

**Beating the Average with Conditional Averages:  
Target Selection using Geo-Demographic Joint Distributions\***

**Jason A. Duan**

Yale School of Management  
jun.duan@yale.edu  
135 Prospect St, PO Box 208200  
New Haven, CT 06520-8200  
(203)-436-8119  
(203)-432-3003

**Sachin Sancheti**

Yale School of Management  
sachin.sancheti@yale.edu  
135 Prospect St, PO Box 208200  
New Haven, CT 06520-8200  
(203)-508-2872  
(203)-432-3003

**K. Sudhir**

Yale School of Management  
k.sudhir@yale.edu  
135 Prospect St, PO Box 208200  
New Haven, CT 06520-8200  
(203)-432-3289  
(203)-432-3003

July 2007

---

\* The authors contributed equally to the paper and are listed in alphabetical order. We thank the participants at the Yale Ph.D. Student Workshop and the Marketing Science Conference in Singapore for their comments.

## **Beating the Average with Conditional Averages: Target Selection using Geo-Demographic Joint Distributions**

### **Abstract**

Managers often face the problem of limited data at the individual customer level. A common practice is to augment the limited available customer level data with averages for the group to which the customer belongs. We demonstrate using a target selection problem that this standard practice of using group (zip code) averages as a proxy for individual information leads to biased inference and erroneous managerial decisions. We therefore propose that firms use “conditional averages,” i.e., rather than use the raw averages for the group, use averages conditional on the available individual information in the firm’s internal databases. However, this is hard to implement in practice because group level joint distributions are unavailable. We develop a flexible and scalable approach to obtain group level joint distributions by augmenting the available group level marginal distributions with joint distribution information from a representative sample of individuals at the aggregate market level that comprises all the groups which form the aggregate market. Our approach to infer joint distributions has a wide range of applications in marketing and empirical industrial organization.

Key Words: Target Selection, Geo-demographics, Database Marketing, Bayesian Estimation, Missing Data Problems

## 1. Introduction

The problem of target selection – identifying potential customers who are likely to be most valuable to a firm or most responsive to a marketing campaign – is a fundamental problem in direct marketing. The standard approach involves relating the value of the firm’s existing customers to their descriptive characteristics using historic data and then selecting new individuals who are likely to be most valuable, based on their descriptive characteristics.

While logically simple, the strategy is often hard to implement in practice because the necessary data is not easily available. With transactional data, firms know the value of their current customers, but have limited descriptive information about them. For example, a bank’s internal database has detailed account data to measure the customer’s profitability, but limited descriptive demographic data. Typically, a bank has the customer’s mailing address and age (obtained from the application for new accounts), but not other information such as income, home value and education level of the customer.

Firms therefore use the services of data intermediaries such as Claritas or Experian to augment their internal databases with descriptive information about customers. But the intermediaries cannot provide data at the individual level, because the FTC mandates that they “mask” individual customer information by reporting it only at a geographically aggregate level (e.g., five or nine digit zip code) for privacy reasons. For the same reason, the Census Bureau also reports their data only at the aggregate census block group level. Firms therefore need to work with aggregate level “masked data” (e.g., *average* demographics of the relevant customer zip code).

Steenburgh et al. (2003) demonstrated that this standard practice of using the average demographics of the customer’s zip code as a surrogate for individual’s specific information

without accounting for unobserved zip code effects exaggerates the precision of the parameter estimates. This is because the practice implicitly assumes that all the variation across zip codes is due to the observed average zip code demographics. Steenburgh et al. (2003) add zip code level random effects that take into account the unobserved variation across zip codes and increase the standard error of the estimated parameters.

In this paper, we introduce the idea that using the standard approach not only leads to spurious precision, but worse, it leads to bias in estimates. The intuition for the bias is as follows. Suppose the most valuable customers tend to be higher income, older customers within any zip code and older customers tend to have higher incomes. Customer age enters the model at the individual level because it is in the internal database, but income enters as the average value for the zip code. Given the positive correlation between age and income, older customers will also have above average incomes for the zip code and younger customers will have below average incomes for the zip code. Ignoring this correlation and using the average zip code income instead of the correct conditional (on age) average causes the residuals to be systematically correlated with demographic variables, thus biasing the estimates. While the Steenburgh et al. (2003) procedure corrects the precision problem with masked data, it does not address the bias problem.<sup>1</sup>

This paper proposes and illustrates a practical procedure to correct the bias (in addition to the precision correction) inherent in the use of masked data. From the intuition above, the key idea is that we should not plug-in the average value of the zip code income, but the conditional average of income given the age that is already available in the firm's internal database. To obtain the conditional average, we need the joint distribution of age and income for each zip

---

<sup>1</sup> Steenburgh et al.'s (2003) university's admissions targeting application does not have the bias problem because their individual specific variables are reasonably uncorrelated with their masked demographic variables.

code. Thus our solution is to “beat the average” with “conditional average” obtained from “geo-demographic joint distributions.” The challenge however is that joint distributions are not available directly. The key methodological contribution of the paper is that we develop a practically feasible technique to infer joint distributions using data that are easily available to firms.

Putler et al. (1996) first addressed the issue of obtaining zip code joint distributions.<sup>2</sup> They use a Bayesian approach to estimate cell probabilities of a contingency table comprised of multiple demographic variables by combining information from the marginal distributions at the zip code level with joint distribution information from a sample of individuals from an aggregate level market (e.g., county, state, MSA etc.). Specifically, they treat the individual sample data at the aggregate level as a prior for the joint distribution and update it for each zip code using the zip code marginal distribution. This approach however is not easily scalable to applications involving many variables with multiple levels for each variable. Consider four variables with 5 levels each. The total number of cells in a contingency table made of these variables would have  $5*5*5*5 = 625$  cells. The large number of parameters to be estimated makes this approach impractical for many practical problems.

We address the dimensionality problem by avoiding the contingency table approach and directly working in a continuous variable framework. Even when variables in geo-demographic datasets are collected and reported in ordinal levels (e.g. income, age), we can use the continuous variable framework by assuming that the observed ordinal data are generated from an underlying latent variable. We then illustrate how to use data augmentation to apply our framework. In addition, if there are categorical variables that cannot be represented as latent continuous

---

<sup>2</sup> For simplicity, we will refer to the local geographic unit as a zip code in the rest of the paper. In practice, the geographic unit may be smaller like a Census Block Group (CBG). It could also be larger, where geographic units may be cities or counties, while the aggregate market of interest could be a state, region or the entire country.

variables (e.g., race, gender), we can combine our approach with the contingency table framework. But a large number of categorical variables will create dimensionality challenges just like Putler et al. (1996). In sum, our approach can deal with many continuous and ordinal variables, and a limited number of categorical variables. This makes our method more practical than extant methods available for a wider range of applications.

Romeo (2005) also proposes a solution to alleviate the dimensionality problem. It is a method of moments based parametric approach where cell probabilities are parametric functions of observed data at the local market and higher level. With his parametric representation, the number of parameters to be estimated does not increase exponentially with the number of cells in the contingency table. However, the identifying assumption used in Romeo (2005) is restrictive in that he equates the covariance in each zipcode to the covariance in the aggregate individual-level sample. Since he assumes this for all the zip codes, this would imply that covariances across zip codes are also the same (and equal to aggregate market covariances), an assumption that is inconsistent with the data, given the wide differences in variances across zip codes.<sup>3</sup> Instead of assuming equal covariances, we make the less restrictive assumption that only correlations between variables are the same across zip-codes. As we describe in detail later, this is the least restrictive assumption we can make while achieving identification with the available data.

In summary, our approach improves on the state of the art in terms of inferring joint distributions in several ways. It is practical in that it is scalable to a large number of variables and can be used with both continuous and ordinal variables. The data requirements are relatively

---

<sup>3</sup> To be precise, Romeo (2005) sets up moment conditions to minimize the difference between zip code and aggregate market covariances. So his model, in principle, does account for different covariances across zip codes, but only due to sampling error. We, on the other hand, are flexible with the covariances and do not impose the condition that they be same across zip codes.

simple given the additional individual level sample is typically available through public datasets or can be collected relatively easily by any firm directly. Finally, our approach optimally combines the information in both the zip code marginal distributions and the aggregate sample, where the joint distribution information uses the information in both types of data in the optimal manner. Rather than treat the aggregate sample distribution as simply a prior (Putler et al. 1996) or equate the covariance of the aggregate sample to the covariance at each zip code (Romeo 2005), we treat the aggregate sample as arising from a distribution that is a (zip code population) weighted sum of all of the joint distributions of the zip codes that constitute the aggregate market. This recognizes that the joint distribution of the zip code marginal distributions also contains information about the joint distribution, which we use in our inference.

We organize the rest of the paper as follows. Section 2 explains the “bias problem” with masked data and how the problem can be solved by using conditional averages. Section 3 describes our procedure to estimate joint distributions which are needed to obtain conditional averages. Section 4 reports the results of a simulation analysis validating the procedure. Section 5 provides an empirical illustration in the context of a bank’s direct mail customer acquisition program. Section 6 concludes.

## **2. The “Bias” Problem using Masked Data and Its Correction Procedure**

As discussed in the introduction, the standard approach used by practitioners for target selection is to append the individual level variables ( $X$ ) with masked variables available at a group (e.g., zip code, census tract) level ( $Z$ ) to explain a variable of interest (e.g., customer profit) below:

$$Y_i = X_i^T \alpha + Z_{j(i)}^T \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \dots\dots (1)$$

where  $i$  indexes individuals and  $j(i)$  indexes the zip code to which  $i$  belongs.  $X_i$  represents the set of individual characteristics available at the individual level and  $Z_{j(i)}$  represents the individual characteristics which are “masked”, i.e., they are only available as summary statistics (e.g., group average, standard deviation) at the group level.

The model described above ascribes all variation across groups to the group level averages. However, there could be other factors varying across groups that are unobserved. To account for the unobserved omitted group effects Steenburgh et al. (2003) propose also including group level random effects as in the following equation:<sup>4</sup>

$$Y_i = X_i^T \alpha + Z_{j(i)}^T \beta + v_{j(i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad \dots\dots (2)$$

$v_{j(i)}$  is the random effect for the group dummies.

Steenburgh et al. (2003) compare model (2) with the standard model (1) and show that the standard errors for  $\beta$  are much lower in the standard model. The addition of random effects removes the spurious confidence in the estimates of  $\beta$ . Thus they correct the “precision” problem in the target selection equation. However, we show below that there is a “bias” problem that still remains and cannot be eliminated through the introduction of random effects  $v_{j(i)}$ .

To understand why the bias occurs, consider the following: Suppose we could observe the unmasked individual level values (denoted by  $Z_i$ ) for the “masked” variables  $Z_{j(i)}$ , then we could estimate the following model:

$$Y_i = X_i^T \alpha + Z_i^T \beta + v_j + e_i, \quad \dots\dots(3)$$

---

<sup>4</sup> Specifically, Steenburgh et al. (2003) estimate the equivalent hierarchical model:

$$Y_i = X_i^T \alpha + \gamma_{j(i)} + \varepsilon_i \text{ where}$$

$$\gamma_{j(i)} = Z_{j(i)}^T \beta + v_{j(i)}$$

The addition of hierarchy per se does not make any difference.

where  $e_i$  is the random error which is uncorrelated with the observed characteristics  $X_i$  and  $Z_i$ .

One can then decompose the  $Z_i$  into a group-level average and an individual-specific deviation from the average.

$$Y_i = X_i^T \alpha + Z_{j(i)}^T \beta + \nu_j + (Z_i - Z_{j(i)})^T \beta + e_i.$$

By letting  $\varepsilon_i = (Z_i - Z_{j(i)})^T \beta + e_i$  be the random error, we can reduce the above model to (2), when  $Z_i$  is not observed and only the masked group level values are available. It is easy to see that using the group average  $Z_{j(i)}$  instead of individual characteristics  $Z_i$  in the model causes the individual-specific deviation being absorbed into the random error  $\varepsilon_i$ . If the observed demographic variables  $X_i$  vary systematically with the unobserved characteristics  $Z_i$ , then  $X_i$  will be correlated with  $\varepsilon_i$ , making the least square estimate of  $\alpha$  inconsistent. Only when  $X_i$  and  $Z_i$  are independent of each other (as happens to be the case in the Steenburgh et al. application), will the model (2) parameters be consistently estimated. However, in most applications, correlation between  $X_i$  and  $Z_i$  is very likely. For example, in our illustrative example, where we model profitability for a commercial bank to target potential zip-codes, demographic variables such as *Age*, *Income* and *Home Value* are all correlated, but *Income* and *Home Value* is available to the bank only as the zip-code average. Indeed, in our application, we find that *Income* and *Home Value* are correlated with *Age*.

Suppose we know or can infer the joint distribution of the observed and “masked” variables for each group. Once we know the joint distribution of  $X_i$  and  $Z_i$  for each group  $j$ , we propose replacing the unobserved  $Z_i$  with the conditional mean of  $Z_i$  given  $X_i$  to obtain

consistent estimates. Let  $E_j(Z_i | X_i)$  be the conditional mean obtained from the joint distribution for group  $j$ . We model

$$Y_i = X_i^T \alpha + E_j(Z_i | X_i)^T \beta + \nu_j + \tilde{\varepsilon}_i \quad \dots\dots(4)$$

where  $\tilde{\varepsilon}_i = (Z_i - E_j(Z_i | X_i))^T \beta + e_i$  according to the ideal model (3).

We now show that Model (4) will provide us consistent estimates of  $\alpha$ . This is because as shown below,  $\tilde{\varepsilon}_i$  will have zero mean given  $X_i$ , and therefore is uncorrelated with all the covariates in the model. Indeed,

$$E(\tilde{\varepsilon}_i | X_i) = E_j\{(Z_i - E_j(Z_i | X_i))^T \beta + e_i | X_i\} = E_j\{Z_i - E_j(Z_i | X_i) | X_i\}^T \beta + E_j(e_i | X_i) = 0$$

and because  $E_j(Z_i | X_i)$  is a function of  $X_i$ ,  $\tilde{\varepsilon}_i$  is uncorrelated with  $E_j(Z_i | X_i)$ .

Model (4) nests model (2) inasmuch as model (4) reduces to model (2) when  $Z_i$  and  $X_i$  are uncorrelated, i.e.,  $E(Z_i | X_i) = Z_{j(i)}$  in this case. We discuss this further in Section 4 with estimation results using simulated data.

### 3. Estimating Joint Distributions

In this section, we develop the procedure to estimate a joint distribution of variables using the marginal distributions of group characteristics and a sample of individuals from the aggregate market comprising of all groups. The individual sample data from the aggregate population provides information on correlations between characteristic variables which is missing in the group level marginal distributions. Nevertheless, we note that the correlations in the aggregate data cannot be directly used as correlations at the group level. This is because the joint distribution of the aggregate population is a mixture of the group level joint distributions

weighted by the population of each group. We first present our method for continuous variables and then discuss how to extend the approach for ordinal categorical variables.

### 3.1 Continuous Variables

Suppose we have  $J$  groups:  $j = 1, \dots, J$ . For each group, we know the group population  $n_j$  and the marginal distributions of characteristics  $X_{j1}, \dots, X_{jK}$  where  $K$  is the total number of variables. The joint distribution of  $X_{j1}, \dots, X_{jK}$  is assumed to be multivariate normal, i.e.,  $[X_{j1}, \dots, X_{jK}] \sim N_K(\mu_j, \Sigma_j)$  where  $\mu_j = [\mu_{j1}, \dots, \mu_{jK}]$  is the mean vector that is known at the group level.  $\Sigma_j$  is a  $K \times K$  covariance matrix that is not completely known.  $\Sigma_j$  can be decomposed into:  $\Sigma_j = D_j R D_j$  where  $D_j$  is the diagonal standard deviation matrix and  $R$  is the correlation matrix. Let each diagonal entry of  $D_j$  be  $\sigma_{jk}$ . We assume that the correlation  $R$  is same across all groups. As discussed in the introduction, this relaxes the assumption of equal covariance across all groups in Romeo (2005), which is typically rejected by the data. Note that the marginal distributions at the group level have no information about the correlation  $R$ , therefore the correlation matrix is to be identified completely from the aggregate survey sample. In the absence of a survey sample from each group, it is impossible to identify a heterogeneous group level correlation; hence our assumption of equal correlation across groups is the only practical assumption given the available data.

Suppose we collect a sample of  $n$  individuals from the aggregate population. For each individual  $i$ , let  $Z_{i\bullet} = [Z_{i1}, \dots, Z_{iK}]$  be the corresponding characteristics. As this sample is drawn from the aggregate population and there is no group indicator in the data, this sample is from a finite mixture distribution:

$$[Z_{i1}, \dots, Z_{iK}] \stackrel{iid}{\sim} \sum_{j=1}^J \omega_j N(\mu_j, \Sigma_j), \quad i = 1, \dots, n, \quad \dots (5)$$

where the mixture weights are given by  $\omega_j = \frac{n_j}{\sum_{j=1}^J n_j}$ . The first two moments of this finite

mixture distribution can be easily obtained from (5). The mean vector of  $Z_{i\bullet}$  is given by

$$E(Z_{i\bullet}) = \sum_{j=1}^J \omega_j \mu_j \quad \text{and the covariance matrix of } Z_{i\bullet} \text{ is}$$

$$\sum_{j=1}^J \omega_j (D_j R D_j + \mu_j \mu_j') - \left( \sum_{j=1}^J \omega_j \mu_j \right) \left( \sum_{j=1}^J \omega_j \mu_j' \right). \quad \dots (6)$$

From (6), the variance of each  $Z_{ik}$  is given by:

$$\text{var}(Z_{ik}) = \sum_{j=1}^J \omega_j \sigma_{jk}^2 + \sum_{j=1}^J \omega_j \mu_{jk}^2 - \left( \sum_{j=1}^J \omega_j \mu_{jk} \right)^2$$

and the covariance between  $Z_{ik}$  and  $Z_{im}$  is

$$\text{cov}(Z_{ik}, Z_{im}) = \sum_{j=1}^J \omega_j (\sigma_{jk} \sigma_{jm} R_{km} + \mu_{jk} \mu_{jm}) - \left( \sum_{j=1}^J \omega_j \mu_{jk} \right) \left( \sum_{j=1}^J \omega_j \mu_{jm} \right) \quad \dots (7)$$

Equation (7) describes the relationship between covariance in the aggregate sample and correlation at the group level. It can be seen that the correlation in the aggregate sample will not generally be equal to the correlation at the group level and therefore cannot be used as an estimate of  $R$  directly. The correlation in the aggregate sample,  $\text{cov}(Z_{ik}, Z_{im}) / \sqrt{\text{var}(Z_{ik}) \text{var}(Z_{im})}$ , comprises not only of within-group correlation ( $R$ ) but group means and variances also well. The variation in the standard deviations and means across groups contribute to the observed correlations of the survey sample. Previous approaches (Putler et al., 1996 and Romeo, 2005) do not take this into account when constructing zip code joint distributions.

Since we already know the group means  $\mu_j$  and variances  $D_j$ , we only need to estimate the within-group correlation. We can construct a likelihood function for the unknown  $R$  using the finite mixture distributions (5) for  $Z_{i\bullet}$ ,  $i = 1, \dots, n$

$$\prod_{i=1}^n \left( \sum_{j=1}^J \omega_j N(Z_{i\bullet} | (\mu_j, D_j R D_j)) \right) \dots\dots (7)$$

We propose a Metropolis-Hastings algorithm to estimate the unknown matrix  $R$ . The details of the algorithm are presented in Appendix A.

Although we assume that the variables are jointly normal, this parametric assumption is less restrictive than it appears because appropriate transformations can make a variable approximately normal. For example, skewed data can be modeled using a lognormal distribution and this can be easily transformed into a normal distribution; discrete ordinal variables can be “transformed” into continuous normal variables using a latent variable approach. We discuss how to model ordinal variables next.

### 3.2 Discrete Ordinal Variables

In many applications, data is available as discrete ordinal variables, i.e. a frequency table of how many individual  $X_{jk}$ 's fall within a range of the k-th variable in zip code  $j$ . See Table 1 for an example of the type of data that is available in the ordinal form from either the census bureau or from market research firms.

**Table 1. Ordinal variables usually reported in zip code demographic data**

Zip Code	Income			Home Value		
	<\$50k	\$50k-\$100k	>\$100k	<\$100k	\$100k-\$250k	>\$250k
06520	45%	35%	20%	20%	60%	20%
06510	25%	65%	10%	15%	70%	15%

Individual characteristics  $X_{jk}$ 's, which are not directly available, can be treated as latent variables that generate the frequency table with fixed cut-off values. The corresponding mean  $\mu_j$  and standard deviation  $D_j$  of  $X_{jk}$ 's can be inferred from a sample of the latent variable conditional on the observed frequency distributions. The data augmentation technique (Tanner and Wong, 1987) provides a convenient tool to sample the latent variables  $X_{jk}$ 's given the frequency distribution of the ordinal categorical variable in the table. This method greatly reduces the dimensionality of the correlation-estimation problem, thanks to the simple correlation structure of the multivariate normal latent variables. This helps us avoid the dimensionality problems faced by previous researchers.

For any variable  $k$  corresponding to zip code  $j$ , assume there are  $M$  categories:  $\{C_k^1, \dots, C_k^M\}$ . We observe the number of individuals  $n_{jk}^m$  (or the proportions) in category  $C_k^m$ . Given  $\mu_j, D_j$  and  $R$ , the latent variables  $X_{jk}$  are sampled as in the multivariate probit model in Chib and Greenberg (1998). To be self contained, we describe the method concisely for our problem setting.

The probability of the  $k$ -th characteristic of individual  $l$  being in category  $C_k^m$  is  $P(\gamma_k^{m-1} \leq X_{jkl} \leq \gamma_k^m)$  where  $\gamma_k^m$ 's are the cut-off values for the categories of variable  $k$  and  $l = 1, \dots, n_j$ . These cut-offs can either be given a priori for numeric variables or be estimated for other ordinal variables (see for e.g., Albert and Chib, 1993). The probability that all characteristics belong to the joint category  $(C_1^{m_1}, \dots, C_K^{m_K})$  is then

$$P\left(\bigcap_{k=1}^K (\gamma_k^{m_k-1} \leq X_{jkl} \leq \gamma_k^{m_k})\right)$$

Conditioning on  $\mu_j$  and  $\Sigma_j = D_j R D_j$ , the  $X_{j\bullet}$  is sampled as

$$\left[ X_{j1l}, \dots, X_{jkl} \right] \sim N_K \left( \mu_j, D_j R D_j \right) I \left( \bigcap_{k=1}^K \left( \gamma_k^{m_k-1} \leq X_{jkl} \leq \gamma_k^{m_k} \right) \right) \quad \dots (8)$$

Just as with continuous variables, zip code level data would not provide information on the correlation matrix. So we ought not to use the samples of  $X_{jkl}$  from (8) to estimate  $R$ . However, we use the sample of  $X_{jkl}$  to estimate mean and variance for the zip code  $j$ , which will be used in (7) for the estimation of  $R$ . Indeed, we may practically sample  $X_{jkl}$  independently from the following truncated univariate normal distribution

$$\left[ X_{jkl} \mid \mu_{jk}, \sigma_{jk} \right] \sim N \left( \mu_{jk}, \sigma_{jk} \right) I \left( \gamma_k^{m_k-1} \leq X_{jkl} \leq \gamma_k^{m_k} \right),$$

which provides the same information on  $\mu_{jk}$  and  $\sigma_{jk}$ .

Conditioning on sample  $X_{jkl}$ , the posterior distribution for  $\mu_j$  and  $D_j$  is proportional to

$$\prod_{k=1}^K \prod_{l=1}^{n_j} N \left( X_{jkl} \mid \left( \mu_j, D_j \right) \right) \pi \left( \mu_j \right) \pi \left( D_j \right).$$

Once  $\mu_j$  and  $D_j$  are sampled, we can impute them into the estimation algorithm described in Section 3.1 to sample the correlation  $R$ .

We propose a Metropolis-embedded Gibbs sampler for group means, standard deviations and correlations. We elaborate on the steps of the sampler in Appendix A.

#### 4. Validating the Procedures – Simulation

We now report the results of a simulation study that validates the procedures. The study has two objectives: (1) to show that our procedure for inferring joint distributions in Section 3.1 “works” in that it can recover the underlying correlation of individual characteristic variables at zip-code level and (2) to show that the use of conditional averages in the place of averages for the masked data helps us obtain consistent estimates for the targeting equation.

### 4.1 Simulated Example for Recovering Joint Distributions

We first select four variables (named as  $X_1$ ,  $X_2$ ,  $Z_1$  and  $Z_2$ ) and their correlations, which are shown in the first row in each cell of Table 2. The four variables are meant to approximately proxy variables such as Age, Income, Years of Education, and Home Value respectively. We simulate 100 zip codes and corresponding means and standard deviations of the four selected variables for each of the zip codes. The means of the four variables across zip codes are sampled from normal distributions with means and variances as indicated:  $N(\log(40), 0.01)$ ,  $N(\log(40000), 0.09)$ ,  $N(\log(15), 0.01)$  and  $N(\log(15000), 0.09)$ . The standard deviations across zip codes are sampled from the following log-normal distributions:  $\log N(0.4, 0.0025)$ ,  $\log N(1, 0.0025)$ ,  $\log N(0.3, 0.0025)$  and  $\log N(1, 0.0025)$ . We then obtain joint distributions using the sampled means and standard deviations, and a common correlation matrix given in Table 2.

Table 2. Correlations in simulated data

	$X_2$	$Z_1$	$Z_2$
	<b>0.1</b>	<b>0.3</b>	<b>0.2</b>
$X_1$	0.058 <i>0.11 [0.089, 0.14]</i>	0.21 <i>0.30 [0.26, 0.33]</i>	0.16 <i>0.21 [0.18, 0.25]</i>
		<b>0.3</b>	<b>0.6</b>
$X_2$		0.22 <i>0.32 [0.28, 0.36]</i>	0.53 <i>0.59 [0.52, 0.64]</i>
			<b>0.3</b>
$Z_1$			0.25 <i>0.30 [0.25, 0.34]</i>

Note: The top row in each cell represents the true value of correlation. The middle row is the correlation obtained from the aggregate market sample. The bottom row is the correlation obtained using our approach, with the 95% posterior interval.

The zip code level distributions constructed above represent continuous variables. In order to simulate the more common scenario of ordinal variables, we sample individuals (1000 - 1500) in every zip code using the joint distributions. We then transform individual data into an ordinal categorical table like Table 1 using ten cutoff values and save only the marginal cell-counts for all four variables. We also randomly select a sample of 2000 individuals (sample without replacement) from all zip codes and save their complete characteristics. This random sample is treated as the aggregate market sample from which the correlations are inferred.

As we state in Section 3.1, it is inappropriate to use the aggregate market sample correlation in place of  $R$ . We compute the sample correlations for the aggregate market sample and present them in the second row in each cell of Table 2. It is obvious that the sample correlations are very poor estimators as we had argued in Section 3.1. For this particular example, a downward bias is observed but in other cases it could be upward depending on the variation of means and variances across zip codes. We then apply our estimation procedure to estimate  $R$  and present the posterior means and 95% posterior predictive intervals for all correlation parameters in the third row in every cell of Table 2. Contrasting the correlations of the aggregate market sample, we can see that our estimates are very accurate and all the predictive intervals cover the real correlations. The simulation study validates that our model recovers the correlations very well and thus effectively the joint distributions for each zip code.

## **4.2 Simulated Example for Targeting**

In this section, we demonstrate the performance of our modeling approach for targeting using simulated profitability data. We still use the 100 zip codes from the simulation in Section 4.1. For each of these zip codes, we randomly select 50 individuals and use the model with

detailed individual level data as in equation (3) to obtain their profitability, conditional on known parameters. We assume that variables  $X_{ij1}$  and  $X_{ij2}$  are observed at the individual level, while  $Z_{ij1}$  and  $Z_{ij2}$  are “masked” and therefore we only know their means and standard deviations at the zip code level. For equation (4), the individual conditional means ( $\tilde{Z}_{ij1}$  and  $\tilde{Z}_{ij2}$ ) are obtained using the joint distribution (correlation matrix) recovered in Section 4.1.<sup>5</sup>

To assess the effectiveness of our approach we compare the results from our proposed model against three benchmark models. The characteristics of the benchmark models and the proposed models are summarized in Table 3.

**Table 3. Features of the Alternative Models**

	Standard Approach (Equation 1)	Steenburgh et al. (2003) (Equation 2)	Equal Covariance (a la Romeo 2005)	Equal Correlation (Proposed Model) (Equation 4)
Random Effect	No	Yes	Yes	Yes
Replace Missing Individual data with	Group Average	Group Average	Conditional Average	Conditional Average
Assumption to obtain Conditional Average	NA	NA	Aggregate covariance equal to zip code covariance	Correlation equal across zip codes

The first benchmark model uses the ‘naïve’ or standard approach wherein zip code averages are treated as masked data and these averages capture all the variation across zip codes. This is also the benchmark model used by Steenburgh et al. (2003). The second benchmark model we estimate is the one proposed by Steenburgh et al. (2003). This model adds an unobserved component of variation across zip codes to the previous model. The third benchmark model uses conditional averages for masked data, but similar to Romeo (2005) assumes that the covariance for each zip code is equal to the covariance in the aggregate market sample. Finally, our proposed model also uses conditional averages for masked data, but assumes that only the

<sup>5</sup> We do not have point estimates for the means, standard deviations and correlations for zip codes, but samples from the posterior distributions of the same. For the purposes of this simulation and the application, we take means of all simulated samples to obtain point estimates and use those to construct joint distributions of all zip codes.

correlation across variables is equal across zip codes. When the observed and masked variables are independent, the conditional averages reduce to simple averages and the last two models reduce to the Steenburgh et al. (2003) model.

The estimation procedure for our proposed model (4) is outlined in Appendix B. The algorithms to estimate the other three models are derivatives of the given algorithm and hence are not provided. We used 5000 iterations for inference and discard the first 1000 as burn-in even though we found good convergence after about 100 iterations. We report the posterior means as parameter estimates, posterior 95% intervals (in parentheses) and log-marginal likelihood in Table 4 for each of the proposed models. The true values are shown in the first column.

We can observe obvious biases in the first two models in the coefficient for  $X_1$ . The coefficient for  $X_2$  is also biased upward for Steenburgh et al. (2003) model. The cause of these biases is that the omission of significant predictive variables  $Z_{ij}$  that introduces correlation between the random error  $\varepsilon$  and the observed variable  $X_{ij}$ . Merely including the random effects  $v_j$  cannot compensate for these omitted effects.

In contrast, the means of the posterior samples from our model (4) are very close to the true parameters and the 95% posterior predictive intervals contain all of the true parameters. This indicates our model nearly recovers the data generating scheme. The log-marginal likelihood, a measure of the fit of the model, also increases from left to right indicating that addition of the random effect alone, as in Steenburgh et al. (2003), does not solve the problem completely. Although the third model with equal zip code and aggregate covariance has better fit than Steenburgh et al. (2003), it does worse than our model with constant correlation assumption, both in terms of fit and recovering the true parameters.

**Table 4. Comparison of results of the four models**

Simulation Results					
	TRUE	Standard Approach	Steenburgh, et al. (2003)	Equal Covariance a la Romeo (2005)	Equal Correlation (Proposed Model)
<i>Intercept</i>	-10	<b>-10.004**</b> [-10.474, -9.527]	<b>-5.894**</b> [-6.293, -5.5]	<b>-9.963**</b> [-11.340, -8.738]	<b>-9.790**</b> [-10.534, -9.062]
<i>X1</i>	1.5	<b>1.658**</b> [1.582, 1.733]	<b>1.685**</b> [1.592, 1.779]	<b>1.450**</b> [1.336, 1.562]	<b>1.450**</b> [1.351, 1.548]
<i>X2</i>	0.1	<b>0.103**</b> [0.097, 0.109]	<b>0.150**</b> [0.130, 0.169]	<b>0.158**</b> [0.139, 0.177]	<b>0.104**</b> [0.091, 0.117]
<i>Z1</i>	1.5	<b>1.293**</b> [1.155, 1.428]	<b>1.518**</b> [1.051, 2.028]	<b>1.340**</b> [0.845, 1.868]	<b>1.480**</b> [1.183, 1.778]
<i>Z2</i>	0.1	<b>0.096**</b> [0.092, 0.100]	<b>0.098**</b> [0.084, 0.111]	<b>0.096**</b> [0.082, 0.111]	<b>0.098**</b> [0.088, 0.108]
$\sigma_\varepsilon^2$	0.01	<b>0.022**</b> [0.021, 0.023]	<b>0.019**</b> [0.018, 0.021]	<b>0.019**</b> [0.018, 0.021]	<b>0.019**</b> [0.018, 0.020]
$\sigma_v^2$	0.0025		<b>0.005**</b> [0.003, 0.009]	<b>0.007**</b> [0.004, 0.011]	<b>0.003**</b> [0.002, 0.004]
<i>Log-Marg. Likelihood</i>		<b>1213.7</b>	<b>1358.8</b>	<b>1361.2</b>	<b>1380.4</b>

Note: (1) The intercept in Steenburgh et al. (2003) model is much lower because variables Z1 and Z2 are normalized as deviations from their means. This is done for identification purposes in the hierarchical model.

(2) \* p < 0.1; \*\* p < 0.05

We also estimated models with simulated data when there is zero correlation between the X and Z variables. As expected, in these models, since our proposed model reduces to the Steenburgh et al. (2003) model, we recover virtually identical estimates that are close to the true values. Overall, the simulation analysis confirms that our procedure is able to recover the true parameters.

## 5. Empirical Illustration

Our empirical illustration addresses a target selection problem for a bank in the Northeastern United States. The bank seeks to understand the demographic determinants of customer profitability for its existing customers in order to identify which zip codes have the highest profitability prospects.

### 5.1 Data

While the bank's internal database has accurate measures of customer transactions, it has limited information about the customer characteristics at the individual level. Typically, customer profitability, age and zip code of residence are the only information available for each customer (i.e.  $Y$  = customer profitability;  $X$  = age). Customer profitability is a measure of the total revenues generated by the customer net of the costs associated with serving the customer. Banks calculate this figure at the individual level to gauge the value of a customer. The other two variables, Age and Zip code of residence, are naturally available to the bank because they are reported to banks at the start of a relationship. Our customer level data contains information on profitability, age and zip code of residence of 1655 customers residing in 40 different zip codes in the state of Connecticut.

For our illustration, we will use two additional variables that are relevant to the bank for targeting: income and home value. However, since these variables are not available at the individual level, we instead use their conditional averages corresponding to the zip code of residence of the customer.

To obtain conditional averages of income and home value for each customer, we need to construct joint distributions of age, income and home value for each of the 40 zip codes in which the customers reside. For this, we need two sources of information. The first is data on marginal distributions of variables for all zip codes in the state of Connecticut. These data would provide information on how age, income and home value are distributed independently in each zip code. However, publicly available census data on such distributions are only available for Census Block Groups (CBGs) and not zip codes. Zip codes are constructs used by the United States Postal Service and there is no one-to-one mapping from CBGs to zip codes.

We therefore use zip-code and CBG equivalence data that can be purchased from third party geographic data providers to obtain approximate zip code level marginal distributions from census data. The zipcode-CBG equivalence data indicate the number of zip codes over which a particular CBG is spread out and vice-versa. For simplicity, we assume that the extent of overlap is uniform. For example, if a CBG is spread over three zip codes, we assume that the population spread is equal over the three zip codes. We do the same if a zip code is spread over multiple CBGs and then aggregate the information for each zip code. Since the census data are reported as ordinal distributions, similar to the format in Table 1, we can obtain the population in each zip code that belongs to a particular ordinal variable category using this simple allocation rule. Note that the approximation is necessary only because of data limitations and not due to any inherent limitations of our approach. If zip code level marginal distributions were directly available, the above approximation would not be required.

Still, the marginal distributions so obtained do not have any information on the association (correlation) between variables. For instance, it is impossible from Table 1 to ascertain the proportion of zip code 06520 that belongs to the (<\$50k) income and (<\$100k) home value category. We therefore use another piece of information - survey data from the census on a sample of individuals across the entire state of Connecticut (i.e. aggregate market sample). Specifically, we use the Public Use Microdata Sample (PUMS) of the American Community Survey. Note that even if such data were not available from the census on variables of interest to marketers, they could cost-effectively conduct their own surveys at the state level on variables of interest. To link zip code distributions with the aggregate market sample, we use zip code populations which are also publicly available.

## 5.2 Inferring Joint Distributions

We first estimate joint distributions for the three demographic variables: age, income and home value. Recall that the aggregate market sample is modeled as a mixture over all zip code distributions. Therefore, although we are only interested in obtaining joint distributions for 40 zip codes, we still have to use data from all zip codes to infer correlations between variables.<sup>6</sup>

The marginal distributions of all three variables across zip codes appear to be skewed and they take only positive values. For these reasons, we model these variables as log-normal distribution, instead of a normal distribution. Therefore  $\log(\text{Age})$ ,  $\log(\text{Income})$  and  $\log(\text{Home Value})$  across zip codes follow a multivariate normal distribution. Also, since zip code data are ordinal, we cannot estimate the joint distributions from this data directly but have to use data augmentation to obtain zip code marginal distributions first. We use Bayesian MCMC based methods for posterior inferences on the unknown correlations. The first 4000 draws were used for burn-in and the next 4000 were used for inference. The details of the estimation algorithm are in Appendix A.

It is worth mentioning here that even though we use only three variables, they are ordinal with 10-13 levels each, still creating a large number of combinations if we use a contingency table approach as in Putler et al. (1996).

Table 5 shows the estimated correlations for variable pairs where the correlation estimates of our model differ from those of aggregate survey. The correlation between Age and Income is estimated to be negative, while the correlation between home value and income is positive. The correlation between age and home value is statistically insignificant.

---

<sup>6</sup> The state of Connecticut has over 1000 zip codes. As a practical matter, we limited ourselves to a random sample of 100 zip codes to infer the correlations. This was done purely for faster computation and does not limit our approach in any manner.

**Table 5. Comparison of correlations in the aggregate survey sample with model results**

	<i>Model Results</i>	<i>Aggregate Market Sample</i>
Age, Income	<b>-0.33*</b> [-0.42,-0.26]	-0.27
Age, Home Value	-0.05 [-0.15, 0.03]	-0.08
Home Value, Income	<b>0.34*</b> [0.28,0.44]	0.47

Note: (1) The intervals reported under Model Results are the intervals containing 95% of the posterior simulated samples. Only 2 out of the 3 possible correlations are significantly different from those obtained from the aggregate survey sample.

(2) All variables are log-transformed.

(3) \*  $p < 0.1$ ; \*\*  $p < 0.05$

Table 5 also shows that correlations obtained from the aggregate survey sample are significantly different from those obtained from our model. Hence, the aggregate survey sample correlations cannot be used in place of zip code correlations directly and doing so could lead to incorrect inference. These correlations are used in conjunction with the zip code means and standard deviations to construct the multivariate normal joint distributions for each of the 40 zip codes of interest. We now apply the joint distributions we have obtained to a real world problem and show how the unavailability of joint distributions at the zip code level could affect managerial decision-making.

### 5.3 Target Selection

We first estimate the four models listed in Table 3 using the bank data and the results are reported in Table 6. This exercise is similar to the one presented in the simulation study, but with real data.

Table 6. Model Comparison

	Estimation Results			
	Standard Approach	Steenburgh, et al. (2003)	Equal Covariance a la Romeo (2005)	Equal Correlation (Proposed Model)
<i>Intercept</i>	0.03 [-0.091, 0.151]	0.024 [-0.077, 0.164]	0.259 [-0.687, 1.150]	<b>0.690**</b> [0.290, 1.104]
<i>Age</i>	-0.04 [-0.089, 0.007]	<b>-0.047*</b> [-0.093, 0.001]	0.037 [-0.260, 0.309]	<b>0.208**</b> [0.072, 0.404]
<i>Income</i>	-0.063 [-0.134, 0.013]	-0.316 [-19.117, 18.909]	0.122 [-0.301, 0.509]	<b>0.276**</b> [0.134, 0.404]
<i>Home Value</i>	<b>0.118**</b> [0.050, 0.180]	0.181 [-20.282, 19.866]	-0.097 [-0.447, 0.253]	-0.189 [-0.397, 0.042]
$\sigma_\varepsilon^2$	<b>0.07**</b> [0.066, 0.074]	<b>0.069**</b> [0.065, 0.073]	<b>0.069**</b> [0.065, 0.073]	<b>0.068**</b> [0.064, 0.072]
$\sigma_v^2$		<b>0.153**</b> [0.110, 0.213]	<b>0.157**</b> [0.113, 0.219]	<b>0.157**</b> [0.112, 0.217]
<i>Log-Marg. Likelihood</i>	<b>-461.9</b>	<b>-451.1</b>	<b>-448.9</b>	<b>-438.9</b>

Note: (1) The intervals reported are the intervals containing 95% of the posterior simulated samples.  
 (2) All variables are log-transformed.  
 (3) \*  $p < 0.1$ ; \*\*  $p < 0.05$

The proposed model has the highest log-marginal likelihood (Gelfand and Dey, 1994) among all the models. Even, the equal covariance model (Romeo, 2005) performs worse than our model indicating that our approach of assuming only equal correlations to obtain joint distributions is superior in terms of fitting the data.

The coefficient on Income is not significant in any of the models except ours. The results of the other three models could lead managers to believe that Income does not have an effect on

customer profitability. On the other hand, the coefficient of Home Value is insignificant after controlling for Age and Income, but the standard model (i.e. the first benchmark model), indicates a significant positive impact of Home Value on profitability, leading to incorrect managerial inference.

As we had discussed earlier, using zip code averages for Income and Home Value when Age, Income and Home Value are correlated causes dependence between Age and  $\varepsilon$ . Standard regression models assume independence of  $\varepsilon$  leading to endogeneity bias in the Age coefficient. This kind of endogeneity bias would occur whenever we have missing data at the individual level and use aggregate averages to fill-in for the missing data, if we do not account for correlations between variables. It can be seen that the coefficient on Age, which is an individual level variable, is biased downward in the benchmark models as compared to our model. The downward nature of the bias is due to negative correlation between Age and Income that we had estimated earlier. In fact the extent of the bias is so severe in the Steenburgh et al. (2003) model that the coefficient is of opposite sign and significant. Thus when there is correlation between the variables observed at the individual level and only at the group level, managerial decisions based on ignoring this correlation can lead to very erroneous decisions. We next illustrate this point using a target selection exercise.

Banks often target geographic units (zip codes, carrier routes, etc) with direct mail to acquire customers. The main challenge that they face is to select geographies with the most profitable customers, given their acquisition budget. Typically, banks would rank markets based on profitability and select the top 'x' percent for targeting. In our target selection application, we take a group of 100 zip codes and rank them based on profitability. Then we look at the top 50 zip codes as our acquisition target. We do this for two models, Steenburgh et al. (2003) and the

proposed model (4), and compare the selected zip codes to see if there are any real managerial consequences. Indeed, we find that out of the 50 zip codes selected by the two models, 45 are different, i.e. 90% of the zip codes are different. This difference is mainly because of the erroneous sign on the Age coefficient in the Steenburgh et al. (2003) model, due to the bias identified in this paper. Banks spend several hundred thousand dollars every month on direct mail; therefore the inefficiency due to such erroneous targeting is very large in absolute terms and can affect the bottom-line substantially.

## **6. Conclusion**

Marketers often have limited data at the individual level; therefore they need to augment the individual data with variables that are available only at an aggregate level to aid decision making. In this article, we describe the case of a direct marketer that has to augment information at the individual customer level with information at the zip code level for a target selection problem. We demonstrate that this standard approach of using aggregate level zip code averages leads to biased inference and erroneous decision making. Specifically, we demonstrate that when the underlying variables that are observed at the individual and zip code level are correlated, the standard approach of using aggregate zip code averages without accounting for the correlations will lead to biased estimates. We therefore recommend the use of conditional averages, where we condition the group averages on information that is observed at the individual level.

To solve this problem, we develop an approach to infer joint distributions for each zip code by combining zip code level marginal distributions with a state-level survey sample which is publicly available or can be relatively easily collected. The approach is easily scalable to deal with typical market scenarios involving a large number of variables, even when the marginal

distribution of a variable is reported only as ordinal information. Nevertheless, there are other opportunities to adapt our solution to more general settings. While we reduce the dimensionality of the problem by making a multivariate normal assumption for the joint distribution of variables, this may not be appropriate for multi-modal distributions. One solution to this problem is to use a mixture of normal distributions, which can be quite flexible, to capture multimodality. Another solution is to use an infinite mixture model or Dirichlet process which circumvents the problem of determination of number of mixture components. Future research needs to address this issue.

Though we have illustrated our technique in the context of a target selection problem, our approach to infer joint distributions has wide applicability across a number of domains in marketing and empirical industrial organization. It is relevant whenever data from several markets are pooled together in estimating consumer demand and yet one needs to model customer heterogeneity appropriately for each market. For instance, Nevo (2001) estimates demand across many local markets as a function of their demographic characteristics. Romeo (2005) illustrates a similar problem where he uses data on the marginal distributions of store local trading area demographic characteristics within a city and a sample of consumers from across the city. Zhu and Singh (2005) study entry of discount stores into different local markets whose attractiveness is a function of demographic characteristics. In general, much of the literature on spatial models should find the joint distribution approach valuable in modeling observed customer heterogeneity in a specific market. We hope the techniques discussed in this paper spawn additional research in direct marketing and more generally in other marketing and empirical industrial organization settings that requires estimating joint distributions of customer heterogeneity within each market.

## References

- Albert, James H. and Siddhartha Chib (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", *Journal of the American Statistical Association*, 88(422), 669-679.
- Barnard, J., McCulloch, R. and Meng, X. (2000), "Modeling Covariance Matrices in terms of Standard Deviations and Correlations with Applications to Shrinkage", *Statistica Sinica*, 10, 1281-311.
- Chen, Ming-Hui and Dipak K. Dey (1998), "Bayesian Modeling of Correlated Binary Responses via Scale Mixture of Multivariate Normal Link Functions", *Sankhya: Series A*, 60(3), 322-343.
- Chib, Siddhartha and Edward Greenberg (1998), "Analysis of Multivariate Probit Models", *Biometrika*, 85(2), 347-361.
- Chintagunta, Pradeep K. and Jean-Pierre Dube (2005), "Estimating a Stockkeeping-Unit-Level Brand Choice Model That Combines Household Panel Data and Store Data", *Journal of Marketing Research*, 42(3), 368-379.
- Gelfand, Alan E. and Dipak K. Dey (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations", *Journal of the Royal Statistical Society: Series B*, 56(3), 501-514.
- Nevo, Aviv (2001), "Measuring Market Power in the Ready-to-Eat Cereal Industry", *Econometrica*, 69(2), 307-342.
- Putler, Daniel S., Kirithi Kalyanam and James Hodges (1996), "A Bayesian Approach to Estimating Target Market Potential with Limited Geo-Demographic Information", *Journal of Marketing Research*, 33(2), 134-149.

- Romeo, Charles (2005), "Estimating Discrete Joint Probability Distributions for Demographic Characteristics at the Store Level Given Store Level Marginal Distributions and a City-Wide Joint Distribution", *Quantitative Marketing and Economics*, 3(1), 71-93.
- Steenburgh, Thomas, Andrew Ainslie and Peder Hans Engebretson (2003), "Massively Categorical Variables: Revealing the Information in Zip Codes", *Marketing Science*, 22(1), 40-57.
- Tanner, Martin A. and Wing Hung Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation", *Journal of the American Statistical Association*, 82(398), 528-540.
- Zhu, Ting and Vishal Singh (2005), "Empirical Analysis of Entry and Location Choice in Discount Retailing", *Working Paper*.

## Appendix

### A. Reconstructing Joint Distributions from Ordinal Data and Aggregate Market Sample

Estimating the correlation matrix  $R$  for *continuous variables* with known group means and variances needs only a Metropolis-Hastings sampler which is detailed in Step 4 of the Gibbs Sampler below. For *ordinal variables* we do not know the zip code means and variances. Hence, they need to be inferred from the data. Let  $j = 1, \dots, J$  be the groups. For any group  $j$ , assume there are categories:  $\{C_k^1, \dots, C_k^{M_k}\}$  for variable  $k$  and we observe the number of individuals

$n_{jk}^m$  (or the proportions) in category  $C_k^m$ . Let  $\sum_{m=1}^{M_k} n_{jk}^m = n_{jk} = n_j \quad \forall k$ . Let  $\mu_{jk}$  and  $\sigma_{jk}^2$  be the mean

and variance for the latent variable  $X_{jk}$ . Let  $D_j = \begin{pmatrix} \sigma_{j1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_{jK} \end{pmatrix}$ .

#### *Prior Distributions*

1.  $\mu_{jk} \sim N(\mu_0, \sigma_{\mu_0}^2)$ , where  $\mu_0$  and  $\sigma_{\mu_0}^2$  are parameters
2.  $\sigma_{jk}^2 \sim \text{Inverse Gamma}(a_0, b_0)$ , where  $a_0$  and  $b_0$  are parameters
3. Following Barnard et al. (2000), we specify a non-informative prior for  $R$  such that each non-one entry of  $R$  has a marginal uniform prior on  $(-1, 1)$ . Barnard et al. (2000) shows that this prior can be derived from an Inverse-Wishart distribution with  $K+1$  degrees of freedom. The prior  $\pi(R)$  is proportional to

$$|R|^{\frac{K(K-1)}{2}} \left( \prod_{k=1}^K |R_{kk}| \right)^{-\frac{K+1}{2}}$$

where  $R_{kk}$  is the k-th principal sub-matrix of  $R$ . Note that we do not need to know the normalizing constant of this prior in the Metropolis-Hastings algorithm for inference on  $R$ .

### ***Full Conditionals for the Gibbs Sampler***

1. The full conditional distribution for  $\mu_{jk}$  is normal,

$$N \left[ \left( \frac{n_{jk}}{\sigma_{jk}^2} + \frac{1}{\sigma_{\mu_0}^2} \right)^{-1} \left( \frac{\sum_{l=1}^{n_{jk}} X_{jkl}}{\sigma_{jk}^2} + \frac{\mu_0}{\sigma_{\mu_0}^2} \right), \left( \frac{n_{jk}}{\sigma_{jk}^2} + \frac{1}{\sigma_{\mu_0}^2} \right)^{-1} \right], \text{ where } X_{jkl} \text{ is the latent continuous}$$

variable

2. The full conditional distribution for  $\sigma_{jk}^2$  is inverse-gamma,

$$IG \left[ a_0 + \frac{n_{jk}}{2}, \frac{1}{2} \left\{ \sum_{l=1}^{n_{jk}} (X_{jkl} - \mu_{jk})^2 + 2b_0 \right\} \right]$$

3. The full conditional distribution for the latent continuous variable  $X_{jkl}$  is truncated normal,

$$N(\mu_{jk}, \sigma_{jk}) I(\gamma_k^{m_k-1} \leq X_{jkl} \leq \gamma_k^{m_k})$$

4. For sampling the correlation matrix  $R$ , we follow the Metropolis Hit-and-Run algorithm developed in Chen and Dey (1998). If  $g$  is the current iteration of the chain, then the proposal distribution for  $R^{(g)}$  is defined as  $R^{(g)} = R^{(g-1)} + H$  where the entries of  $H$  are sampled as follows:

- a. Sample i.i.d  $N(0,1)$  variables  $\zeta_{12}, \zeta_{13}, \dots, \zeta_{K-1,K}$ ;

b. Sample a signed distance  $d$  from  $N(0,1)$  truncated to  $\left(-\frac{\xi^{(g-1)}}{\sqrt{2}}, \frac{\xi^{(g-1)}}{\sqrt{2}}\right)$ , where

$\xi^{(g-1)}$  is the least eigenvalue of  $R^{(g-1)}$

c. Let  $H_{kk'} = \frac{\zeta_{kk'} d}{\left(\sum_{p=1}^{K-1} \sum_{q=1}^K \zeta_{pq}^2\right)^{\frac{1}{2}}}$ .  $R^{(g)}$  be accepted with probability

$$\min \left\{ 1, \frac{\prod_{i=1}^n \left( \sum_{j=1}^J \omega_j N\left(Z_{i\bullet} \mid \left(\mu_j, D_j R^{(g)} D_j\right)\right) \right) \pi\left(R^{(g)}\right) \left( \Phi\left(\frac{\xi^{(g)}}{\sqrt{2}\sigma_d}\right) - \Phi\left(\frac{-\xi^{(g)}}{\sqrt{2}\sigma_d}\right) \right)}{\prod_{i=1}^n \left( \sum_{j=1}^J \omega_j N\left(Z_{i\bullet} \mid \left(\mu_j, D_j R^{(g-1)} D_j\right)\right) \right) \pi\left(R^{(g-1)}\right) \left( \Phi\left(\frac{\xi^{(g-1)}}{\sqrt{2}\sigma_d}\right) - \Phi\left(\frac{-\xi^{(g-1)}}{\sqrt{2}\sigma_d}\right) \right)} \right\}$$

where  $\Phi$  is the standard normal cumulative distribution function.

## B. Sampling Algorithm for the Empirical Illustration

Let there be  $j = 1, \dots, J$  zip codes. In each zip code we observe  $i = 1, \dots, n_j$  customers and their corresponding characteristics  $X_{ij}$ . Consider the model described in (4) and let  $\tilde{Z}_{ij} = E_j(Z_i \mid X_i)$ .

The likelihood function for this model is

$$\prod_{j=1}^J \prod_{i=1}^{n_j} N\left(Y_{ij} \mid \left(X_{ij}^T \alpha + \tilde{Z}_{ij}^T \beta + v_j, \tilde{\sigma}_j^2\right)\right)$$

### Prior Distributions

1.  $\alpha \sim N(\mu_\alpha, \Sigma_\alpha)$ , where  $\mu_\alpha, \Sigma_\alpha$  are prior parameters
2.  $\beta \sim N(\mu_\beta, \Sigma_\beta)$ , where  $\mu_\beta, \Sigma_\beta$  are prior parameters
3.  $\tilde{\sigma}^2 \sim IG(a, b)$ , where  $a, b$  are prior parameters

4.  $\sigma_v^2 \sim IG(a_v, b_v)$ , where  $a_v, b_v$  are prior parameters

$$\text{Let } \eta = (\alpha, \beta), \mu_\eta = (\mu_\alpha, \mu_\beta) \text{ and } \Sigma_\eta = \begin{bmatrix} \Sigma_\alpha & \\ & \Sigma_\beta \end{bmatrix}$$

### Full Conditionals for the Gibbs Sampler

1) The full conditional for  $\eta$  is normal,  $N\left(H_\eta \left( \sum_{j=1}^J W_j \Omega_j^{-1} \tilde{Y}_j + \Sigma_\eta^{-1} \mu_\eta \right), H_\eta\right)$

$$\text{where } H_\eta = \left( \sum_{j=1}^J W_j \Omega_j^{-1} W_j^T + \Sigma_\eta^{-1} \right)^{-1}, Y_j = (Y_{1j}, \dots, Y_{n_j, j})^T, W_j = \begin{bmatrix} X_{1j} & \cdot & \cdot & \cdot & X_{n_j, j} \\ \tilde{Z}_{1j} & \cdot & \cdot & \cdot & \tilde{Z}_{n_j, j} \end{bmatrix},$$

$$\tilde{Y}_j = Y_j - \nu_j \mathbf{1}_{n_j} \text{ and } \Omega_j = \tilde{\sigma}_j^2 I_{n_j}$$

2) The full conditional for  $\tilde{\sigma}^2$  is inverse-gamma,

$$IG \left[ a + \frac{\sum_{j=1}^J n_j}{2}, b + \frac{1}{2} \left\{ \sum_{j=1}^J \sum_{l=1}^{n_j} (Y_{lj} - X_{lj}^T \alpha - \tilde{Z}_{lj}^T \gamma - \nu_j)^2 \right\} \right]$$

3) Let  $\tilde{Y}_j = Y_j - X_j^T \alpha - \tilde{Z}_j^T \beta$ ,  $\nu = (\nu_1, \dots, \nu_J)$  and  $\Omega_j = \tilde{\sigma}_j^2 I_{n_j}$ . The full conditional for  $\nu$  is

$$N\left( (\Lambda + \Sigma_\nu^{-1})^{-1} \zeta, (\Lambda + \Sigma_\nu^{-1})^{-1} \right), \quad \text{where } \Lambda_{J \times J} = \text{diag}\left(\text{tr}(\Omega_1^{-1}), \dots, \text{tr}(\Omega_J^{-1})\right) \quad \text{and}$$

$$\zeta = \left( \mathbf{1}_{n_1}^T \Omega_1^{-1} \tilde{Y}_1, \dots, \mathbf{1}_{n_J}^T \Omega_J^{-1} \tilde{Y}_J \right)$$

4) The full conditional for  $\sigma_v^2$  is  $IG\left(a_v + \frac{J}{2}, b_v + \frac{1}{2} \nu^T \nu\right)$