Statistica Neerlandica (2003) Vol. 57, nr. 1, pp. 46–57

List augmentation with model based multiple imputation: a case study using a mixed-outcome factor model

Wagner A. Kamakura[1] and Michel Wedel[2]*

[1]*Fuqua School of Business, Duke University, P.O. Box 90120, Durham, NC 27708, USA*

[2]*University of Michican Business School, 701 Tappan Street, Ann Arbor, MI 48109, USA*

This study concerns list augmentation in direct marketing. List augmentation is a special case of missing data imputation. We review previous work on the mixed outcome factor model and apply it for the purpose of list augmentation. The model deals with both discrete and continuous variables and allows us to augment the data for all subjects in a company's transaction database with soft data collected in a survey among a sample of those subjects. We propose a bootstrap-based imputation approach, which is appealing to use in combination with the factor model, since it allows one to include estimation uncertainty in the imputation procedure in a simple, yet adequate manner. We provide an empirical case study of the performance of the approach to a transaction data base of a bank.

Key Words and Phrases: factor analysis, simulated likelihood, multiple imputation.

1 Introduction

Observations that are missing due to the design of studies are the rule rather than the exception in marketing. Examples occur in data fusion, split questionnaires, time sampling and sub-sampling. Figure 1, taken from KAMAKURA and WEDEL (2000), presents the structure of these missing data-designs. In each of these conditions, data are *missing intentionally,* i.e., they are specified missing in the design of the study to reduce respondent burden, increase the response rate or reduce costs of data collection. When data are intentionally missing, they are usually MAR (Missing at Random) and the Missing Data Generating Mechanism is ignorable (LITTLE and RUBIN, 1987, SCHAFER, 1997).

In this paper we focus on the particular problem of sub-sampling; a problem that often occurs in Database Marketing (DBM). DBM involves building, organizing, supplementing and mining customer transaction databases to increase the accuracy

*m.wedel@eco.rug.nl

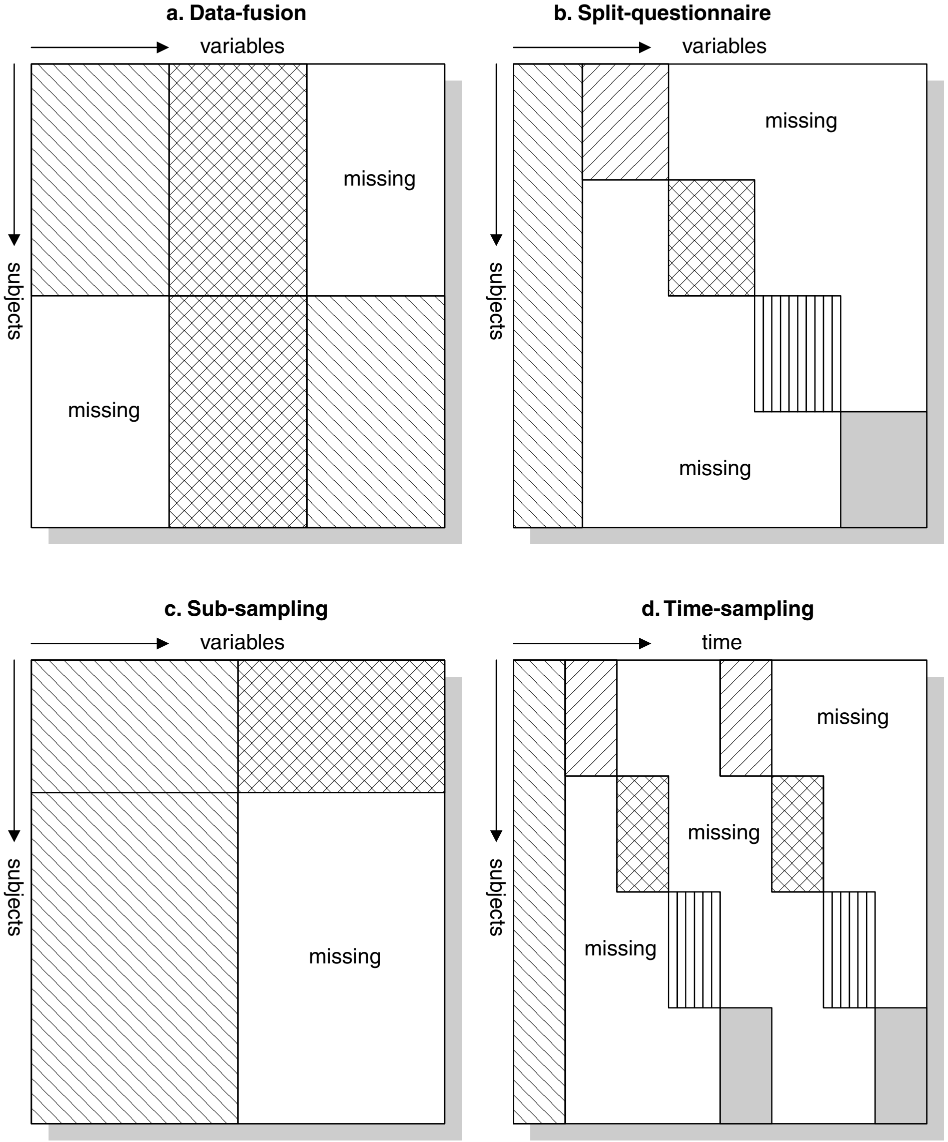

Fig. 1. Marketing data missing by design.

of marketing efforts (Goodman, 1992). Many companies nowadays record their transactions with each individual customer and store those in customer transaction databases. Such practice is quite common among firms in the financial, leisure and telecommunications sectors, and in particular among online (Internet) companies. A properly compiled, cleaned and maintained transaction database can be used as a list for targeted marketing efforts. For those purposes, transaction data are enriched with supplementary data. ZIP-level geo-demographic and life-style data are often used for that purpose. Customer transaction records, often consisting of purchase indicators or counts, are linked to the additional data at the individual customer level based on ZIP-codes. In many cases data on the use of products and services from competitors, and "soft data" such as customer satisfaction, provide important additional insights to the company, but are lacking in the transaction database and 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. The required set of supplementary variables is measured on a sample from the database, frequently using metric or rank-order measurement scales. Such sub-sampling for soft data collection is applied regularly in direct marketing, and a similar situation occurs when test-mailings are conducted for a sub-sample of the house-list. The information from these surveys or test mailings, however, is needed for all customers and the list needs to be augmented with the auxiliary data.

List augmentation is a special case of missing data imputation, or data fusion. From a statistical perspective, problems in list augmentation are presented by the enormous amount of missing data and by the vastly different measurement scales of the data. Usually the sub-sample is small relative to the size of the database. However, as is apparent from Figure 1, the structure of the list augmentation problem may yet lend itself to effective imputation. The reason is that the quality of the imputations is strongly dependent on the common fusion variables, the variables that are present in both the complete and incomplete data records. Unlike in traditional data fusion problems where one has to rely on relatively weak demographic fusion variables, in list augmentation the customer transaction data themselves serve as fusion variables, since they are often strongly associated with the partially observed auxiliary data.

Extensive overviews of methods for imputing missing data have been presented by, for example, LITTLE and RUBIN (1987) and SCHAFER (1997). State-of-the art imputation methods are model-based, and involve multiple imputations, drawn from the predictive distribution of the data, given the model estimates. Thus imputation models need to be specified, which allows one to impute the missing data multiple times, where multiple imputation allows one to gauge the accuracy of the imputation procedure.

Model-based imputation using regression-type models suffers from the drawback that the model needs to be tailored to the particular data in question on a case-by-case basis, which has limited the use of (designed) incomplete data in large-scale marketing problems. SCHAFER (1997) provided a general set of approaches to imputation, based on the full covariance matrix of the observed variables. Along these lines, we present an approach to impute high-dimensional data based on a parsimonious factor representation. LITTLE and RUBIN (1987, pp. 148–149) already addressed the problem of missing data in factor analysis. Our approach builds upon LITTLE and RUBIN (1987), SCHAFER (1997), KAMAKURA and WEDEL (2000) and WEDEL and KAMAKURA (2001). The latter authors proposed to use a factor model for data imputation. We conjecture that the factor model parsimoniously captures the covariance of completely and partially observed data, required for augmentation, while it deals with the mixed measurement scales for the observed data. We first review the factor model and propose a nonparametric Bootstrap-based imputation approach. The Bootstrap approach is appealing to use in combination with the factor model, since it allows one to include estimation uncertainty in the imputation procedure in a simple, yet adequate manner. We apply the imputation model and procedure in a case study on list segmentation.

2 Mixed outcome factor model for list augmentation

2.1 *Model description*

We assume that a firm has conducted a survey among a random sample of its customers. Data from this sample survey is collected to augment the customer list. Let $n = 1,\ldots,N$ denote customers and $j = 1,\ldots,J$ variables. The J variables are measured on different scales, such as ordinal scales, ratio-scales, counts, and binary scales. We assume the J observations, $y_j = (y_{nj})$, to be realizations of random variables, distributed according to a member of the exponential family (McCullagh and Nelder, 1989). This allows us to accommodate the types of data encountered in house-lists in a single framework, by assigning each observed variable j its own distribution, optimally matching the support of the selected distribution to the measurement scale of each of the transaction variables.

We aim at providing a low dimensional map of the observed variables $y_n = (y_{nj})$ from subject n. We assume that the J observations on each individual, y_{nj}, are conditionally independent given factor scores x_{np}, and specify the conditional distribution as $f_Y(y_n|\eta(x_n)) = \prod f_{Y_j}(y_{jn}|\eta(x_n))$, with $E\lfloor Y_j|x_n\rfloor = h_j(\eta(x_n))$, where $h_j(\cdot)$ is the canonical link function^{j}for variable j, and

$$\eta(x_n) = \lambda_0' + x_n\Lambda'. \tag{1}$$

In equation (1) x_n is the nth row of an unobserved vector of i.i.d. normally distributed $(N \times P)$ factor scores X, Λ a $(J \times P)$ matrix and λ_0 a $(J \times 1)$ vector of fixed, but unknown, weights. The expectation of the random outcome vector for each subject is mapped onto a lower P-dimensional subspace, $\eta(x_n)$ defining that map, its dimensionality being unknown a priori. We specify the 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 factor model. They arise because one can add a vector of scalars to x_n and subtract it from λ_0. Rotation invariance arises because one can post-multiply x_n and Λ with an orthogonal rotation matrix T. Without imposing further constraints on Λ (i.e., setting a $(P \times P)$ sub matrix to the identity matrix), the factor model is not identified.

Our model fits in with 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). Note that the distribution in (1) presents the conditional distribution of Y, given the latent variables X. For the sake of illustration, 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, occurring in the application below. Then the conditional distribution of the observed data is

$$\begin{aligned} f_Y(y_n|\eta(x_n)) = {} & \prod_{j=1}^{J_N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\sigma^{-2}(y_{nj} - \eta_{nj})^2\right] \prod_{j=J_N+1}^{J_N+J_B} \frac{\exp[y_{nj}\eta_{nj}]}{1+\exp[\eta_{nj}]} \\ & \times \prod_{j=J_N+J_B+1}^{J_N+J_B+J_P} \frac{\exp[y_{nj}\eta_{nj} - \exp[\eta_{nj}]]}{y_{nj}!} \\ & \times \prod_{j=J_N+J_B+J_P+1}^{J_N+J_B+J_P+J_R} \binom{K_j - 1}{y_{nj} - 1} \frac{\exp[(y_{nj}-1)\eta_{nj}]}{(1+\exp[\eta_{nj}])^{K_j-1}}. \end{aligned} \tag{2}$$

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

2.2 *Estimation and selection*

We partition the observation vector as $y_n = (\overset{\frown}{y}_n, \breve{y}_n)$, with the corresponding sets of variables being $C = \overset{\frown}{C} \cap \breve{C}$, where we assume the first subset of variables to be observed. Also, without loss of generality, 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 as

$$L\left(\Xi|\breve{Y}\right) = \prod_{n=1}^{M} \int \prod_{j=1}^{J} f_Y\left(\breve{y}_{nj}|\eta(x_n), \Xi\right) f_X(x_n)dx_n. \tag{3}$$

Note that in (3) we ignore the missing data generating mechanism and replace the product over N (all subjects in the *list*) by a product over M (all subjects in the *sample*). We may use only complete data because the missing data are MAR (LITTLE and RUBIN, 1987). Our estimator is not efficient compared with one based on all N observations, but we consider the loss in efficiency less important than the substantial gain in ease and speed of estimation, which is particularly important in analysing large business transaction databases for imputation purposes, in particular when applied in combination with the Bootstrap as detailed below.

The estimation of the factor model is not feasible with standard numerical algorithms for maximizing the likelihood function, given the potentially high-dimensional integration involved in the likelihood. Development of simulated likelihood (SML) estimation has made the approximation of such integrals possible. Such simulation methods were introduced by MCFADDEN (1989), an overview is provided by STERN (1997). In SML based estimation one draws S values z_n^s from $f_X(x_n)$ and uses the simulated log likelihood

$$\tilde{L}\left(\Xi|\breve{Y}\right) = \sum_{n=1}^{M} \ln \sum_{s=1}^{S} \prod_{j=1}^{J} \tilde{f}_Y\left(\breve{y}_{nj}|\eta(z_n^s); \Xi\right) \Big/ S \tag{4}$$

as an approximation to (3). The value of Ξ that maximizes (4) is the SML estimator. It holds that $\tilde{L}\left(\Xi|\widehat{Y}\right) \to L\left(\Xi|\widehat{Y}\right)$ as $S \to \infty$, from the strong law of large numbers. SML provides consistent estimators if $S \to \infty$ as $M \to \infty$, since the simulated likelihood (4) is a consistent simulator of the likelihood (3). The bias is of order $1/S$. However, finite values of S are sufficient to obtain good properties of the estimates (LEE, 1995, WEDEL and KAMAKURA, 2001).

Models with different values for P cannot be compared using likelihood ratio (LR) tests, since the asymptotic Chi-square distribution of the LR test of the P-factor model versus the $P+1$-factor model does not hold (ANDERSON, 1980). We therefore compare models with different values for P on the basis of the BIC information statistic: $BIC = -2\ln\tilde{L}\left(\hat{\Xi}|\,\widehat{Y}\right) + K\{\ln(N)\}$, with K the effective number of parameters (SCHWARTZ, 1978).

2.3 *Bootstrap-based imputation*

In order to augment the transaction database, we draw each missing observation repeatedly from its posterior predictive distribution, given the model estimates, and the values of the observed data for the subject in question. The posterior distribution of the missing data is

$$f_Y\left(\breve{y}_n|\widehat{y}_n\right) = \int f_Y\left(\breve{y}_n|\eta(\hat{x}_n), \hat{\Xi}\right) f_X\left(\hat{x}_n|\widehat{y}_n\right) f_\Xi\left(\hat{\Xi}|\widehat{y}_n\right) \mathrm{d}\hat{x}_n \mathrm{d}\hat{\Xi}. \quad (5)$$

Here, $\hat{x}_n$ is the vector with the posterior estimates of the factor scores for customer n. As explained above, the factor model is not identifiable without further restrictions on the matrix Λ, so that $f_\Xi(\hat{\Xi}|\widehat{y}_n)$ can then not be obtained. Rather than resolving this issue by imposing identification constraints, we use a nonparametric bootstrap procedure (EFRON and TIBSHIRANI, 1994). To that end, we draw b=1,…, B sub samples from the original data with replacement and re-estimate the model for each sub sample, yielding estimates of the parameters $\hat{\Xi}^b$ and posterior estimates of the factor scores $\hat{x}_n^b$. We then draw from $f_X(\hat{x}_n^b|\widehat{y})$, centred around the posterior mean of the factor scores and generate an imputation by drawing from $f_Y\left(y_n|\eta\left(\hat{x}_n^b\right), \hat{\Xi}^b\right)$. We use a single imputation for each Bootstrap sample. Note that in this procedure it is not needed to evaluate the bootstrap estimate of the covariance matrix of the estimates, and it is also not needed to impose identifying constraints on the model. In the application we may use a fairly small number of bootstrap samples, in the range B = 5–10, since such numbers of imputations are generally considered sufficient (LITTLE and RUBIN, 1987).

3 Application

3.1 *Data and study design*

In order to illustrate the proposed approach for list augmentation, we apply and test it on a sample of 5,550 customers of a major commercial bank in Brazil. These data were previously analyzed in part by Kamakura *et al.* (2002). We have data gathered from personal interviews, as well as transaction data from the bank's internal records, for each of these customers. We are interested in investigating the quality of the multiple imputations and therefore, rather than using the entire transaction database, we use the complete data only. Starting from the complete data, we simulate the situation of list augmentation by deleting the survey data for a subset of the subjects. We use the following variables from the bank's internal records (assumed distribution in parenthesis):

- Number of transactions per month (*Poisson*)
- Contribution of the account (*Normal*)
- Volume of deposits in the bank (*Normal*)
- Number of years using the bank (*Normal*)
- Number of visits to the bank in the past 6 months (*Poisson*)
- Age (*rank-order Binomial*)
- Gender (*Bernoulli*)
- Usage indicators for 22 financial services within the bank (*Bernoulli*).

We estimate the proposed mixed data factor analyzer on a random sample of 2,000 customers, for whom we assume that we have full information about their transactions with the bank, as well as external (survey) information on two variables:

- Share of wallet – bank's percentage of the customer's total financial applications, usually known as "share of wallet" (*Normal, after logit transform*)
- Whether the customer would recommend the bank to a friend, as an indicator of "customer satisfaction" *(4-point rank-order binomial)*.

After calibrating the model on the internal and external data for these $N = 2{,}000$ customers, we then apply the estimated model to the remaining customers, for whom we assume the survey data (share of wallet and customer satisfaction) to be missing. We do that for $B = 10$ bootstrap samples drawn randomly with replacement from the calibration sample. Based on the parameter estimates from each of the bootstrap samples, we impute "share of wallet" and customer satisfaction in the holdout sample of 3,550 customers, based solely on their internal records. Since we have the survey data for the holdout customers as well, this allows us to investigate the performance of the imputation 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, it can be applied to augment the entire list. This exercise is particularly important to the bank, because banks usually have good information about their relationships with their own customers, but they have limited information on the relationships these customers might have with competing banks, as well as on their satisfaction. We demonstrate how the bank would be able to use the proposed factor model to identify customers who are "at risk," or with low levels of satisfaction.

3.2 *Results*

Estimation of the factor model from the complete calibration data ($N = 2{,}000$) for $P = 1$ to 4 factors lead us to choose the model with $P = 3$ factors ($P = 2$: BIC = 82472.0, $P = 3$: BIC = 81981.8, $P = 4$: BIC = 82012.0). Parameter estimates of the $P = 3$ model are shown in Table 1. The three factors show a distinct pattern of variable weights. The *first factor* has large weights for the number of transactions and for the satisfaction variable to be imputed (Would you recommend the bank?). The *second factor* has large negative weights for typical banking services such as loans, savings, checking, debit/credit cards, as well as insurance services. This second factor also has a large negative weight for the satisfaction variable, indicating that customer with a high score on this second factor are less likely to recommend the bank to friends. The third factor shows many large weights, in particular for investment and insurance services, but also for traditional banking services such as installment loans, savings, credit card, special checking, private manager and safety

Table 1. Parameter estimates for the complete data ($N = 2{,}000$)*.

Variable	Type	$\hat{\lambda}_0$	$\hat{\sigma}$	$\hat{\lambda}_{j1}$	$\hat{\lambda}_{j2}$	$\hat{\lambda}_{j3}$
Number of transactions per month	Counts	3.58		**0.58**	−0.46	**−0.60**
Contribution of the account	Continuous (standardized)	−0.01	0.59	−0.18	0.08	−0.24
Volume of deposits in the bank	Continuous (standardized)	−0.03	0.51	−0.12	0.04	−0.24
Number of years using this bank	Continuous (standardized)	0.04	0.87	−0.07	−0.16	**−0.50**
Percentage of total applications[1]	Continuous (standardized)	−0.02	0.94	0.05	−0.20	0.28
Number of visits to branch	Counts	0.03		0.00	**−0.72**	−0.01
Would you recommend your main branch[1]	Ordinal	1.08		**−0.95**	**−1.14**	−0.49
Age	Ordinal	0.48		−0.44	−0.01	**−0.53**
Gender	Binary	0.71		−0.02	−0.30	**−0.57**
Savings	Binary	0.32		−0.21	−0.42	**−0.70**
Credit card	Binary	−1.79		0.16	**−0.86**	**−0.85**
ATM card	Binary	1.63		0.31	−0.31	−0.01
Phone banking card	Binary	−1.68		0.06	−0.16	**−0.64**
CD	Binary	−1.51		−0.31	**−0.50**	**−1.78**
Special checking	Binary	0.73		0.32	**−1.05**	**−1.57**
Safety box	Binary	−3.73		−0.20	−0.33	**−1.33**
PC banking	Binary	−3.53		0.06	−0.20	**−0.83**
Auto bill payment	Binary	−2.31		−0.05	**−0.56**	0.32
Personal loans	Binary	−2.93		0.29	**−1.26**	−0.30
Mortgage	Binary	−5.37		0.00	**−1.23**	−0.26
Installment loan	Binary	−5.76		0.05	**−1.86**	**−0.86**
Farming credit	Binary	−6.53		0.10	0.00	**−2.13**
Mutual fund	Binary	−3.22		0.14	−0.45	**−1.35**
Investment fund	Binary	−0.86		0.19	−0.56	**−1.33**
Commodities fund	Binary	−2.25		0.07	−0.03	**−1.56**
Annuities fund	Binary	−3.51		−0.23	−0.37	**−1.56**
Private manager	Binary	−3.94		−0.26	**−0.59**	**−0.97**
Gold	Binary	−3.62		−0.23	**−0.77**	**−1.10**
Car insurance	Binary	−3.26		0.23	**−1.57**	**−1.52**
Home insurance	Binary	−3.63		0.18	**−1.61**	**−1.19**
Life insurance	Binary	−1.34		0.26	**−1.19**	−0.45

*Boldface type indicates large ($>|0.5|$) factor weights.

[1] Survey variable, to be imputed.

box. Customers using a wide range of services offered by this bank are likely to have a negative score on this factor. A negative score on this factor is also related to a large number of transactions per month and long tenure as customer of this bank. This third factor is not strongly associated with the two imputation variables "satisfaction" and "share of wallet" (percentage of total applications).

The pattern of weights of the two imputation variables shows that the satisfaction variable has large weights for the first two factors, while share of wallet is not strongly associated with any of the three factors. This indicates that satisfaction measurements from the survey correlate stronger with the internal transaction records from the banks database, which may leverage in imputations of higher quality as compared with share of wallet. This illustrates the diagnostic value of the mixed outcome factor model for data imputation: inspection of the weights of the variables to be imputed diagnoses the potential quality of the imputations.

We simulate the application of the model for list augmentation, using the bootstrap-based imputation procedure described above. We impute the share of wallet variable and the probability of being dissatisfied (would not recommend the bank to friends) based on the factor model using the eight bootstrap samples (we report the imputed probabilities rather than the 0/1 outcome variable itself, since we found that to be more informative). We focus on the probability of not recommending, because this helps the bank identify customers "at risk" who should be contacted to prevent attrition. Figure 2 shows the distribution of the standardized imputations across the ten bootstrap samples, with the true frequency distribution in the imputation sample. In each of the bootstrap samples, the distribution of the imputed values covers the true frequencies of being satisfied/dissatisfied fairly well for both variables. The figure suggests, however, that the imputations for the share of wallet variable may exhibit too little variability.

For the two imputed variables we compute measures of imputation performance. For share of wallet, the correlations between the true and imputed values range from 0.261 to 0.271, with an average of 0.265. These correlations are somewhat low, which may be partially caused by the variability in the imputations, the measurement error in the "true values" in the imputation sample, and by the fact that the factor model does not show a strong relationship between any of the latent factors and share of wallet. For satisfaction, as a measure of selectivity of the predictive model, we generate power curves of the cumulative proportion of customers who actually said they would not recommend the bank to friends against the cumulative proportion of customers according to the imputed probability for not recommending the bank. These power curves are shown in Figure 3 for the ten different bootstrap-based imputations. A power curve on the 45-degree line indicates a lack of predictive power, while perfect predictive power is obtained if the model sorts all customers perfectly in decreasing order of true likelihood of usage. The graphs are all well above the 45° line, underlining the quality of the multiple imputations in this case. These results indicate that the proposed factor model could be useful in identifying potentially dissatisfied customers, based solely on internal records.

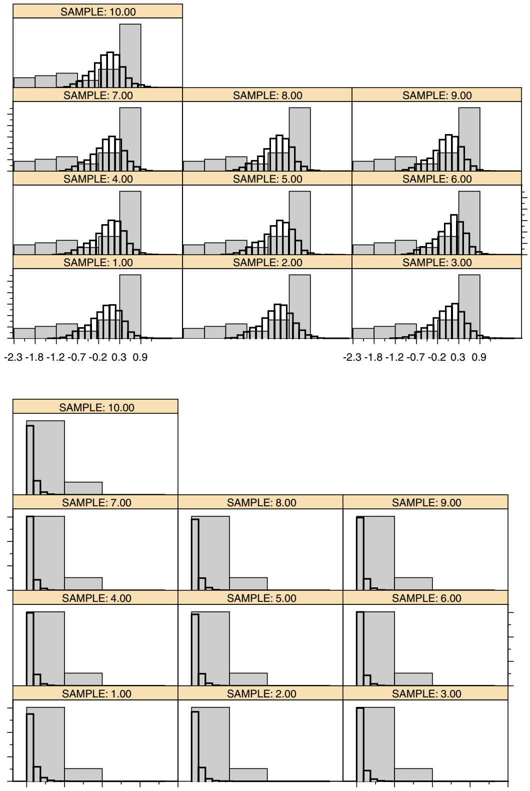

Fig. 2. Distribution of the share of wallet (top panels) and satisfaction (bottom panels), observed values (shaded bars) and imputations (transparent bars) across the $B = 10$ bootstrap samples.

4 Discussion

In database marketing, list augmentation has become increasingly popular. Companies' customer transaction databases are increasingly used for marketing rather than for only accounting purposes. However, the information in the transaction base is often too limited and companies conduct market research among samples of customers to enrich their databases. We propose the use of a mixed outcome factor model for purposes of data augmentation that enables one to impute the survey data, collected among a sample, for all customers in the database.

The proposed procedure offers the advantages of accommodating the different measurement scales of variables usually encountered in transaction databases, of providing a low dimensional representation of the variables, and of enabling one to diagnose the extent to which the model may provide adequate imputations, based on the factor weights of the imputation variables. We have proposed a bootstrap-based imputation approach, which is appealing to use in combination with the factor model. It allows one to include estimation uncertainty in the imputation procedure in a robust manner, without the need to impose identifying constraints on the

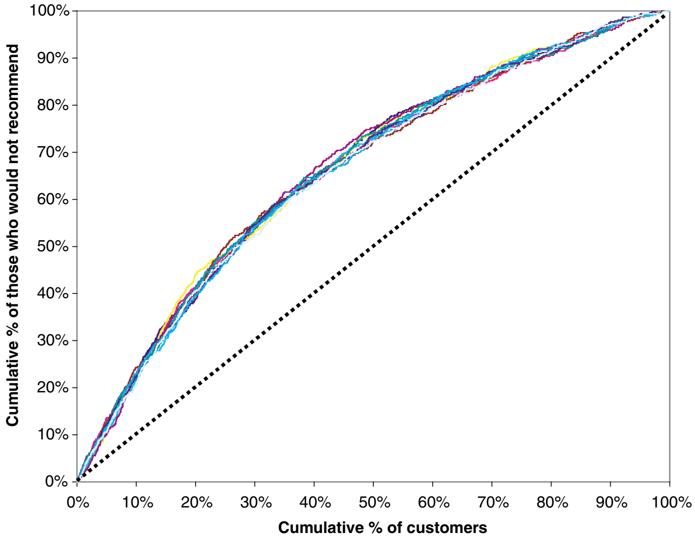

Fig. 3. Power curves for the cumulative proportions of customers who actually would not recommend the bank against that according to the imputed probability for not recommending the bank, for $B = 10$ bootstrap samples.

(primarily) exploratory factor model nor the need to obtain (invert) the information matrix of the parameters. We provide an empirical case study of the approach and we apply it to a transaction data base of a bank. We simulate a list augmentation problem, by deleting part of the data and imputing them with our factor model. This allows us to illustrate the diagnostic value of the factor model for and its performance in list augmentation. However, our investigation of the quality of the imputations in terms of the distribution of the imputed values is somewhat limited. Subsequent research should address the issue of objective validation further by investigating the properties of the resulting inferences based on the imputed data, as in Rubin (1987, Ch. 4). A drawback of the approach is the estimation time involved in the bootstrap procedure including simulated likelihood estimation. However, we believe this problem to further decrease over the years as computation power of desktop computers increases.

References

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

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

Efron, B. and R. J. Tibshirani (1994), *An introduction to the bootstrap*, Chapman and Hall, London.

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

KAMAKURA, W. A. and M. WEDEL (2000), Factor analysis and missing data, *Journal of Marketing Research* **37**, 490–498.

KAMAKURA, W. A., F. DE ROSA, F. M. WEDEL and J. A. MAZZON (2002), Cross-selling financial services with database marketing, *International Journal of Research in Marketing* **19**, forthcoming.

LEE, L. F. (1995), Asymptotic bias in simulated maximum likelihood estimation of discrete choice models, *Econometric Theory* **15**, 437–483.

LITTLE, R. J. A. and D. B. RUBIN (1987), *Statistical analysis with missing data*, Wiley, New York.

MCCULLAGH, P. and J. A. NELDER (1989), *Generalized linear models*, Chapman and Hall, New York.

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

MOUSTAKI, I. and M. KNOTT (2000), Generalised latent trait models, *Psychometrika* **65**, 391–411.

RUBIN, D. B. (1987), *Mulple imputation of nonresponse in surveys*, Wiley, New York.

SCHAFER, J. L. (1997). *Analysis of incomplete multivariate data*, Chapman and Hall, London.

SCHWARTZ, G. (1978), Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

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

WEDEL, M. and W. A. KAMAKURA (2001), Factor analysis with mixed observed and latent variables in the exponential family, *Psychometrika* **66**, 515–530.

Received: March 2002. Revised: August 2002.