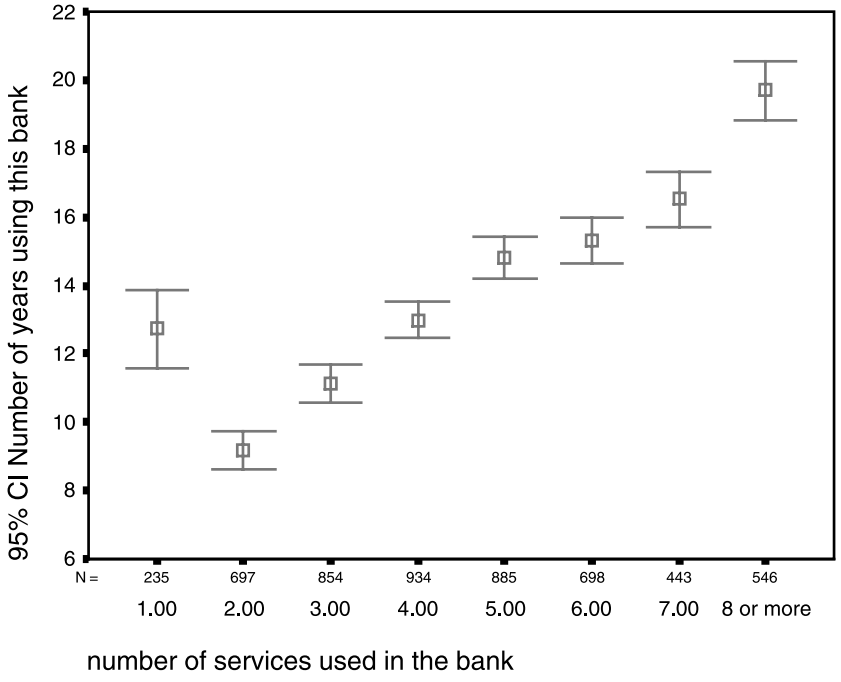

Fig. 1. Number of years of using the bank plotted against the number of services used, with 95% confidence intervals.

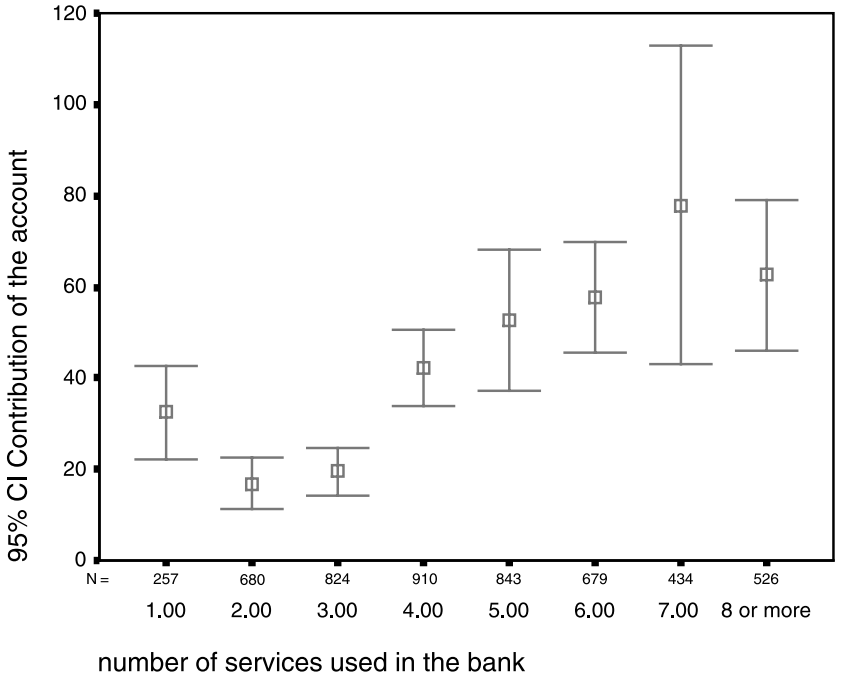

Fig. 2. Profitability of the account plotted against the number of services used, with 95% confidence intervals.

Despite its importance for relationship marketing, cross-selling has received limited attention in the academic literature. Most of the literature focuses on methodology for identifying common acquisition patterns of products by customers based on their usage or ownership data. The problem is to infer the longitudinal pattern of acquisition across various products or services, when only cross-sectional data are available on usage or ownership. One of the earliest attempts is the study by Paroush (1965), who uses Guttman's (1950) coefficient of reproducibility as an indicator of the order of acquisition implied by cross-sectional data. Paroush's study has been replicated and extended by Hebden and Pickering (1974), Kasulis, Lusch, and Stafford (1979), and Stafford, Kasulis, and Lusch (1982).

However, the models used in these studies were not explicitly developed to implement cross-selling. Kamakura et al. (1991) propose a uni-dimensional latent-trait model that makes probabilistic predictions that a consumer would use a particular product or service, based on their ownership of other products/services and on the characteristic of the new one. They apply this latent-trait model to survey data on the use of financial services. However, the approach requires that the firm knows about each customer's usage of services from both the firm and its competitors, something unlikely to be observed in practice. In most cases, information on ownership of competitive products is available only when collected as a sample of a firm's customers. Such incomplete data cannot be analyzed with the model of Kamakura et al. Moreover, their specification is limited since it assumes that a single unobserved dimension adequately summarizes the variation of the variables contained in the transaction database and it can only handle binary (0/1) variables, whereas transaction databases usually contain a wide variety of different variables, such as counts, choices, ranks, and classifications.

To accommodate these requirements for a parsimonious model for the description of cross-buying and its use for cross-selling purposes, we extend the recent literature on factor analysis for non-normal variables and exploit its strengths in the imputation of missing data. Our approach builds on recent work in factor analysis for non-normal variables, in particular that by Bartholomew and Knott (1999), Kamakura and Wedel (2000), Moustaki and Knott (2001), and Wedel and

Kamakura (2001). We extend that work in two ways. First, by developing a factor analyzer for mixed outcome data, simultaneously dealing with missing observations. Previous work in this area, as cited above, has not accommodated such mixed outcome data, where some variables pertain to choices, others to ratings, some others to rank-ordered variables, and others to counts. Such a mix of data types is fairly typical in customer transaction databases and its proper analysis is a non-trivial exercise. It is important to accommodate the measurement scales of the variables in forecasting the success of cross selling efforts, where predictions need to be confined to the proper support. A second extension of past work on factor analysis is that we deal with missing data that arise due to sub-sampling. Again, this situation arises fairly often in customer transaction databases, where the transaction data is augmented with a survey among its customers. In addition, the approach that we propose next offers advantages over the one that has been postulated by Kamakura et al. (1991) in that it accommodates a much broader range of distributions of observed variables, allows for multiple dimensions, and allows for predictions that extend beyond the information available within a firm's customer database.

3. A mixed data factor analyzer for identifying cross-selling prospects

Customer-oriented businesses have a wealth of customer information at their disposal, generated from their data production systems. Harnessing this rich source of customer level transaction information is increasingly important to marketers. Database marketing (DBM) involves building, organizing, supplementing, and mining customer transaction databases to increase the accuracy of marketing efforts by enabling the identification of the best prospects for marketing efforts (Goodman, 1992; Labe, 1994). Many DBM efforts have been ineffective, however, since the database is only used as a mailing list and the possibilities for integration of marketing and computer systems are not effectively exploited (Shaw, 1993). Two causes of this undesirable state of affairs can be identified. First, in many cases, detailed transaction data pertaining to the company in question are compiled, possibly enriched with ZIP-level Geo-Demographic data, but critical data on the use of products and services from competitors, and "soft data" such as customer satisfaction, are lacking. These often need to be collected in separate surveys. Due to the survey costs, such data are usually only collected from a sample of customers in the database. Yet, this type of information is needed for all customers for the effective implementation of one-to-one marketing. Second, the development of methods for the extraction of information for strategic marketing purposes has lagged behind the development of techniques for the construction and maintenance of the databases. Too few efforts have been made to tailor these methods to optimally match the structure of the database or the substantive marketing problem.

To effectively cross-sell its products/services, the marketer must find dependencies among product/service ownership, i.e. must identify the structure in customers' cross-buying behavior. In particular, one is interested in the likelihood that a particular customer will buy certain products or services that s/he does not own yet, given ownership of other products and services. We develop next a mixed data factor analyzer that is tailored to analyze cross-buying for the implementation of cross-selling based on customer transaction data and identifies the best prospects for each service.

3.1. Description of the factor analyzer

We assume that a firm has access to a customer transaction database and has conducted a survey among a random sample of its customers. Data from this sample survey serves to supplement the customer database, providing, in particular, information about usage of services from competitors. Thus, for a representative sample of its customers, the firm has complete information. Let $n = 1,\ldots,N$ denote customers in the database and $j = 1,\ldots,J$ represent observed variables. These J variables are measured on a variety of scales. In the application below, for example, income and education are rated on ordinal scales, volume of customer transactions on a ratio-scale, the total number of transactions is a discrete count, and service usage is measured with binary indicators. We assume the J observations, $y_j=(y_{nj})$, to be realizations of random variables, distributed in the exponential

family of distributions. The exponential family is a very general class of distributions, including both continuous and discrete distributions, which allows us to accommodate the various types of data typically encountered in DBM in a single framework, by assigning each observed variable j its own distribution. For example, binary indicators of service usage can best be modeled with a Bernoulli distribution, numbers of transactions with a Poisson distribution, rating scales with a rank-order binomial distribution, and the volume of transactions with a normal distribution. The exponential family allows one to optimally match the support of the selected distribution to the assumed measurement scale of the transaction variables. This is particularly important in predicting service usage for cross-selling, since individual-level predictions need to be logically consistent with each variable's measurement scale.

We aim at identifying a low-dimensional map of the observed variables that identifies the most salient features of these data and allows for graphical representation. X is the $(N \times P)$ stochastic matrix representing that low (P)-dimensional space, where we assume that the elements of X are independently distributed across subjects according to a standard normal distribution. We specify the conditional distribution of the observations for one particular subject:

$$f(y_n \mid x_n) = \prod_{j=1}^{J} \exp\left[\frac{y_{nj}\eta_{nj} - a_j(\eta_{nj})}{\phi_j} + b_j(y_{nj}, \phi_j)\right], \tag{1}$$

$$\eta_n = \lambda_0' + x_n\Lambda'. \tag{2}$$

Here, y_n=(y_{nj}) is a vector of observed data from customer n, x_n is the n-th row of an unobserved vector of i.i.d. normally distributed $(N \times P)$ quantities X, Λ a $(J \times P)$ matrix, and λ_0 a $(J \times 1)$ vector of fixed, but unknown, weights, is a dispersion parameter that applies for certain distributions in the exponential family such as the normal, $a_j(\cdot)$and $b_j(\cdot)$ functions depending on the particular distribution for the variable j (McCullagh & Nelder, 1989). Eq. (2) shows that the expectation of the observation vector for each subject is mapped onto a lower-dimensional subspace: $\eta(x_n)$ defining that map. Note that the specification of the distributions in Eq. (1) implies: $E[y_n \mid x_n] = h(\eta(x_n))$, with $h(\cdot)$ a canonical function, depending on the distribution of the data (it is the log-function for the Poisson and the logit-function for the binomial distribution, for example). Also note that the J observations on each individual, y_{nj}, are conditionally (but not marginally) independent, given x_n. Since x_n is normally distributed, so is $\eta(x_n)$.

Our model provides a factor analyzer, since the reduced P-dimensional space spanned by captures the salient features of the data and lends itself to a graphical representation of the weights that define the map. We specify the subject-specific map to have a prior normal distribution across subjects: $x_n \sim N_p(0,1)$. The use of the standard normal distribution for the latent variables alleviates scale and translation invariance of the model. Those arise because one can add a vector of scalars to x_n and subtract a vector of constants from λ_0, or one can post-multiply x_n and Λ with the inverse of a diagonal matrix T, which yields the same model, as in standard factor analysis.

The factor analyzer provided in Eqs. (1) and (2) is a powerful approach, since it maps observed variables of a wide variety of measurement scales nonlinearly onto a latent feature space of reduced dimension that lends itself for identification of important aspects of the data through graphical display. We view our model as one allowing for convenient graphical display of the structure of data, without a necessary interpretation of the factors as "latent dimensions". While we see our approach as useful for data reduction and data-mining, similar to PCA, we think that one should be careful in interpreting the results of factor analyses of behavioral data as latent perceptions or intentions. The reason is that, in making inferences on latent dimensions extracted from measurements of behavior, one makes strong claims with respect to the underlying process. Thus, contrary to the application of factor analysis to the analysis of measurement scales specifically designed for the identification of latent dimensions, the application of our tool to customer transaction databases is one where one is not primarily interested in a behavioral interpretation of the latent dimensions, but rather in a convenient low-dimensional graphical display of the structure in the data. However, the maps themselves are interpretable as we will show below.

Note that the distribution in Eq. (1) presents the conditional distribution of the data Y, given the latent variables X. To illustrate the form of the expression, assume that there are $J=J_N+J_B+J_P+J_R$ variables, with, respectively, a normal, Bernoulli, Poisson, and rank-order binomial distribution, as in the application below. Then the conditional distribution of the observed data given latent variables takes the following form:

$$f(y_n \mid \eta(x_n)) = \prod_{j=1}^{J_N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\sigma^{-2}(y_{nj}-\eta_{nj})^2\right] \times \prod_{j=J_N+1}^{J_N+J_B} \frac{\exp[y_{nj}\eta_{nj}]}{1+\exp[\eta_{nj}]} \times \prod_{j=J_N+J_B+1}^{J_N+J_B+J_P} \frac{\exp[y_{nj}\eta_{nj}-\exp[\eta_{nj}]]}{y_{nj}!} \times \prod_{j=J_N+J_B+J_P+1}^{J_N+J_B+J_P+J_R} \binom{K_j-1}{y_{nj}-1} \times \frac{\exp[(y_{nj}-1)\eta_{nj}]}{(1+\exp[\eta_{nj}])^{K_j-1}}. \quad (3)$$

Here, $\eta_{nj}=\lambda_{0j}+x_n\lambda_j'$, where λ_j is the j-th row of Λ, and K_j is the number of scale points of rank-order rating scale j.

It is of interest that our model requires the conditional distribution of the data, given the factor scores, to be in the exponential family. However, since the factor scores themselves follow a normal distribution, the marginal distribution of the data—obtained by integrating over the factor score distribution—is in general not in the exponential family and will accommodate overdispersion. In addition, our model assumes the observed variables to be conditionally independent, given the factor scores. However, our model accommodates marginal dependence of the variables, since they depend on the same unobserved factor scores. We consider these important features in modeling marketing data.

3.2. Estimation using SML

The unconditional distribution of the observations is obtained by integrating out the unobserved variables in Eq. (3). The likelihood of the factor analyzer, providing the support of the data for the parameters, is obtained as the product of that expression over all N observations. However, in applications to cross-selling, the observation vector is complete only for a sample of the subjects in the database, being obtained both from the database and the supplementary survey. For the remaining customers in the database, part of the data is missing and, for those subjects, we partition the observation vector as $y_n=(\hat{y}_n,\check{y}_n)$, with the corresponding sets of variables being $C=\hat{C}\cap\check{C}$, where we assume the first subset of variables to be observed without loss of generality. Also, we assume the customers to be ordered such that for the first M subjects complete data are available, while for the remaining N-M subjects the data are incomplete. The observed data likelihood is obtained by integrating the joint distribution of the observed and missing data over the distribution of the missing data in the likelihood:

$$L(\Xi \mid Y) = \prod_{n=1}^{N} \int\int \prod_{j\in\hat{C}} f(\hat{y}_{nj} \mid \eta(x_n),\Xi) \times \prod_{j\in\check{C}} f(\check{y}_{nj} \mid \eta(x_n),\Xi)\mathrm{d}\check{y}_{nj} f(x_n)\mathrm{d}x_n, \quad (4)$$

where we collect all parameters in Ξ. However, since the data are missing at random (MAR), the survey being conducted among a random sample of the database, this expression is equivalent to the simpler observed data likelihood:

$$L(\Xi \mid \hat{Y}) = \prod_{n=1}^{M} \int \prod_{j=1}^{J} f(\hat{y}_{nj} \mid \eta(x_n),\Xi) f(x_n)\mathrm{d}x_n. \quad (5)$$

Note that, in Eq. (5), we may ignore the missing data generating mechanism and replace the product over N (all subjects in the *database*) by a product over M (all subjects in the *sample*). We may ignore the missing data generating mechanism and use only complete data because the missing data are MAR, being under control of the researcher, the estimators based on Eq. (5) being unbiased (Little & Rubin, 1987).

The estimation of the factor analyzer is not feasible with standard (numerical) algorithms for maximizing the likelihood function, given the potentially high-dimensional integration involved in the likelihood. However, simulated likelihood (SML) estimation has

made the approximation of such integrals possible. Such simulation methods were introduced by McFadden (1989) and an overview is provided by Stern (1997). The problem is to evaluate the log-likelihood (Eq. (5) in the general case where x_n is a P-dimensional normal random variable. The idea of simulation is to draw S random variables z_n^S from $f(x_n)$ and use the approximation:

$$\tilde{L}(\Xi \mid \hat{Y}) = \sum_{n=1}^{M} \ln \sum_{s=1}^{S} \prod_{j=1}^{J} \tilde{f}(\hat{y}_{nj} \mid \eta(z_n^s); \Xi)/S \qquad (6)$$

instead of Eq. (5). The value of Ξ that maximizes Eq. (6) is the SML estimator. SML provides consistent estimators if $S \to \infty$ as $M \to \infty$. Then the simulated likelihood (6) is a consistent simulator of the likelihood (5). The bias in the estimates is of order $1/S$. However, finite values of S are sufficient to obtain good properties of the estimates. We use $S = 100$ (Lee, 1995).

3.3. Model selection and prediction

In most applications of the factor analyzer, the number of dimensions P is treated as unknown and needs to be determined empirically. Models with different numbers of factors cannot be compared using standard likelihood-based tests, since the asymptotic χ^2 distribution of the LR test of the P-factor model versus the $P+1$-factor model does not hold (Anderson, 1980). In order to determine the number of latent factors, we compare the solutions with different numbers of factors on the basis of the consistent Akaike information criterion (CAIC) (Bozdogan, 1987) and choose the solution with the lowest CAIC.

In order to predict/impute the missing data for all subjects in the transaction database, we compute the posterior expectation of these missing data, given the model estimates, and the values of the observed data for the subject in question:

$$E[\check{y}_{nj}] = \int \check{y}_{nj} f(\check{y}_{nj} \mid \eta(\hat{x}_n), \hat{\Xi}) \mathrm{d}\check{y}_{nj} \qquad (7)$$

Here, $\hat{x}_n$ is a vector with the posterior estimates of the factor scores for customer n; the integrals are again computed through repeated draws from the distributions in question. Currently, our imputations are based on the expected value in Eq. (7), but multiple imputations obtained as draws from the predictive distribution of the variable in question, with expectation as in Eq. (7), can also be generated (Little & Rubin, 1987).

4. Empirical illustration

4.1. Database marketing in the financial industry

In the US, the recent repeal of the Glass-Steagall Act lead to a wave of mergers in financial markets, blurring the distinction between banks, insurers, and brokerage firms. These mega mergers lower the barriers among financial industries (Shesbunoff, 1999). Conglomerates may capture all aspects of a consumer's financial needs, from checking accounts to life insurance and one-stop shopping for financial services will become common. On the demand side, consumers want to spend less and less time with a financial service provider; electronic banking and e-trading have reduced the opportunity for personal selling and the Internet has made information search less costly and financial markets more transparent to consumers. These developments have stimulated banks to shift from a product focus to a customer focus. As the cost of acquiring new customers increases, financial institutions are coming to the conclusion that their current customers are by far the best prospects for the sales of current and new services and attempt to consolidate service sales from their customers by implementing customer relationship management. DBM is viewed in banking as one of the most powerful marketing tools, but its success depends on the availability of databases (Onge, 1999). The level of penetration of electronic banking has propelled electronic storage of customer transactions, which is now routine in the entire financial industry. Therefore, the financial services industry presents all conditions to the successful implementation of DBM.

4.2. Internal and external data

In order to illustrate the proposed approach for the cross-selling of services, we apply it to a sample of 5550 customers of a major commercial bank in Brazil.

For each of these sampled customers, we have data that were gathered in a personal interview, as well as transaction data from the bank's internal records. For this particular study, we use the following variables from the bank's internal records (assumed distribution in parenthesis):

- Number of transactions/month (Poisson)
- Volume of deposits in the bank (normal)
- Education (rank-order binomial)
- Age (rank-order binomial)
- Gender (Bernoulli)
- Ownership of automobile, telephone, fax, and personal computer (Bernoulli)
- Personal income (rank-order binomial)
- Usage indicators for 22 financial services within the bank (Bernoulli). These include four types of services:
  - *Conveniences*: ATM card, phone banking, PC banking, safety box, private manager, and automatic bill payment
  - *Investments*: special checking, savings, certificate of deposit, mutual fund, annuities fund, investment fund, commodities fund, and gold
  - *Risk management*: life insurance, car insurance, and homeowner's insurance
  - *Credit*: mortgage, installment loan, credit card, personal loan, and farming credit

These internal data are supplemented with survey data on each customer's usage of the same 22 financial services from competing vendors (Bernoulli). Note that most of these financial services can be owned from multiple banks by the same customer. Table 1 provides a summary description of the variables in the study.

For this application, we use the complete data on the sample of 5550 customers, since we want to validate our procedure. We estimate the proposed mixed data factor analyzer on a sample of 1387 of these customers. This sample is a random sample taken from all customers and is representative of the entire database and is large enough for reliable estimation of the parameters of our model in reasonable computation times. We then apply the estimated model to the remaining 4163 customers, for whom we assume the survey data on competitive ownership to be missing. We thus predict their likelihood to use each of the 22 services from competing firms, based solely on these customers' internal records. This is an important problem for the bank in itself, since competitive ownership is only known for a subset of its customers and our procedure allows one to forecast it for all customers in the database. Since in our application we have the survey data for the hold-out customers as well, this allows us to investigate the performance of the procedure, by comparing the imputed values to the "true" values of the survey variables. Our objective is to demonstrate that, once the model is estimated on a combination of internal and external data for a sub sample, it can be applied to the firm's entire customer database to predict whether customers satisfy their needs for specific financial services elsewhere.

4.3. Results

Estimation of the factor model to the data from both sources leads us to choose the model with three factors ($P=1$: CAIC = 81,899, $P=2$: CAIC = 76,571, $P=3$: CAIC = 75,622, $P=4$: CAIC = 75,872). In addition to this model, we also estimated the latent-trait model previously proposed by Kamakura et al. (1991), as well as a three-dimensional binary data factor model (Bartholomew & Knott, 1999) for comparison. Note that our application of the latent trait model proposed by Kamakura et al. is an extension of their approach, because we augment the customer database with external (survey) data and utilize the model to make predictions about usage of the services from competing vendors.

The factor weights representing the reduced space map are graphically displayed as vector termini in Figs. 3 and 4, classified by type of service (credit, investment, risk management, or convenience services). Since pictorial information is more quickly processed and better remembered than verbal or numerical information (Spence & Lewandowski, 1990), the graphs allow for efficient communication with bank managers, who can quickly grasp the dependencies of service usage and identify implied cross-selling opportunities. The graphical display facilitates dissemination of the results within the company. Thus, we emphasize low-dimensional representation of the data and graphical display, rather than substantive interpretation of the factors themselves.

Table 1
Summary statistics

	Frequency	Valid percent
Education		
Incomplete elementary	265	4.8
Elementary	471	8.5
Junior high	424	7.7
High school	1282	23.2
College or more	3079	55.8
Total	5521	100.0
Age		
20 or less	53	1.0
21–30 years	719	13.0
31–40 years	1272	23.0
41–50 years	1514	27.4
51–60 years	1080	19.5
More than 60 years	896	16.2
Total	5534	100.0
Monthly income		
<US$400	784	14.2
US$400–699	784	14.2
US$700–999	678	12.3
US$1000–1499	745	13.5
US$1500–1999	546	9.9
US$2000–2999	727	13.2
US$3000–5000	846	15.3
>US$5000	418	7.6
Total	5528	100.0

Service	*N*	Mean	Standard deviation
Number of transactions/month	5550	43.81	38.64
Volume of deposits in the bank	5550	6336.72	24,698.26
Gender	5538	65.1%	47.7%
Own an automobile	5550	83.6%	37.0%
Own a phone	5550	83.4%	37.3%
Own a fax	5550	12.4%	33.0%
Own a PC	5550	29.5%	45.6%
Savings (competitor)	5550	60.5%	48.9%
Savings	5550	54.7%	49.8%
Credit card (competitor)	5550	34.1%	47.4%
Credit card	5550	17.0%	37.6%
ATM card (competitor)	5550	60.6%	48.9%
ATM card	5550	81.7%	38.6%
Phone banking card (competitor)	2750	35.8%	48.0%
Phone banking card	2750	21.2%	40.9%
CD (competitor)	5550	27.2%	44.5%
CD	5550	26.1%	43.9%
Special checking (competitor)	5550	55.1%	49.7%
Special checking	5550	60.3%	48.9%
Safety box (competitor)	2750	8.5%	28.0%
Safety box	2750	6.0%	23.8%
PC banking (competitor)	2750	8.4%	27.8%
PC banking	2750	4.9%	21.6%
Auto bill payment (competitor)	5550	39.5%	48.9%
Auto bill payment	5550	11.1%	31.4%
Personal loans (competitor)	5550	7.2%	25.9%
Personal loans	5550	7.5%	26.3%
Mortgage (competitor)	5550	7.0%	25.4%
Mortgage	5550	1.3%	11.2%
Installment loan (competitor)	5550	4.0%	19.6%
Installment loan	5550	1.4%	11.6%
Farming credit (competitor)	2750	2.1%	14.2%
Farming credit	2750	1.4%	11.8%
Mutual fund (competitor)	2750	14.4%	35.1%
Mutual fund	2750	9.3%	29.1%
Investment fund (competitor)	5550	38.4%	48.6%
Investment fund	5550	34.3%	47.5%
Commodities fund (competitor)	5550	22.1%	41.5%
Commodities fund	5550	15.5%	36.2%
Annuities fund (competitor)	5550	9.8%	29.8%
Annuities fund	5550	6.8%	25.2%
Private manager (competitor)	2750	10.3%	30.4%
Private manager	2750	4.0%	19.6%
Gold (competitor)	2750	8.0%	27.1%
Gold	2750	6.3%	24.3%
Car insurance (competitor)	5550	17.3%	37.8%
Car insurance	5550	8.8%	28.3%
Home insurance (competitor)	5550	11.2%	31.5%
Home insurance	5550	7.0%	25.5%
Life insurance (competitor)	5550	21.5%	41.1%
Life insurance	5550	25.9%	43.8%

The plots allow managers to quickly identify effective strategies for cross-selling, i.e. they enable managers to target services to customers who currently use them from competitors or have a high predicted probability of usage, but have not yet acquired the service within or outside the bank. Those services are identified from the similarities in the weights of the internal and external service items in Figs. 3 and 4, respectively. For example, the fact that the vectors representing *credit card (crdt crd)* and *personal loans (prsnl loan)* within the bank are close

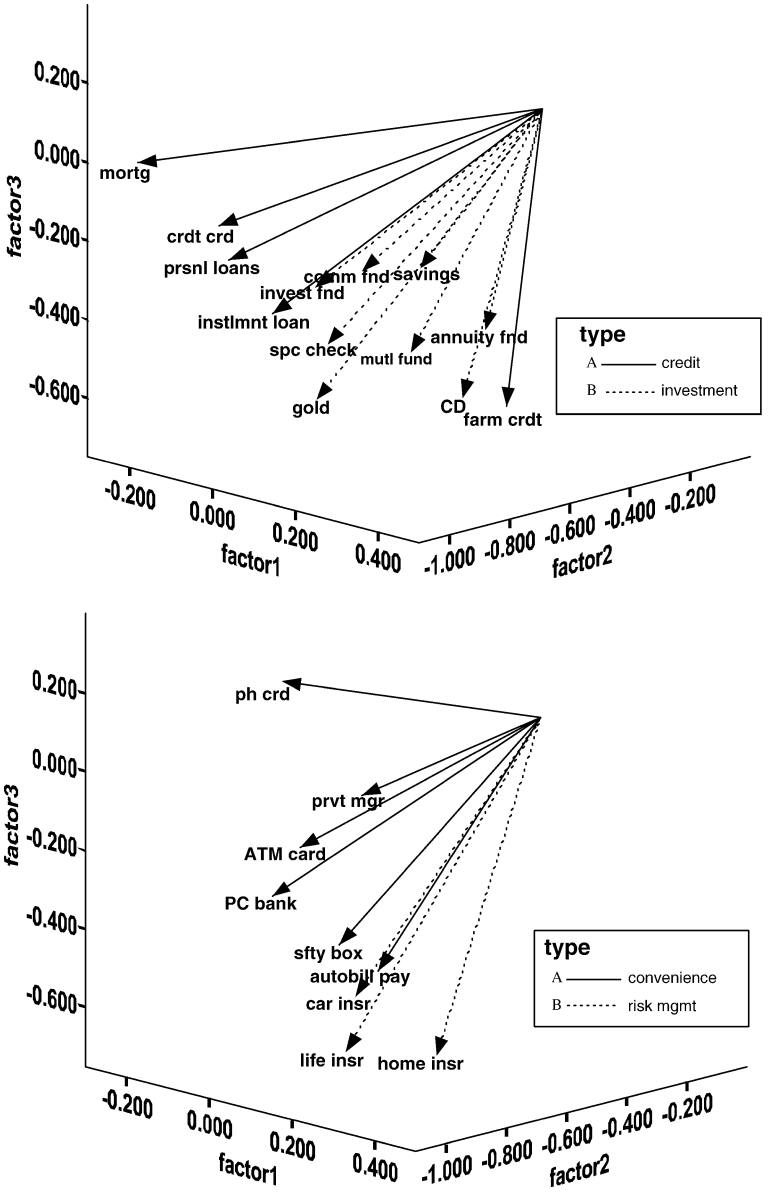

Fig. 3. Weights for service usage outside the bank.

to each other indicates that a customer who uses *personal loans* with the bank would be a good prospect for the bank's *credit card*, if she does not use it yet. The same conclusion can be drawn for *home insurance (home insr)* and *car insurance (car insr)*, indicating cross-selling opportunities.

Similarly, the fact that the vectors representing *annuities funds (annuity fnd)* within and outside the bank point in the same direction indicates that a customer with a high propensity to use this service might have it from multiple sources. Therefore, s/he would represent a good prospect for strategies that

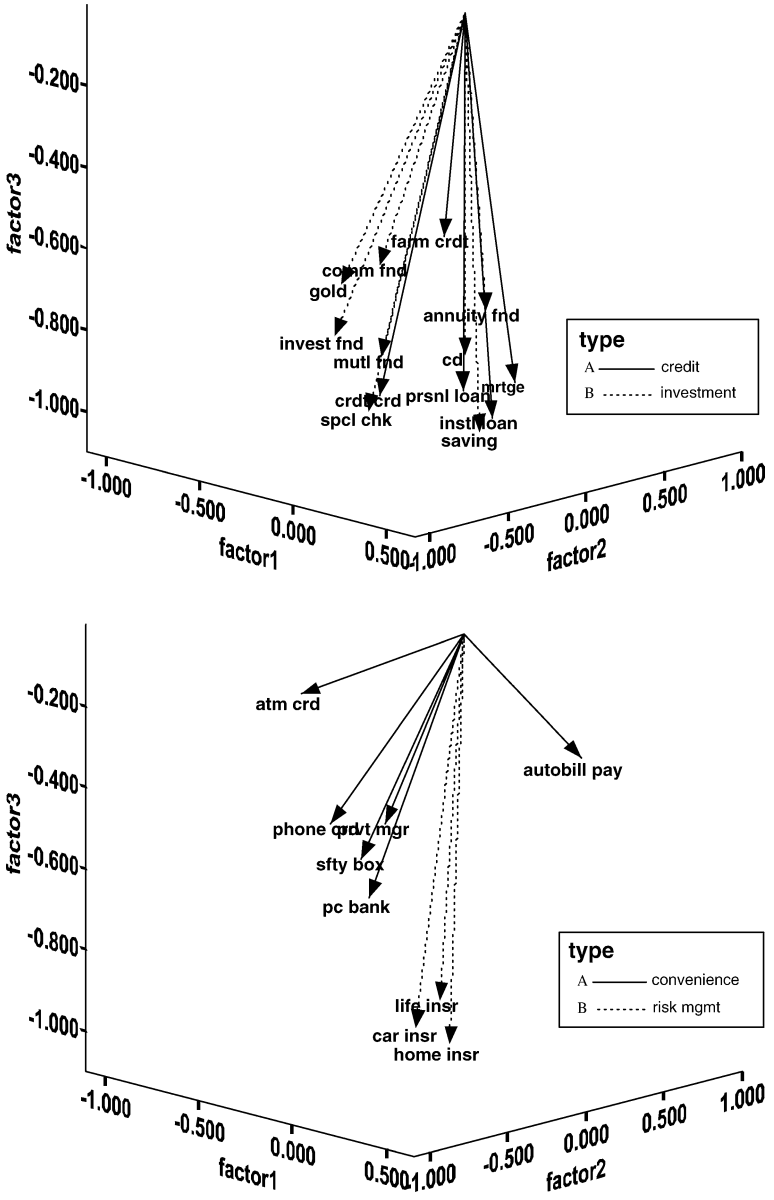

Fig. 4. Weights for service usage within the bank.

induce switching (if s/he does not yet use the service within the bank) or a higher "share of wallet" (if s/he already uses the bank's annuity fund). The opposite conclusion can be drawn about *home mortgage (mortg)*; the maps indicate that usage of this service within the bank is unrelated to usage at another institution. These patterns of joint usage of these services within and outside the bank can be potentially useful to develop a cross-selling program. However, for such a program to be effectively implemented, the question arises to what extent the model makes use of the inter-relationships among these

services to produce accurate predictions, which we consider next.

Fig. 5 shows how the latent space relates to customer demographics. Comparing this figure with the previous ones, one can see that usage of financial services in general is correlated to the demographic variables that indicate income (income, education, ownership of durables, etc.). As one would expect, ownership of information-technology durables (PC and fax) is highly collinear with education and volume of deposits is highly collinear with the number of transactions per month.

4.4. Out of sample tests and comparison

As a validation test, we simulate its application in the identification of customers who use various financial services at competing financial institutions. With this particular purpose in mind, we compute the predictive probabilities that each customer uses the financial services outside the bank for each of the 4163 customers in our hold-out sample, using only information from the internal records as in Eq. (7). Again, this is an important problem for banks since competitive ownership tends to be available only for a sample of the customers based on a survey and our procedure allows one to predict it for all subjects in the transaction database. Since in our particular application we also have the survey data available for the hold-out customers, this enables us to validate the predictive performance out of sample using the actual information obtained from the survey.

As a measure of selectivity of the predictive model, we generate power curves of the cumulative proportion of actual users of the service (observed data) against the cumulative proportion of customers with a certain predicted usage probability for that service (out of sample forecasts). These power curves are shown in Fig. 6, comparing the performance of the three models for selected financial services from the database. The plot for the *private manager* service, for example, shows that the 30% of customers that are predicted by the three-factor binary model to have the highest probability (based solely on service usage data within the bank), account for more than 70% of all users of that service outside the bank. For our proposed model, that number is 80%. For most services, the power curve for our model lies above that for the

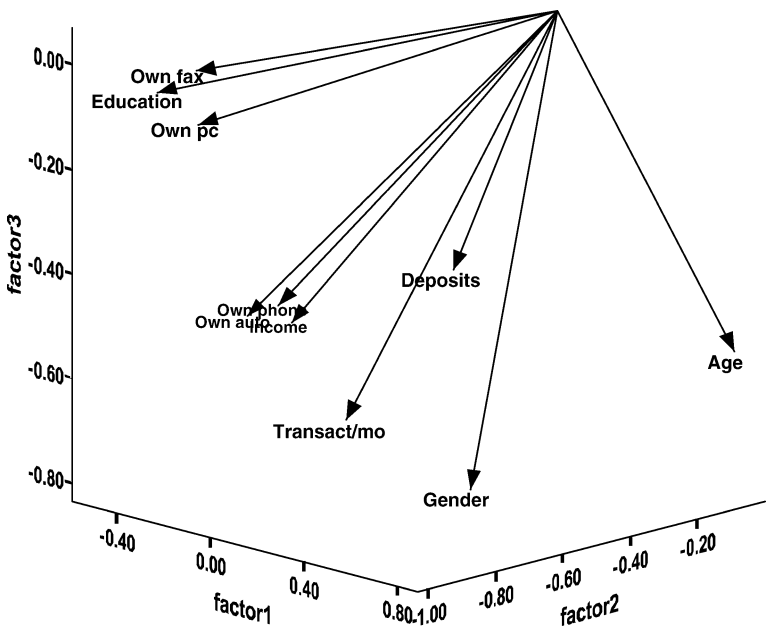

Fig. 5. Weights for other internal variables.

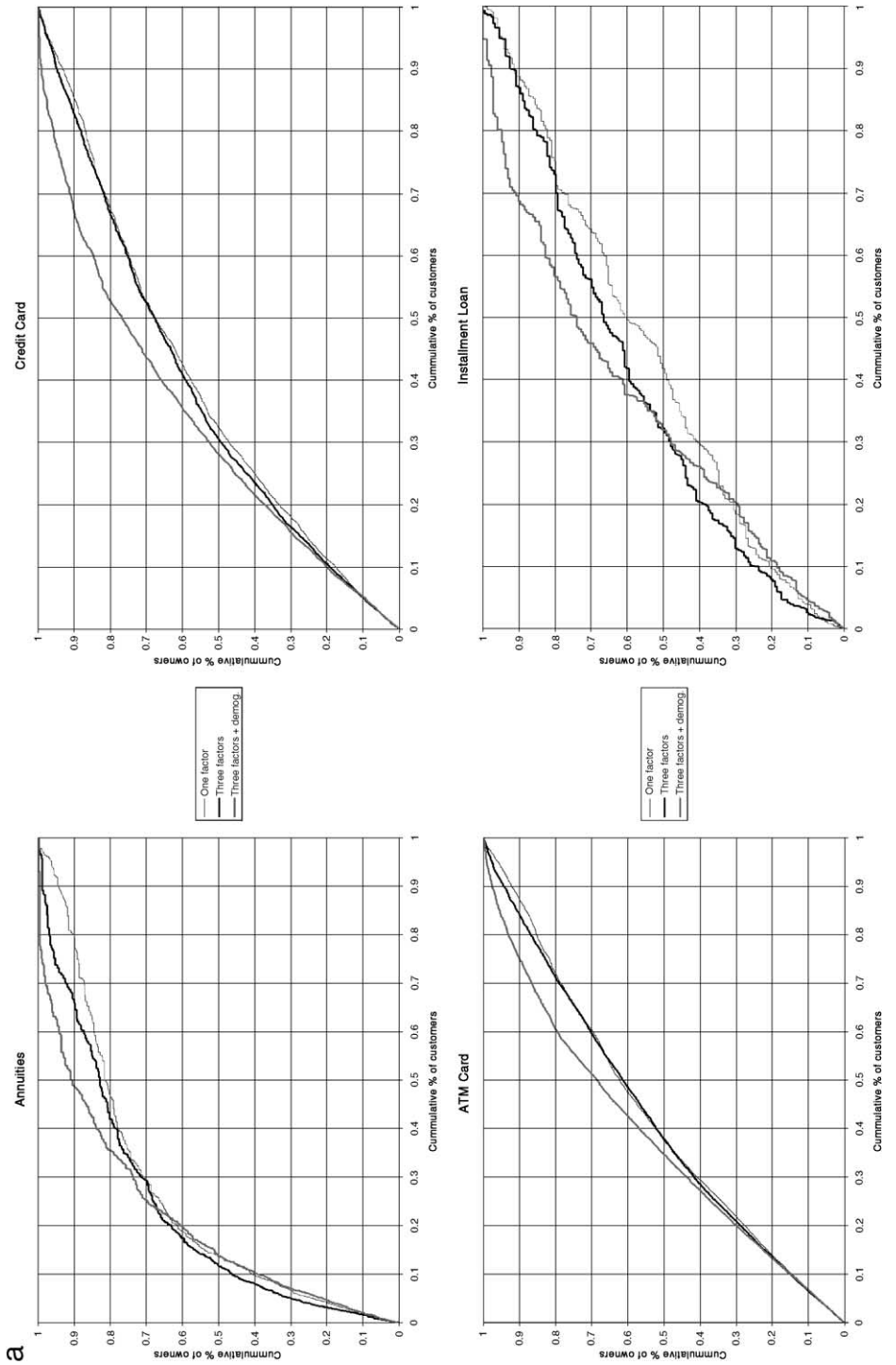

Fig. 6. Power curves for service usage outside the bank.

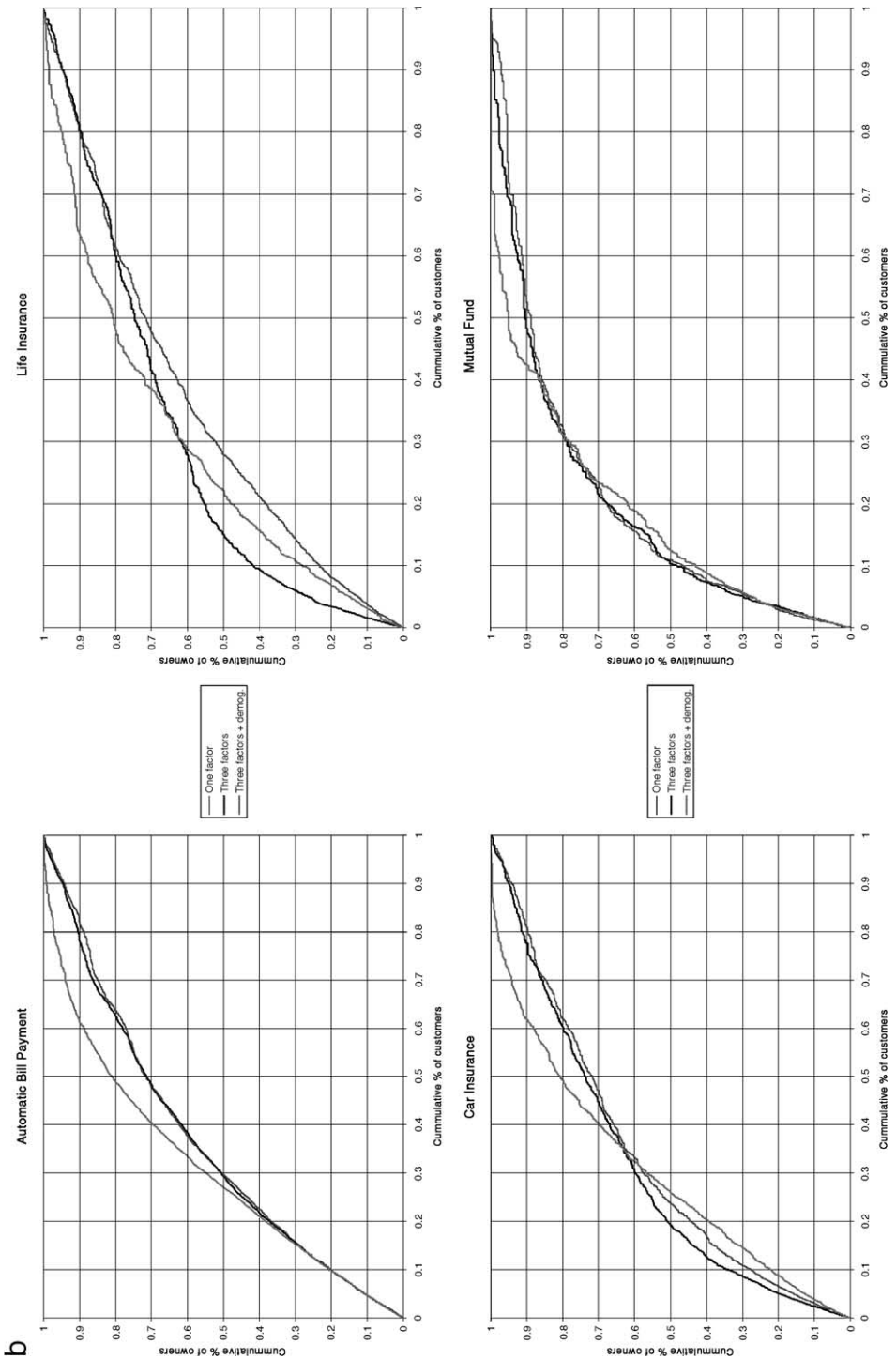

Fig. 6 (*continued*).

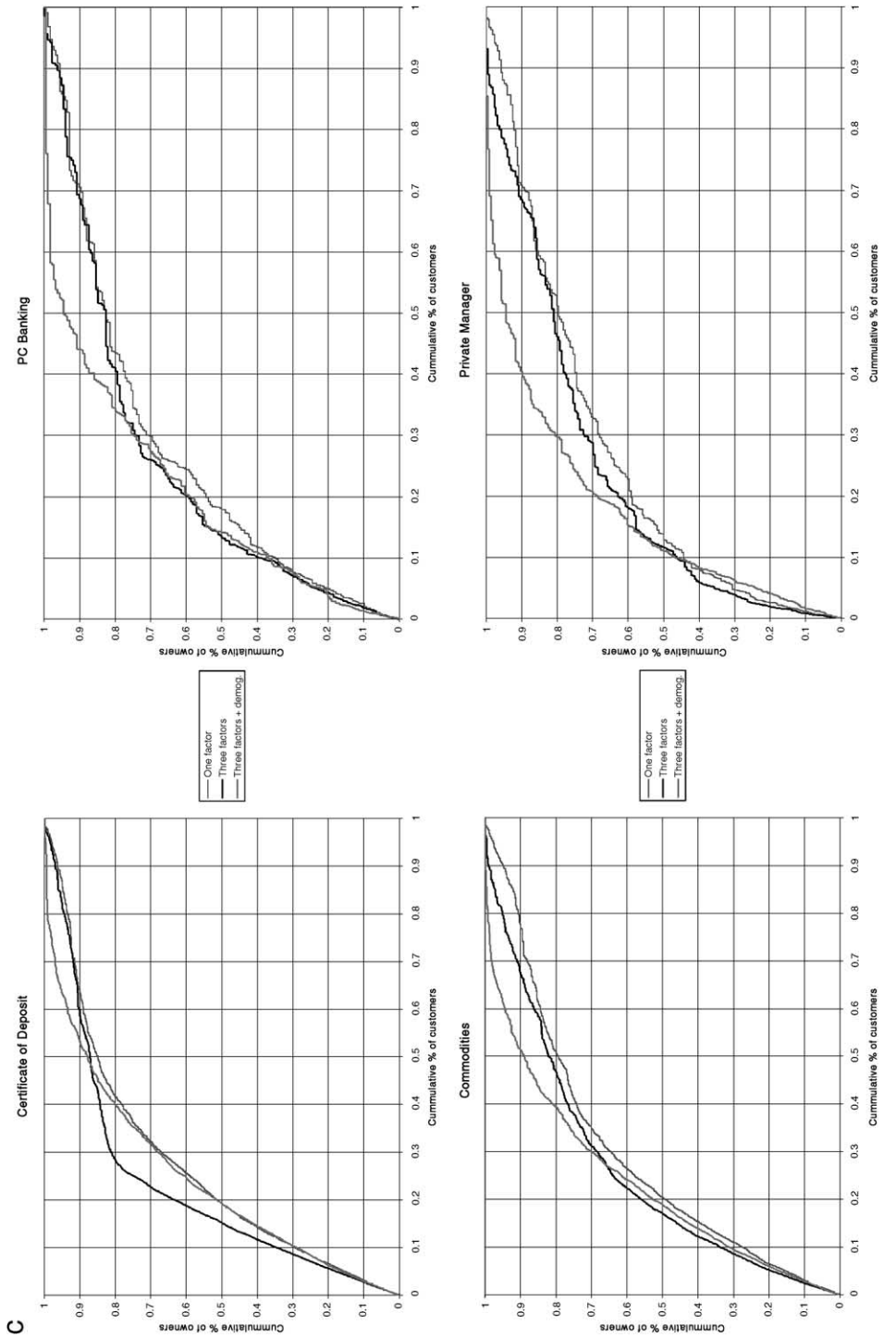

Fig. 6 (*continued*).

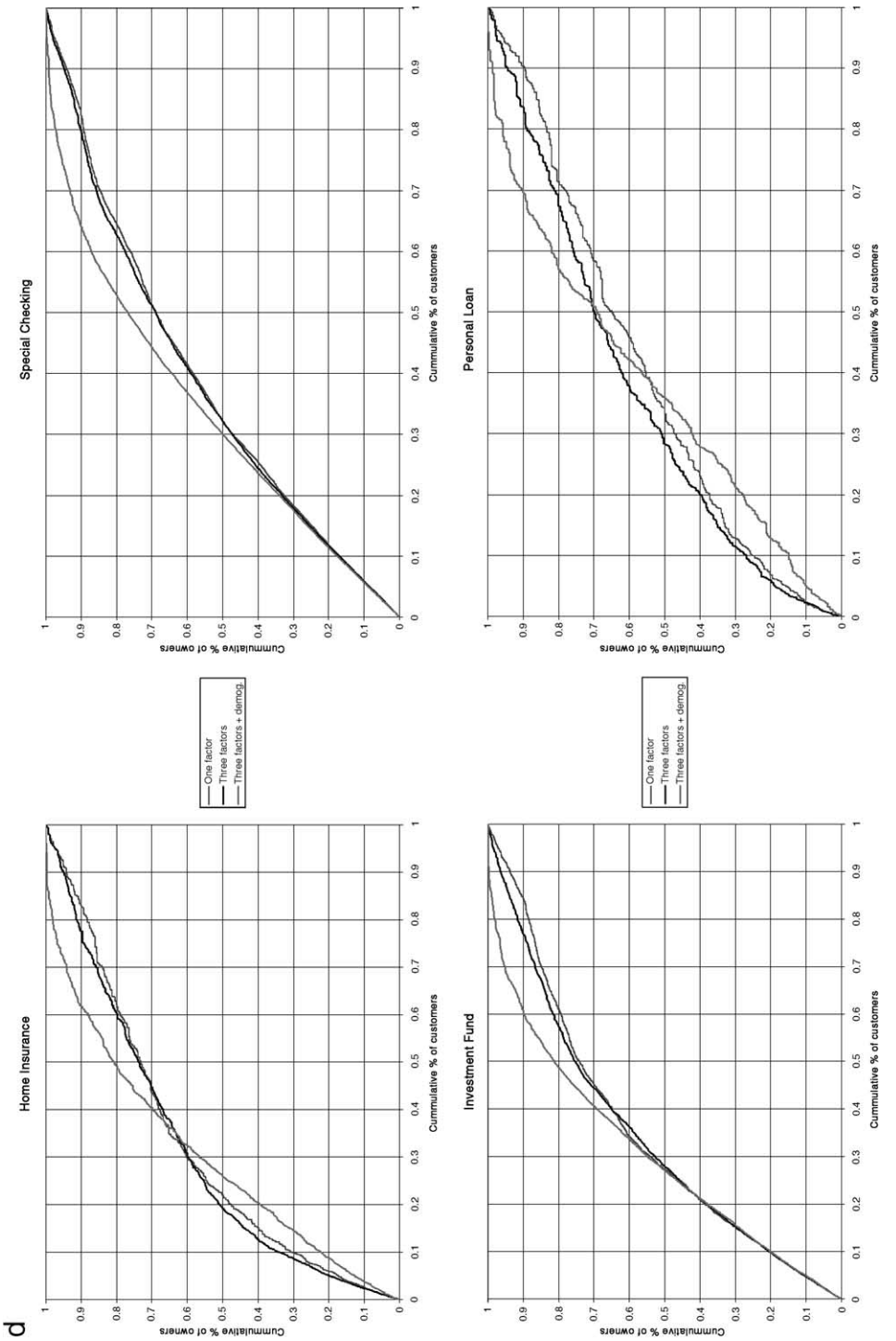

Fig. 6 (*continued*).

two competing models. We note that large improvements are attained for cutoffs around 40–50%. This means that the improvement of our model over the two competing models increases substantially if cross-selling is considered for larger proportions of the database. Note that this holds for any predictive model because, as the penetration for a product/service reaches saturation, there is not much to gain in using a more discriminating predictive model.

As a summary measure of selectivity, we compute the Gini coefficient from the power curve for each of the services. The Gini coefficient is a measure of concentration, indicating the extent to which usage of the service outside the bank is concentrated among those customers who were predicted to have a high probability of doing so. We compute this index for service j as $G(j) = \sum_n v_{nj} - \hat{v}_{nj} / \sum_n 1 - \hat{v}_{nj}$, where $\hat{v}_{nj}$ is the proportion of the sample of customers who have a predicted probability of usage for the service equal or greater than customer n's, and v_{nj} is the proportion of actual users of the service who are ranked equal or higher than customer n in their usage probability. This measure of concentration equals the ratio of the area between the power curve and the 45° line over the total area above this line. An index equal to zero indicates a lack of predictive power, while a value of one is obtained if the model sorts all customers perfectly in decreasing order of true likelihood of usage. Table 2 shows the Gini coefficients for all 22 financial services. The table reveals that our model yields a substantially higher Gini coefficient than the $P=1$ latent trait model in all cases, but one. It yields a higher Gini coefficient than the $P=3$ latent trait model in all cases, but four. Thus, the predictive performance of our model is superior to that of the two competing models (note also that the application of the three-dimensional binary factor model to DBM and imputation of missing observations with each of the three models is itself new and that we had to extend the two competing procedures to deal with the missing data structure of the application to enable comparison). In absolute terms, the Gini coefficient is very high ($G(j)>0.6$) for at least six services (safety box, PC banking, farming credit, mutual fund, private manager, and gold) and reasonably high for at least six others ($G(j)>0.4$). These services appear from our analyses to be important for developing cross-selling activities.

Table 2
Gini coefficients for service usage outside the bank

Service	$P=1$, binary	$P=3$, binary	Full model
Savings	0.11	0.13	0.18
Credit card	0.22	0.24	0.35
ATM card	0.15	0.16	0.24
Phone banking card	0.36	0.47	0.44
CD	0.47	0.55	0.52
Special checking	0.24	0.25	0.34
Safety box	0.64	0.66	0.73
PC banking	0.48	0.52	0.61
Auto bill payment	0.28	0.29	0.39
Personal loans	0.21	0.29	0.26
Mortgage	0.15	0.16	0.26
Installment loan	0.14	0.24	0.30
Farming credit	0.56	0.55	0.63
Mutual fund	0.60	0.62	0.64
Investment fund	0.30	0.32	0.39
Commodities fund	0.41	0.48	0.54
Annuities fund	0.48	0.55	0.59
Private manager	0.49	0.55	0.66
Gold	0.72	0.72	0.68
Car insurance	0.34	0.38	0.40
Home insurance	0.35	0.41	0.43
Life insurance	0.30	0.38	0.37

4.5. Identifying the best prospects for cross-selling

The best prospects for the cross-selling of a particular service are those customers who have a high predicted propensity to use the service within the bank, but do not yet use it. However, these customers may already use the (same) service at a competing financial institution and, therefore, must be persuaded to switch service providers. While the bank does not have perfect information as to whether these customers use the service at a competing institution (it has that information only for the sample of subjects included in the survey), it can use our factor analyzer to compute their propensity to do so and the predicted probability. Fig. 7 illustrates this for four different services, showing the predicted probabilities of usage within and outside the bank, among all customers who do not use the service within the bank, ranked in decreasing order of their potential as cross-selling prospects. For example, Fig. 7 shows that the top 30% prospects for the cross-selling of *phone banking cards* have predicted usage probabilities that are equal or greater than 80%. On the other hand, these same customers have very similar probabilities of being current users of the service at a competing institution. The situation is

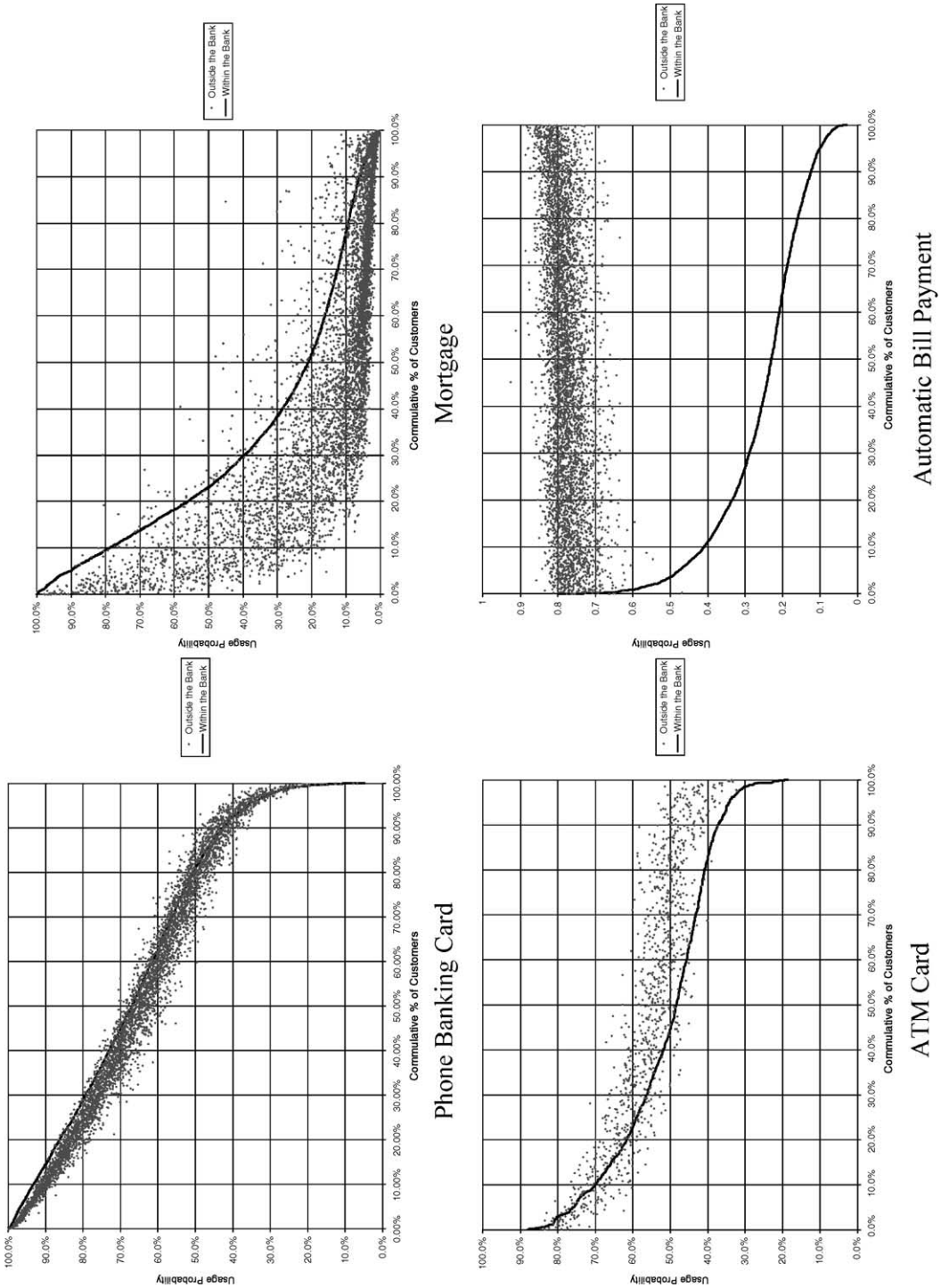

Fig. 7. Usage probabilities among non-using customers.

more severe for *automatic bill payment*. For this service, the top 30% prospects for cross-selling have predicted usage probabilities of only 30% or greater. Moreover, most customers have a higher probability of being current users at a competing institution than to use this service within the bank. Therefore, the selection of prospects for cross-selling purposes will depend on whether the bank can persuade these customers to switch away from competitors.

5. Summary and conclusions

The information revolution in the past couple of decades has caused a proliferation of customer databases, often leading to injudicious applications of direct marketing techniques, canvassing the market with ineffective sales pitches, increasing consumer resistance to "junk mail" and telemarketing, and reducing the profitability of marketing activity. Appropriate use of DBM enables firms to effectively leverage on knowledge about current customers. This maximizes the yield of the sales effort, minimizes the risk of annoying the customer with uninteresting offers, and strengthens the ties between the firm and the customer. As a consequence of the information revolution, in particular in the financial sector, firms have amassed vast amounts of behavioral and demographic data about their customers in data warehouses. Effectively utilizing this source of information requires the application of methods concisely tailored to the requirements posed by the structure and contents of the database and the marketing needs of the company.

In this study, we have presented a novel tool to support cross-selling with database marketing that we believe meets those needs. The method is tailored to a situation where the transaction database is augmented with information on ownership of products and services from competitors, collected through a survey. Our mixed data factor analyzer allows the firm to predict its customers' likely buying behavior beyond the products and services currently owned from the firm. These probabilistic predictions form the basis for selecting the best prospects for the cross-selling of new products or services. The flexibility of the approach in matching the assumed distribution with the measurement scale of the observed variables is particularly important for cross-selling predictions to be logically consistent. A major advantage of the factor analyzer for strategic use within companies is that its use and implementation is almost entirely based on graphical representation of the dependencies in the data. The identification of services for effective cross-selling and the selection of cross-selling targets can all be accomplished by graphical means, which greatly facilitates ease and speed of use by managers.

The estimation of the proposed model on a random sample combining data from the customer database and from external sources is computationally intensive, due to the reliance on simulation-based estimation methods. Calibration of the model on complete transaction databases of millions or even several hundreds of thousands of customers is currently not feasible. With current technology, it is feasible to estimate the model on sample sizes in the order of tens of thousands, which are typical of large-scale customer surveys. However, the model need not be calibrated on the entire database but only on a sample from it, which, as we have shown, yields reliable estimates of the model parameters. Once the model is calibrated on the sample, the implementation on the entire database is relatively fast and easy. In other words, scalability is not a main concern in the implementation of the proposed model, as long as it can be calibrated on a sample of the customer database.

References

Anderson, E. B. (1980). *Discrete statistical models with social science applications*. New York: North Holland.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. Oxford: Edward Arnold.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Day, G. S. (2000). Capabilities for forging customer relations. *MSI Working Paper 00-118*, MSI, Cambridge, MA.

Goodman, J. (1992). Leveraging the customer database to your competitive advantage. *Direct Marketing*, *55*(8), 26–27.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, et al. (Eds.), *Measurement and prediction: studies in social psychology in World War II*. Princeton, NJ: Princeton University (pp. 60–90).

Hebden, J. J., & Pickering, J. F. (1974). Patterns of acquisition of consumer durables. *Oxford Bulletin of Economics and Statistics*, *36*, 67–94.

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for

cross-selling of financial services. *International Journal of Research in Marketing*, *8*, 329–349.

Kamakura, W. A., & Wedel, M. (2000, November). Factor analysis and missing data. *Journal of Marketing Research*, *37*, 490–498.

Kasulis, J. L., Lusch, R. F., & Stafford Jr., E. F. (1979, June). Consumer acquisition patterns for durable goods. *Journal of Consumer Research*, *6*, 47–57.

Labe Jr., R. P. (1994). Database marketing increases prospecting effectiveness at Merrill Lynch. *Interfaces*, *24*(5), 1–12.

Lee, L. -F. (1995). Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. *Econometric Theory*, 437–483.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman and Hall.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrika*, *57*, 995–1026.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411.

Onge, J. S. (1999). Direct marketing credos for today's banking. *Direct Marketing*, *61*(11), 56–58.

Paroush, J. (1965). The order of acquisition of consumer durables. *Econometrica*, *33*(1), 225–235.

Spence, I., & Lewandowski, S. (1990). Graphical perception. In J. Fox, & J. S. Long (Eds.), *Modern methods of data analysis*. London: Sage (pp. 1–57).

Shaw, R. (1993). Making database marketing work. *Journal of Information Technology*, *8*, 110–117.

Shesbunoff, A. (1999). Winning CRM strategies. *ABA Banking Journal*, *91*(10), 54–58.

Stafford, E. F., Kasulis, J. J., & Lusch, R. F. (1982). Consumer behavior in accumulating household financial assets. *Journal of Business Research*, *10*, 397–417.

Stern, S. (1997, December). Simulation-based estimation. *Journal of Economic Literature*, *35*, 2006–2039.

Wedel, M., & Kamakura, W. A. (2001). Factor analysis with observed and latent variables in the exponential family. *Psychometrika*, *66*(4), 515–530.