

# Predicting Individual Response with Aggregate Data: A Conditional Means Approach

Jason A. Duan<sup>a,\*</sup>, Sachin Sancheti<sup>b</sup>, K. Sudhir<sup>b</sup>

<sup>a</sup> McCombs School of Business, University of Texas at Austin

<sup>b</sup> Yale School of Management, Yale University

## Abstract

Researchers often predict individual behavior using a combination of individual and aggregate variables (e.g., zip code means). We show that in a linear model coefficients associated with individual variables will be inconsistently estimated if those variables are correlated with the underlying individual variables summarized to obtain the aggregate group means. Consistent estimates can be obtained if the group means are conditioned on observed individual variables. We introduce an approach to infer group level joint distributions to obtain such conditional estimates. The approach is illustrated in an application predicting the profitability of a bank's customers.

Keywords: aggregate variable, estimation bias, Gaussian latent variable, shrinkage estimator.

*JEL* classification: C13, C15

---

\* Corresponding author: Jason A. Duan, One University Station B6700, Austin, TX 78712. Tel.: (512) 232-8323

## 1. Introduction

Researchers in economics, sociology, marketing and public health often face the problem of limited data on individual characteristics. For example, marketers have access to detailed behavioral data on consumers through loyalty programs, online transactions, etc. but certain descriptive demographic data important for segmentation and targeting, such as household income and family size, may not be available. Similarly, in surveys of health and financial behavior, important explanatory variables like individual's education or occupation, may be missing. The unavailability of crucial individual characteristics often proves a hindrance to empirical research.

To overcome this problem, researchers commonly augment available individual data with mean socioeconomic characteristics of the corresponding geographic group (census tract, zip-code, etc) and use the aggregate information in lieu of the missing data. This kind of aggregate group information is usually obtained from secondary sources such as the census, data intermediaries like Claritas or Experian, and alternative surveys collecting the relevant information.

However, the standard approach of using aggregate proxies for individual level data is problematic. For example, Steenburgh et al. (2003) demonstrates that using mean zip code demographics without accounting for unobserved zip code effects exaggerates the precision of parameter estimates. This is because the approach implicitly assumes that all the variation across zip codes is due to the observed average zip code demographics. The authors 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. Similarly, a very recent paper, van Dijk and Paap

(2008), shows that using OLS to estimate individual response with aggregate explanatory variables is inefficient and proposes a latent variable approach to improve efficiency.,

In this paper, we argue that using the standard approach not only has efficiency issues, but it may also lead to bias in estimates. The intuition for the bias is as follows. Suppose a researcher wants to estimate the effect of customer's age and income on profitability but only has information on age at the individual level. Following standard practice, customer age enters the model at the individual level, but income enters as the mean value for the zip code. Further suppose that the most profitable customers tend to be higher income, older customers within any zip code and older customers tend to have higher incomes. 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 mean zip code income causes the residuals to be systematically correlated with age, thus biasing the estimates. Geronimus et al. (1996) have also discussed the cause of this inherent bias and assessed its magnitude using two datasets from the Panel Study of Income Dynamics and the National Maternal and Infant Health Survey. However, they do not propose any solution.

Our solution to this problem is based on the observation that using group-level conditional means given observed individual level variables will result in residuals that are not correlated with any of the observed individual level variables. In the example above, using conditional mean of income given age instead of the unconditional mean removes the correlation between age and the residuals. Therefore, if we know the conditional means of the missing individual characteristics given the observed variables, we should be able to consistently estimate all the individual effects.

In practice, group-level conditional means can be obtained from the group-level joint distributions of individual variables. The challenge however is that group-level joint distributions are not available directly. For example, census data are mostly reported as contingency tables of individual variables and most surveys do not have enough observations sampled from each group to estimate joint distributions. Moreover, individual characteristics can be a mixture of continuous and discrete ordinal variables and defining a parametric joint distribution for them can be difficult.

Putler et al. (1996) addressed the issue of obtaining zip code joint distributions for categorical variables. 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. 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 because the number of the parameters that need to be estimated grows exponentially.

Romeo (2005) proposes a solution to alleviate the exploding 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 aggregate 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. This assumption is typically problematic because we

know from the data that there is considerable heterogeneity in the variances across different zip codes.

The key methodological contribution of this paper is that we propose a practically feasible approach to obtain group-level joint distributions of individual characteristics using survey samples. Since individual characteristics can be both continuous and discrete ordinal, our approach uses Gaussian latent variables to transform individual characteristics into a sample from a multivariate Gaussian distribution, similar to Gaussian Copulas (e.g., Clemen, et al. 1999, Pitt, et al. 2006). The dependence structure, i.e. the correlation matrix is then estimated either using an empirical Bayes method or a finite mixture method. The approach is both flexible, in that it does not have the restrictive assumptions like Romeo (2005), and scalable, in that it alleviates the dimensionality problem of Putler, et al. (1996) through the use of Gaussian latent variables. Liechty et al. (2004) proposes a Bayesian correlation estimation method for correlation matrices of Gaussian random variables. However, the computational cost of that method is too high if there is a large number of groups and hence many high-dimensional correlation matrices are to be estimated. Simulated examples indicate that our method performs very well with a reasonably large number of groups and variables.

The rest of the paper will be organized as follows. Section 2 explains the bias and inconsistency problem when aggregate data are used as proxies and show how the problem can be solved by using conditional means. Section 3 describes the procedures to estimate joint distributions, which are needed to obtain conditional means. Section 4 reports the results of simulation analyses validating the procedures. Section 5 provides an empirical illustration in the context of a bank's customer profitability problem. Section 6 concludes.

## 2. The Bias Problem Using Aggregate Data and Its Correction Method

As discussed in the introduction, the standard approach is to append the observed individual level variables ( $X$ ) with variables available at the group level (e.g., zip code, census tract) ( $Z$ ) to proxy for the missing individual variables of interest:

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

where  $i$  indexes individuals and  $j(i)$  indexes the group 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 that are only available as summary statistics (e.g., mean, standard deviation) at the group level. For most of the cases in the literature,  $Z_{j(i)}$  is the group-level mean.

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 or omitted group effects Steenburgh et al. (2003) propose including group level random effects as in the following equation:<sup>1</sup>

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

$\nu_j$  is the parameter (fixed or random effect depending on the situation) for group dummies.

Steenburgh et al. (2003) compare model (2) with the standard model (1) and address the efficiency issue by showing that the standard errors for  $\beta$  are much lower in the standard model. The addition of random effect removes the spurious confidence in the estimate of  $\beta$ . Thus they

---

<sup>1</sup> Specifically, Steenburgh et al. (2003) estimate the equivalent hierarchical model:

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

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

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

correct the “precision” problem in estimation. However, we show below that there is a bias problem that remains and cannot be eliminated through the introduction of  $\nu_j$ .

To understand why the bias occurs, consider the following: Suppose we could observe the individual-level values (denoted by  $Z_i$ ) for the missing variable  $Z_{j(i)}$ . Then instead of Models (1) and (2) we would certainly prefer to use the model of full information

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

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

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

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

By letting  $\varepsilon_i = (Z_i - Z_{j(i)})^T \beta + e_i$  be the random error, we reduce the above model to (2), where  $Z_i$  is not observed and only the group-level values are available. It is easy to see that using the group mean  $Z_{j(i)}$  (or group median and other estimates of the mean in practice) instead of individual characteristics  $Z_i$  in the model causes the individual-specific deviation to be 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  $\alpha$  biased and asymptotically inconsistent. Only when  $X_i$  and  $Z_i$  are uncorrelated, will  $\alpha$  in model (2) be consistently estimated. However, in many applications, correlation between  $X_i$  and  $Z_i$  is not negligible. For example, in our empirical illustration, demographic variables such as *Age*, *Income* and *Home Value* are all correlated, with *Income* available only as a zip-code mean. The cause and magnitude of the bias in estimation using this approach has been

assessed by Geronimus et al. (1996). In this paper, we propose a bias correction method using available secondary individual data sampled across groups.

Suppose we know how to evaluate the group-level conditional mean  $E_j(Z_i | X_i)$  given the observed  $X_i$ . We can certainly use  $E_j(Z_i | X_i)$  as the proxy instead for  $Z_i$  in the model

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

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

It is rather straightforward to show that Model (4) provides us an unbiased least square or ML (assuming normality) estimator for  $\alpha$ . As shown below,  $\tilde{\epsilon}_i$  will have zero mean given  $X_i$ , and therefore is uncorrelated with  $X$  in the model. Indeed,

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

Model (4) also enables unbiased estimation of the original  $\beta$  in the full model (3). To see that  $\tilde{\epsilon}_i$  is also uncorrelated with  $E_j(Z_i | X_i)$ , it is trivial to verify that

$$E\{(Z_i - E_j(Z_i | X_i))[E(Z_i | X_i) - E(Z_i)]\} = 0.$$

In contrast, neither Model (1) nor (2) can provide unbiased least square estimator or MLE for  $\alpha$ . We therefore have the following proposition.

**Proposition 1.** *The least square estimator derived from model (4) is unbiased and consistent for the original  $\alpha$  and  $\beta$  in the real model (3) that generates the data. The maximum likelihood estimator for  $\alpha$  and  $\beta$  if we assume normality for  $\tilde{\epsilon}_i$  and  $e_i$  is also unbiased and consistent.*

The proof is given in Appendix A. Our simulated examples in Section 4 confirm this finding.

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

In practice, we generally do not know the exact conditional mean function  $E_j(Z_i | X_i)$ . However, we often have a secondary data source which has samples with both  $X_i$  and  $Z_i$  jointly observed for the individuals in every group  $j$ . Denote the secondary data as  $X_i'$  and  $Z_i'$ , which are generally not a subset of or matched in identity to primary  $X_i$  and  $Z_i$ . If we assume  $(X_i, Z_i)$  and  $(X_i', Z_i')$  are independently and identically sampled from the population in group  $j$ , this secondary dataset can help us estimate an unbiased conditional mean function  $\widehat{E_j(Z_i | X_i)}$  and hence we can use this function to make unbiased prediction on  $Z_i$  given  $X_i$ . That is, if the unbiased predictor has  $\text{p} \lim_{\tilde{n}_j \rightarrow \infty} \widehat{E_j(Z_i | X_i)} = E_j(Z_i | X_i)$  almost everywhere for  $X_i$  where  $\tilde{n}_j$  is the sample size of the secondary data, we may use  $\widehat{E_j(Z_i | X_i)}$  in the following model

$$Y_i = X_i^T \alpha + \left[ \widehat{E_j(Z_i | X_i)} \right]^T \beta + \nu_j + \tilde{\epsilon}_i \quad (5)$$

**Proposition 2.** *The least square estimator derived from Model (5) is unbiased and consistent for the original  $\alpha$  and  $\beta$  in Model (3) that generates the data. The maximum likelihood estimator derived from Model (5) is unbiased and consistent if we assume normality for  $\tilde{\epsilon}_i$  and  $e_i$ .*

The proof is given in Appendix A.

The simplest approach to estimate  $\widehat{E_j(Z_i | X_i)}$  is of course the linear regression model

$Z_i' = X_i' \eta_j + \omega_{ij}$ . Our goal will be achieved if we can have unbiased estimation for  $\eta_j$ . That is if

$E(\hat{\eta}_j) = \eta_j$ , then we can use the unbiased predictor  $\widehat{E_j(Z_i | X_i)} = X_i \hat{\eta}_j$  in the regression model

(5). However, we often have a multivariate  $Z_i$  which has both continuous and discrete variables.

We propose a general two-step approach using Gaussian latent variables to estimate the joint distribution of  $X_i$  and  $Z_i$ , and the conditional mean of  $Z_i$  given  $X_i$  in Section 3.

To obtain asymptotically unbiased estimator for  $\alpha$  and  $\beta$ , we only need a consistent estimator for  $E_j(Z_i | X_i)$ , i.e.,

$$\text{p} \lim_{n'_j \rightarrow \infty} \widehat{E_j(Z_i | X_i)} = E_j(Z_i | X_i)$$

almost everywhere for  $X_i$ , where  $n'_j$  is the sample size of the secondary data, we may still use  $\widehat{E_j(Z_i | X_i)}$  in the regression model (5). The LSE and MLE for  $\alpha$  and  $\beta$  are asymptotically unbiased as sizes of the primary sample  $n_j$  and secondary sample  $n'_j$  both goes to infinity.

A fundamental requirement for using our approach is that the matrix  $\left[ X, \widehat{E_j(Z_i | X_i)} \right]^T \left[ X, \widehat{E_j(Z_i | X_i)} \right]$  must be invertible. This is not a stringent requirement as the data on multiple groups with heterogeneous joint distributions of  $Z_i$  and  $X_i$  for different groups will satisfy this requirement with probability one. We will use a simulated example and two real data examples in Section 4 to illustrate the bias correction power of our approach.

In the next section, we will discuss the difficulties a researcher may encounter when she is to estimate  $E_j(Z_i | X_i)$ , as both  $X$  and  $Z$  can be multivariate and the random components of  $X$  and  $Z$  can be continuous or discrete random variables. We propose a general approach to infer the joint distribution of  $X_i$  and  $Z_i$  for each group using secondary data. Once we know the joint

distribution 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.

### 3. Estimating conditional means using secondary data

The secondary data consist of individual information on variables  $X_i$  and  $Z_i$ , but not the response variable  $Y_i$ , sampled from the relevant groups (e.g. zip codes). The sources of this data usually include: (i) publicly available survey sample of individuals (e.g., American Community Survey) and (ii) a private survey conducted by the researcher. The secondary data usually come in the following forms:

a).  $X_i$  and  $Z_i$  are continuous variables and there is a relatively large sample in each group  $j$ . This is the simplest situation where a multivariate linear regression model can estimate  $E_j(Z_i | X_i)$ . We will omit further discussion on this case as the method is very similar to the conditional mean imputation method in missing data analysis (e.g., Little 1992) and it has been thoroughly studied in the literature.

b). There is a large sample in each group  $j$ , but  $X_i$  and  $Z_i$  have mixed continuous and discrete ordinal variables, so a mixed linear or generalized linear model is not very feasible in this case. We propose a two-stage Gaussian latent variable approach (Copula) to estimate the joint distribution of  $X_i$  and  $Z_i$ . The conditional mean  $E_j(Z_i | X_i)$  is obtained from the joint distribution. We discuss this case in Section 3.1.

c). There are very few samples of  $X_i$  and  $Z_i$  in each group  $j$ , which renders the estimation of the joint distribution for every group  $j$  inefficient or impossible. However, variables  $X_i$  and  $Z_i$ , though belonging to different groups, are the same explanatory variables, such as

income, home value, education, etc. Hence, they are likely to have similar correlation across all groups. Moreover, if we have a large number of groups, the total sample size of  $X_i$  and  $Z_i$  is large. Hence, implementing a full Bayesian correlation estimation approach such as Liechty et al. (2004) is excessively time-consuming in computation. We introduce a more practical shrinkage estimation procedure to obtain joint distributions and hence  $E_j(Z_i | X_i)$  for all groups using the pooled sample. This method is discussed in Section 3.2.

d). The group identities of the observations in the secondary data are not known (e.g., American Community Survey). This case has the most limited information. Hence we impose a slightly restrictive assumption that the correlation between  $X_i$  and  $Z_i$  across groups is the same and estimate joint distributions using a finite mixture approach. This is in contrast to the assumption in Romeo (2005) that the covariance is identical.  $E_j(Z_i | X_i)$  can then be very easily obtained. We discuss this case in Section 3.3.

### **3.1 Gaussian latent variable method for joint distribution estimation**

If  $Z_i$  and  $X_i$  comprise of both continuous and discrete ordinal variables, it is not feasible to build a multivariate regression model for  $Z_i$  on  $X_i$  to obtain  $E_j(Z_i | X_i)$ . Moreover, certain variables in  $X_i$  and  $Z_i$  may have nonlinear relationships. A nonlinear or nonparametric regression model for  $Z_i$  on  $X_i$  may not only be technically and computationally difficult, but also subject to model misspecification. Therefore, we propose to use the approach of latent Gaussian variables (Gaussian Copula) to estimate the joint distribution of  $X_i$  and  $Z_i$ . The conditional mean  $E_j(Z_i | X_i)$  can be directly derived or computed from the joint distribution.

Suppose we have a secondary sample of  $n_j$  individuals in group  $j$  where we have complete information on individual characteristics  $X_i$  and  $Z_i$ . This sample of individual will help us estimate the correlation between  $X_i$  and  $Z_i$ .

Marginal distributions of characteristics  $X_{i1}, \dots, X_{iK_1}$  and  $Z_{i1}, \dots, Z_{iK_2}$ , where  $K_1$  and  $K_2$  are the dimensions of  $X_i$  and  $Z_i$ , are assumed to be known in this case. We make this assumption because the inference for marginal distributions is relatively easy for both continuous and discrete variables using either the secondary dataset or some other data with group-level marginal distribution information (e.g., the census data). Various parametric and nonparametric statistical methods (e.g., kernel density estimation, etc.) for this have been thoroughly studied in the statistics literature. Therefore we omit discussion on marginal distribution estimation here. For continuous variables in  $X_i$  and  $Z_i$ , we can use the following one-to-one transformation to obtain the latent standard Gaussian variables. Suppose  $X_{ik}$  is continuous and cannot be modeled as Gaussian for the reason that it is, for example, positive or skewed or has heavy tails. Let  $F_k$  denotes the marginal cdf of  $X_{ik}$ . The latent standard Gaussian variable  $\tilde{X}_{ik}$  is constructed as  $\tilde{X}_{ik} = \Phi^{-1}(F_k(X_{ik}))$  where  $\Phi$  denotes the cdf of the standard normal distribution.

Since  $X_{ik}$  is ordinal, suppose there are  $M_k$  categories:  $\{C_k^1, \dots, C_k^{M_k}\}$ . We observe the number of individuals  $n_{jk}^m$  (or the proportions) in category  $C_k^m$ . The corresponding latent variable  $\tilde{X}_{ik}$  is constructed as follows. The probability of the  $k$ -th characteristic of individual  $i$  being in category  $C_k^m$  is

$$P\left(\frac{\gamma_k^{m-1} - \mu_k}{\sigma_k} < \tilde{X}_{ik} \leq \frac{\gamma_k^m - \mu_k}{\sigma_k}\right) \quad (6)$$

where  $\gamma_k^m$ 's are the cut-off values for the categories of variable  $k$ ;  $\mu_k$  and  $\sigma_k^2$  are the corresponding mean and variance of  $X_{ik}$ .

The joint distribution of the latent variables  $\tilde{X}_{i1}, \dots, \tilde{X}_{iK_1}$  and  $\tilde{Z}_{i1}, \dots, \tilde{Z}_{iK_2}$  is assumed to be multivariate normal with unit variance, i.e.,  $N_K(0, R_j)$  where  $R_j$  is a  $K \times K$  correlation matrix that is to be estimated. There are multiple methods for estimating  $R_j$ . Please refer to Clemen and Reilly (1999) for a survey of nonparametric order-rank correlation assessment approaches. A statistical method for estimating correlation using the transformed secondary data is summarized as follows. We adopt the Bayesian approach with the normal likelihood from  $\tilde{X}_{ik}$  and  $\tilde{Z}_{il}$  and a prior for  $R_j$ . The posterior of  $R_j$  is simply sampled given all the  $\tilde{X}_{ik}$  and  $\tilde{Z}_{il}$  in each iteration from

$$\prod_{i=1}^{n_j} N_K(\tilde{X}_i, \tilde{Z}_i | 0, R_j) \pi(R_j), \quad (7)$$

where  $\pi(R_j)$  is the prior distribution of the correlation matrix  $R_j$ . We use a non-informative prior for correlation matrices specified by Barnard et al. (2000). The technical details are provided in Appendix B.

A key question is whether this latent variable approach will lead to asymptotically unbiased or consistent estimator for  $E_j(Z_i | X_i)$ . Here we proposed a two stage approach where we assume marginal densities are either known or can be consistently estimated in the first step. Our second step involves only standard Gaussian latent variables and the correlation estimation is certainly consistent. For general results on consistency using Gaussian Copula models with parametric or nonparametric marginal density estimators, please refer to Oakes (1982) and

Bouezmarni and Rombouts (2009). For the statistical efficiency of the two-stage method, see Joe (2005). For the details of Bayesian inference for Gaussian Copulas, see Pitt et al. (2006).

### 3.2 Shrinkage estimation for the small group-level sample problem

In the following sections, we will assume our secondary data  $X_{i1}, \dots, X_{iK_1}$  and  $Z_{i1}, \dots, Z_{iK_2}$  are multivariate normal (non-normal variables can be transformed using Gaussian latent variables in the first stage as described in Section 3.1.) The major challenge here is that there are very few secondary data samples in each group to estimate the joint distribution (i.e. the correlation matrix). This often occurs when we have a survey done over a large area (e.g., a state) so that there are a large number of groups (e.g. hundreds of zip-codes in the state) but each group consequently has a small number of samples (e.g., a few thousand observations in total and hence only about a dozen in each zip-code). Let  $K = K_1 + K_2$  be the dimension of the data. Let the sample size of the secondary data in each group be  $n_j$ , which is assumed to be small relative to  $K$ . Note that in certain groups,  $n_j$  can be even smaller than  $K$ . We still assume we either know marginal means and variances of the variables in  $X$  and  $Z$  or we can estimate them using the other datasets which have large samples of marginal observations for the variable in  $X$  and  $Z$  (e.g., the census).

Let  $S_j = 1/(n_j - 1) [X_j - \bar{X}_j, Z_j - \bar{Z}_j] [X_j - \bar{X}_j, Z_j - \bar{Z}_j]^T$ , where  $\bar{X}_j$  and  $\bar{Z}_j$  represent the matrices of sample means for all the variables in groups  $j$ , be the sample covariance matrix. Let  $\hat{D}_j = \text{diag}(S_j)^{1/2}$  be a diagonal matrix of sample standard deviation and  $\hat{R}_j = \hat{D}_j^{-1} S_j \hat{D}_j^{-1}$  be sample correlation matrix. When the sample size  $n_j$  is small relative to  $K$ , the sample correlation

matrix is a very poor estimator for population correlation because of low efficiency. Moreover, if  $n_j < K$ , the resulting sample correlation matrix is singular.

Now  $X$  and  $Z$  are common variables across all groups (e.g., income, education, home value, etc.). Therefore, the correlations between these variables are very likely to be similar across different groups. In this section, we will propose a shrinkage estimator approach where the estimator of the correlation matrix of each group is a weighted average of the within-group sample correlation matrix and a matrix that incorporates the correlation information from all the other groups.

To simplify notation, let  $G_j = [X_j, Z_j]$  be the data matrix for group  $j$  and denote the  $i$ -th row vector of  $G_j$  by  $G_j^i$ . We will use a more restrictive model which assumes the variables in all the groups have a common correlation matrix to induce a ‘prior’ correlation matrix for each group. Let the covariance matrix of group  $j$  be  $\Sigma_j = D_j R D_j$ , where  $D_j$  is the diagonal matrix of heterogeneous standard deviations across groups and  $R$  is the common correlation matrix. The log-likelihood function given this assumption and the multivariate normal assumption is proportional to

$$\sum_{j=1}^J \left\{ \sum_{i=1}^{n_j} (G_j^i - \mu_j)^T D_j^{-1} R^{-1} D_j^{-1} (G_j^i - \mu_j) \right\} - \sum_j n_j \log |D_j R D_j| \quad (8)$$

$R$  is a positive definite matrix who diagonal elements are one. When  $\mu_j$  and  $D_j$  are known (usually they can be more efficiently estimated using the primary data), the maximum likelihood inference for  $R$  involves a very difficult constraint maximization within a high-dimensional *elliptic tetrahedron* (Rousseeuw and Molenberghs 1994) due to positive definiteness. We therefore propose two more feasible consistent estimators for  $R$  as follows.

The first is the moment estimator proposed by Liang and Zeger (1986). Let  $\hat{e}_j^i = G_j^i - \bar{G}_j$ , where  $\bar{G}_j = [\bar{X}_j, \bar{Z}_j]$ , and

$$\hat{R}_M = \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{n_j} \hat{D}_j^{-1} \hat{e}_j^i \hat{e}_j^{iT} \hat{D}_j^{-1}, \quad (9)$$

where  $\hat{D}_j = \text{diag}(\hat{\Sigma}_j)^{1/2}$  and  $\hat{\Sigma}_j = 1/n_j \sum_{i=1}^{n_j} \hat{e}_j^i \hat{e}_j^{iT}$  is the MLE for the group-level covariance matrix of group  $j$ . The estimator  $\hat{R}_M$  is a weighted average of sample correlation matrices across all  $J$  groups:

$$\hat{R}_M = \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{n_j} \hat{D}_j^{-1} \hat{e}_j^i \hat{e}_j^{iT} \hat{D}_j^{-1} = \sum_{j=1}^J \frac{n_j}{N} \cdot \left\{ \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{D}_j^{-1} \hat{e}_j^i \hat{e}_j^{iT} \hat{D}_j^{-1} \right\}, \quad (10)$$

where the term in the bracket is the group level sample correlation matrix. Note the group-level sample correlation matrix may be singular as  $n_j$  can be smaller than  $K$ . However, as long as  $N$  is greater than  $K$ , the estimator  $\hat{R}_M$  is positive definite with probability one. This moment estimator is consistent and easy to compute.

The second estimator we propose is a Bayes estimator (posterior mean) using the likelihood function in (8) and a non-informative prior on the correlation matrix  $R$  proposed in Barnard et al. (2000). We denote this estimator by  $\hat{R}_B = E(R | G_1, \dots, G_J)$ .<sup>2</sup> Since there is not a closed-form formula for this estimator, we use the Markov Chain Monte Carlo method to simulate samples from the posterior distribution and estimate the posterior mean  $\hat{R}_B$  by the sample average of a burned-in and thinned chain. A Metropolis-Hastings sampler for the correlation matrix is provided in Appendix B.

---

<sup>2</sup> This method is often referred as the *empirical Bayes estimator* in the statistics literature.

The shrinkage estimator for the group-level  $R_j$  is derived from an implicit Bayes estimator using either  $\hat{R}_B$  or  $\hat{R}_M$  as prior information from the previous step. We will use  $\hat{R}$  to represent either  $\hat{R}_B$  or  $\hat{R}_M$ . Suppose we use the pseudo ‘conjugate prior’  $\Sigma_j \sim \text{Inv-Wishart}(\nu, \nu \hat{D}_j \hat{R} \hat{D}_j)$  for the group level covariance  $\Sigma_j$  and let the matrix  $A_j = \sum_{i=1}^{n_j} (G_j^i - \bar{G}_j)(G_j^i - \bar{G}_j)^T$ . Now  $A_j$  is the sufficient statistic for  $\Sigma_j$  and has a  $\text{Wishart}(n_j - 1, \Sigma_j)$  distribution. Therefore, the posterior distribution for  $\Sigma_j$  using Bayes Theorem is  $\text{Inv-Wishart}(n_j - 1 + \nu, A_j + \nu \hat{D}_j \hat{R} \hat{D}_j)$  and posterior mean is

$$\begin{aligned} \tilde{\Sigma}_j &= \frac{1}{(n_j - 1 + \nu) - K - 1} \{A_j + \nu \hat{D}_j \hat{R} \hat{D}_j\} \\ &= \frac{n_j - 1 + \nu}{(n_j - 1 + \nu) - K - 1} \left\{ \hat{D}_j \left[ \frac{n_j - 1}{n_j - 1 + \nu} \hat{R}_j + \frac{\nu}{n_j - 1 + \nu} \hat{R} \right] \hat{D}_j \right\} \end{aligned} \quad (11)$$

since  $A_j = \sum_{i=1}^{n_j} (G_j^i - \bar{G}_j)(G_j^i - \bar{G}_j)^T = (n_j - 1) \hat{D}_j \hat{R}_j \hat{D}_j$

Let the shrinkage estimator  $\tilde{R}_j$  for  $R_j$  be the matrix in the central bracket above

$$\tilde{R}_j = \frac{n_j - 1}{n_j - 1 + \nu} \hat{R}_j + \frac{\nu}{n_j - 1 + \nu} \hat{R}. \quad (12)$$

$\tilde{R}_j$  is a weighted average of group-level sample correlation  $\hat{R}_j$  and  $\hat{R}$ , which draws correlation information from all the groups. This shrinkage estimator allows one to borrow information on correlation between the same variables from all the other groups. Note  $\tilde{R}_j$  is a positive definite with all diagonal elements being one with probability one because it is a weighted average of a

positive semi-definite matrix ( $\hat{R}_j$ ) and a positive definite one ( $\hat{R}$ ). This is also a consistent and efficient estimator because it is a Bayes estimator.

The remaining step to complete this shrinkage estimator is to select the degree-of-freedom parameter  $\nu$  in the prior. From equation (12), we know  $\nu$  represents the relative weight assigned to the prior information on the correlation matrix that we have obtained from the pooled data of all the groups. Intuitively, if the sample correlation matrices  $\hat{R}_j$ 's are very close to each other and hence they have small variation, we should assign higher weight to  $\hat{R}$  to achieve higher efficiency. If  $\hat{R}_j$ 's have large variation, then the prior assumption that all the variables have the same correlation across all group is less likely to be true and we should assign lower weight to  $\hat{R}$ . Selecting  $\nu$  using a Bayesian approach is a very difficult problem because all the off-diagonal elements in  $\hat{R}_j$  are correlated. Here we will propose an approximate approach based on the previous observation to select  $\nu$ .

The shrinkage estimator (12) can also be written as

$$\tilde{R}_j = \frac{1/\nu}{1/\nu + 1/(n_j - 1)} \hat{R}_j + \frac{1/(n_j - 1)}{1/\nu + 1/(n_j - 1)} \hat{R}. \quad (13)$$

Note that  $1/(n_j - 1)$  is the asymptotic variance of the Fisher's  $z$ -transform of the sample correlation given the real correlation coefficient. Therefore, we select  $1/\nu$  to be the variance of the real correlation parameters across all groups. This is because if we have an approximate Gaussian model based on the asymptotic distribution for the  $z$ -transform of correlation parameters as follows

$$z(\hat{R}_j^{k,l}) \sim N(z(R_j^{k,l}), 1/(n_j - 1)) \text{ and } z(R_j^{k,l}) \sim N(z(\rho^{k,l}), \tau^2), \quad (14)$$

where  $R_j^{k,l}$  is the real correlation between the  $k$ -th and  $l$ -th variable in group  $j$  and  $\hat{R}_j^{k,l}$  is the corresponding sample correlation and  $\rho^{k,l}$  is the prior information on this correlation, then the Bayes estimator for  $z(\hat{R}_j^{k,l})$  is

$$z(R_j^{k,l}) = \frac{\tau^2}{\tau^2 + 1/(n_j - 1)} z(\hat{R}_j^{k,l}) + \frac{1/(n_j - 1)}{\tau^2 + 1/(n_j - 1)} z(\rho^{k,l}), \quad (15)$$

which is comparable to (13) (see Daniels and Kass, 2001). Unfortunately, even an approximate closed-form model like (14) for the entire correlation matrix is technically intractable. The empirical Bayes estimator shrinking towards a correlation matrix in Daniels and Kass (2001) does not guarantee positive definiteness. Hence, we apply the same logic as in (13) to select  $1/\nu$  to be the variance of the elements in  $\hat{R}_j$  given  $\hat{R}$ . From (14), we have  $z(\hat{R}_j^{k,l}) \sim N(z(\rho^{k,l}), 1/(n_j - 1) + \tau^2)$  and we select  $\rho^{k,l} = \hat{R}_j^{k,l}$   $\tau^2 = 1/\nu$ . A moment estimator for

$1/\nu$  is simply

$$\widehat{1/\nu} = \frac{\sum_{k < l} \sum_{j=1}^J \left\{ \left[ z(\hat{R}_j^{k,l}) - z(\hat{R}_j^{k,l}) \right]^2 - 1/(n_j - 1) \right\}}{JK(K-1)/2} \quad (16)$$

The empirical properties of this method are evaluated in our simulated data examples in Section 4.1.

### 3.3 Finite mixture model for the secondary data with missing group identities

This is the case that has the most limited information so the only solution is to assume that correlations are same across groups. The secondary data are samples drawn from the  $J$  groups under study and because the group identities of these samples are not known, we use a finite mixture model to estimate the common correlation matrix. Note that although we assume

common correlations across groups, the joint distributions are still heterogenous as means and variances vary across groups.

For group  $j$ , we again assume we have multivariate normal joint distribution (for the original variables or the latent variables after the transformation in Section 3.2), i.e.,  $N_K(\mu_j, \Sigma_j)$  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 that is known or has already been estimated from marginal primary data.  $R$  is the unknown common correlation matrix. Suppose we know the group population  $n_j$  and the total population  $\sum_j n_j$ . Because the secondary sample is drawn from the aggregate population and there is no group indicator in the data, this sample is from a finite mixture distribution:

$$[G_{i1}, \dots, G_{iK}] \stackrel{iid}{\sim} \sum_{j=1}^J \omega_j N(\mu_j, \Sigma_j), \quad i = 1, \dots, n, \quad (17)$$

where the mixture weights are given by  $\omega_j = n_j / \sum_{j=1}^J n_j$ .

We can construct a likelihood function for the unknown  $R$  using the finite mixture distributions (17) for  $G_{i\bullet}$ ,  $i = 1, \dots, n_G$

$$\prod_{i=1}^{n_G} \left( \sum_{j=1}^J \omega_j N(Z_{i\bullet} | (\mu_j, D_j R D_j)) \right) \quad (18)$$

The Metropolis-Hastings algorithm to estimate the unknown matrix  $R$  is again presented in Appendix B.

#### 4. Validating the Procedures – Simulation Studies

We now report the results of two simulated data examples that validate the procedures in Section 3. The first example has two objectives: (1) to show that the shrinkage estimation procedure in Section 3.2 can recover the underlying correlation of the variables at the group level and demonstrate its empirical performance, and (2) to show that the use of group-level conditional mean for the missing individual variables helps us obtain consistent estimates for the effects at the individual level and the improvement for out-of-sample prediction. We use the second example to illustrate the use of normal latent variable (Gaussian Copula) and common correlation estimation described in Sections 3.1 and 3.3 respectively.

#### 4.1 Example for Shrinkage Estimation

We first select five normal random variables (named as  $X_1, X_2, X_3, X_4, X_5$ , representing, say, age, education, home value, income and savings) and 100 groups (e.g. 100 zip-coded areas). These variables have heterogeneous means and standard deviations across groups. To facilitate shrinkage estimation, we assume that the correlation matrices across groups are also heterogeneous but similar to each other. There are 10 correlation coefficients between these 5 variables in each group. The mean and standard deviation of each correlation coefficient across 100 groups are presented in the first column of Table 1. That is, for group  $j$ , the random variables  $\bar{X}^j = (X_1^j, X_2^j, X_3^j, X_4^j, X_5^j)$  have  $N(\mu^j, D_j R_j D_j)$  distribution, where  $\mu^j = (\mu_1^j, \mu_2^j, \mu_3^j, \mu_4^j, \mu_5^j)$ ,  $D_j$  is a  $5 \times 5$  diagonal matrix of standard deviations and  $R_j$  is a  $5 \times 5$  matrix of correlations.

We simulate a secondary dataset of 4-40 individuals with complete observations of  $(X_1^j, X_2^j, X_3^j, X_4^j, X_5^j)$  in each group. Note that if the sample size of group  $j$  is very small, the sample correlation matrix is a very poor estimator for the real correlation matrix. Indeed, most of

our groups have only 15 to 25 samples. If the sample size is 4 (there are 3 groups having only 4 samples in the example presented here), then the sample correlation matrix is singular. In total, we have 2206 samples allocated into 100 groups. We applied the shrinkage estimation in equation (13) using both the moment estimator  $\hat{R}_M$  and Bayes estimator  $\hat{R}_B$ .  $\hat{R}_M$  is calculated using (10) whereas  $\hat{R}_B$  is the posterior mean of 3000 samples from the Metropolis-Hastings chain in Appendix B. We present the means and standard deviations of the 10 correlation coefficient from 100 groups in Table 1.

Table 1 shows that both shrinkage estimation procedures (using either the moment estimator or Bayes estimator) can recover the true correlation matrices and perform equally well. Based on mean square error loss, the shrinkage estimators reduce the loss by 20% to 90% from the sample correlation estimator. Based on Stein's loss function

$$L(\Sigma, \hat{\Sigma}) = \text{trace}(\hat{\Sigma}\Sigma^{-1}) - \log(\det(\hat{\Sigma}\Sigma^{-1})) - p \quad ,$$

the shrinkage estimators reduce the loss by 40% to 100% (the Stein's loss is infinity if the estimated matrix is singular).

To test the effect of using conditional means as proxies for observed values for the regression model (4), we conduct the following simulated experiment. For the 100 groups, we sample 100 individuals with completely observed  $\bar{X}^{ij} = (X_1^{ij}, X_2^{ij}, X_3^{ij}, X_4^{ij}, X_5^{ij})$  from the same normal distribution  $N(\mu^j, D_j R_j D_j)$  for individual  $i$  of groups  $j$ . The response variable  $Y_i^j$  is generated using the simple regression model

$$Y_i^j = \alpha_0 + \alpha_1 X_1^{ij} + \dots + \alpha_5 X_5^{ij} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad .$$

The true values of the parameters  $\alpha_0, \dots, \alpha_5$  are presented in Table 2. We use 50 observations in each group to fit the models and leave 50 observations for out-of-sample prediction. For

estimation we assume that variables  $X_3^{ij}$  and  $X_4^{ij}$  are missing at the individual level. Only group level means and standard deviations of  $X_3^{ij}$  and  $X_4^{ij}$  are known. We fit three models to the simulated data: (i) the benchmark model where group-level means are treated as proxies for individual level observations

$$Y_i^j = \alpha_1 X_1^{ij} + \alpha_2 X_2^{ij} + \alpha_3 \bar{X}_3^j + \alpha_4 \bar{X}_4^j + \alpha_5 X_5^{ij} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \quad (19)$$

(ii) the model with group-level random effect  $\nu_j$  recommended by Steenburgh et al. (2003)

$$\begin{aligned} Y_i^j &= \alpha_1 X_1^{ij} + \alpha_2 X_2^{ij} + \alpha_5 X_5^{ij} + \nu_{j(i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \\ \nu_{j(i)} &= \beta_0 + \beta_1 \bar{X}_3^j + \beta_2 \bar{X}_4^j + \eta_j, \quad \eta_j \sim N(0, \sigma_\eta^2) \end{aligned} \quad (20)$$

and (iii) the model using group-level conditional means

$$\begin{aligned} Y_i^j &= \alpha_1 X_1^{ij} + \alpha_2 X_2^{ij} + \alpha_3 E_j(X_3^{ij} | X_1^{ij}, X_2^{ij}) + \alpha_4 E_j(X_4^{ij} | X_1^{ij}, X_2^{ij}) + \alpha_5 X_5^{ij} + \eta_j + \varepsilon_i, \\ \eta_j &\sim N(0, \sigma_\eta^2), \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \end{aligned} \quad (21)$$

The group level conditional means  $E_j(X_3^{ij} | X_1^{ij}, X_2^{ij})$  and  $E_j(X_4^{ij} | X_1^{ij}, X_2^{ij})$  are obtained using the group-level correlation matrix estimated by the Bayes shrinkage estimator. Maximum likelihood estimates for the regression parameters are compared in Table 2.

The results in Table 2 demonstrate the bias issue we raise in Section 2. The directions of the biased estimates for  $X_1^{ij}$  and  $X_2^{ij}$  confirm our intuition and analysis. That is,  $\alpha_1$  is biased downwards in the benchmark model (19) and the random effect model (20) because the error contributed by the missing  $X_3^{ij}$  is negatively correlated with  $X_1^{ij}$ . Likewise,  $\alpha_2$  is biased upwards since the missing  $X_4^{ij}$  and  $X_2^{ij}$  have a significant positive correlation. Note the estimate for  $\alpha_5$  is not biased in any of the models since  $X_5^{ij}$  has very low correlation (close to 0 by design, see Table 1) with the missing  $X_3^{ij}$  and  $X_4^{ij}$ . The corresponding estimates in Table 2 are for the  $\beta_1$

and  $\beta_2$  parameters in the hierarchical model (Steenburgh et al. 2001) instead. In contrast to the existing models in the literature, our model using the imputed conditional means recovers the true effects at the individual level almost perfectly.

We compare the fit of these three models using model selection criteria such as AIC and BIC, which are shown in Table 2. It is obvious that our model (21) is the most favored model among the three as it has the lowest AIC (6693.62) and BIC (6739.44) by far. Model (20) with random effects still fits the data much better than the benchmark model (19). We can compare out-of-sample predictive power of these models. As we have mentioned before, we simulated 100 samples within each of the 100 groups and hold out 50 samples in every groups for out-of-sample prediction. The benchmark model, which treats the group-level mean as a proxy for the individual effect, also performs most poorly on the out-of-sample prediction. We also calculate the square root mean square errors (RMSE) of the predicted values against the real hold-out samples for all three models (19), (20) and (21). The RMSE of the benchmark model (19) is 0.84, whereas the RMSE of the model with group-level random effects is 0.72 and the RMSE of our model (21) is 0.60. Based on our simulated example, the random effect model does a better job than the benchmark model on predicting the out-of-sample outcomes. And our model outperforms the random effect model by reducing the RMSE by 16.7% for the simulated dataset. Therefore, we conclude our model (21) is the best model among the three because it correctly estimates the parameters and has the highest predictive power.

#### **4.2 Example for the Normal Latent Variable Model and the Mixture Model**

In this example, 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 3. We again simulate 100

groups 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:  $N(3.7, 0.01)$ ,  $N(10.6, 0.09)$ ,  $N(2.7, 0.01)$  and  $N(9.6, 0.09)$ . The standard deviations across groups 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 3.

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 using ten cutoff values and save only the marginal cell-counts for all four variables. We also randomly select a sample of 2000 individuals (sampled without replacement) from all zip codes and save their complete characteristics. This random sample is treated as the aggregate sample from which the correlation matrix is estimated.

Note that it is inappropriate to use the pooled sample correlation of the secondary data to estimate the common correlation  $R$ , when the group identities in the secondary data are missing. The missing group identity also renders the moment estimator in Section 3.1 invalid. We compute the sample correlations for the aggregate market sample and present them in the second row in each cell of Table 3. It is obvious that the sample correlations are very poor estimators. 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 mixture model and Bayesian estimation procedure in Section 3.3 to estimate  $R$  and present the posterior means and 95% posterior predictive intervals for all correlation parameters in Table 3. The

detailed sampling chain is provided in Appendix C. In contrast to the sample correlations of the pooled sample, we can see that our estimates are very accurate and all the predictive intervals cover the real correlations.

## **5. Empirical Illustration**

Our real data example uses a customer dataset from a bank in northeastern USA. The bank seeks to understand the effect of demographic and geographic variables on customer profitability using the information collected from its existing customers in order to predict the profitability of new customers. We demonstrate how the conditional means approach corrects the bias in the standard approach and improves model prediction of profitability of new customers.

### **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 2377 customers residing in 73 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. Home value of each customer (household that owns at least a home) is also obtained by the bank by a certain appraisal method. However, household income is not available at the individual level; we will instead use its conditional mean given the customer's age and home value in the zip code of residence of the customer.

To obtain conditional mean of income given age and home value for each customer, we will apply the method in Section 3 to construct joint distributions of age, income and home value for each of the 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 the 73 zip codes. These data would provide information on how age, income and home value are distributed 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 zip-code-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, illustrated in Table 4, 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 4 to ascertain the proportion of zip code 06\*\*0 that belongs to the (<\$50k) income and (<\$100k) home value category. We therefore use another piece of information - survey data for a sample of individuals (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 publicly available, marketers 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. Note that the aggregate market sample is modeled as a mixture over all zip code distributions in Section 3.3. Therefore, although we are only interested in obtaining joint distributions for the 73 zip codes in our dataset, we still have to use data from all zip codes to infer correlations between variables.<sup>3</sup>

The marginal distributions of all three variables (denoted by Age, Income and Home\_Value) 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

---

<sup>3</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.

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 the Gaussian latent variable method and Bayesian estimation described in Section 3.2 and 3.3 to simulate and compute the Bayesian estimator (posterior mean) for the unknown correlations. The details of the estimation algorithms are provided in Appendices B and C. For our data example, we first simulate 4000 draws that were used for burn-in and keep the next 10000 draws for inference.

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  $\log(\text{Age})$  and  $\log(\text{Income})$  is estimated to be negative, whereas the correlation between  $\log(\text{Home\_Value})$  and  $\log(\text{Income})$  is positive. The 95% posterior interval of the correlation between  $\log(\text{Age})$  and  $\log(\text{Home\_Value})$  contains zero. Therefore, we can infer that it is not statistically significant at the 95% level.

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 73 zip codes of interest.

### **5.3 Regression Analysis**

In this section, we estimate and compare two models: the model proposed in Steenburgh et al. (2003)

$$\begin{aligned}\log(\text{Profit})_i &= \alpha_1 \times \log(\text{Age})_i + \alpha_2 \times \log(\text{Home\_Value})_i + \nu_{j(i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \\ \nu_{j(i)} &= \alpha_0 + \beta_1 \times \overline{\log(\text{Income})}_j + \eta_j, \quad \eta_j \sim N(0, \sigma_\eta^2)\end{aligned}\tag{22}$$

and our model using the group-level conditional means

$$\begin{aligned}\log(\text{Profit})_i &= \alpha_0 + \alpha_1 \times \log(\text{Age})_i + \alpha_2 \times \log(\text{Home\_Value})_i + \alpha_3 \times \overline{\log(\text{Income})}_{ij} \\ &+ \eta_j + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad \eta_j \sim N(0, \sigma_\eta^2)\end{aligned}\tag{23}$$

where the subscript  $i$  denotes the  $i$ -th individual in our data and  $j$  is the  $j$ -th zip code in which the  $i$ -th individual lives. The variable  $\overline{\log(\text{Income})}_j$  is the average log-income of zip code  $j$  and  $\overline{\log(\text{Income})}_{ij}$  is the conditional mean of individual  $i$ 's log-income given that she lives in zip code  $j$  and her observed log-age and log-home value. Individual  $\overline{\log(\text{Income})}_{ij}$  is computed using the posterior mean correlation obtained in Section 5.2 and the zip-code-level means and standard deviations in the census data (secondary data). We will not compare the results of these two models with the model that only uses the zip-code average  $\overline{\log(\text{Income})}_j$  as a direct proxy for the individual-level data without the random effects, as it is quite obvious that this simplistic model will have the poorest fit from the previous literature and our simulation studies.

We estimate the two models (22) and (23) using the bank data. We randomly select about two thirds (1641 out of 2377 samples) of the sampled customers to fit the models and leave the remaining 736 for out-of-sample prediction comparison. This exercise uses the same algorithm as the simulated data example presented in Section 4.1, but with real data. Again maximum likelihood estimates of the model parameter are computed and the results are reported in Table 6.

We note that the proposed model (23) has lower AIC and BIC than model (22) from the table, so it compares favorably in model fit for the data.

More importantly, model (22) has severely biased estimates for regression coefficients, which lead to completely different model interpretations. Firstly, because  $\log(\text{Age})$  and  $\log(\text{Income})$  are negatively correlated (correlation estimate is -0.33), the coefficient on  $\log(\text{Age})$  is biased downward to zero, if we use the zip-code-level mean of log-income. Because of this downward bias,  $\log(\text{Age})$  becomes statistically non-significant in this example, which leads to the wrong interpretation that age has no effect on profitability. Secondly,  $\log(\text{Income})$  is positively correlated with  $\log(\text{Home\_Value})$  (the correlation estimate is 0.34) and hence it biases the coefficient on  $\log(\text{Home\_Value})$  upward, which leads to an exaggerated estimate for the effect of home value. Lastly, using only few levels of zip-code-level mean income also causes the estimate for the effect of log-income to be statistically nonsignificant. In contrast, our model (23) has significant and positive estimates for the effects of  $\log(\text{Age})$  and  $\log(\text{Income})$ . The estimated effect of  $\log(\text{Home\_Value})$  is also lower than that of model (22), which corrects the upward bias.

We also assess the predictive power of models (22) and (23) using hold-out samples (736 out of 2377 customers) aforementioned and test whether the biased estimates will affect prediction. The prediction is conducted using the fitted values of the coefficients and random effects in table and the observed demographic variables and zip-code information. We compute the square roots mean square errors (RMSE) for both models (22) and (23) using the observed profit for the 736 hold-out samples and predicted profit. The RMSE for model (22) is 0.468 whereas the RMSE for model (23) is 0.447. Therefore our model reduces the RMSE by 4.5%. In conclusion, our model (23) outperforms model (22) in both model fit and out-of-sample prediction, and more importantly, it corrects biased estimates and potentially erroneous interpretations.

## 6. Conclusion

Researchers often augment individual data with variables that are available only at an aggregate level (e.g., average income for the zip code). We demonstrate that the standard approach of using aggregate group-level data as proxies for unobserved individual-level data leads to biased and inconsistent inference. Specifically, when the underlying variables that are observed at the individual and group level are correlated, the standard approach of using aggregate means or medians without accounting for the correlations will lead to biased estimates. We therefore recommend the use of group-level conditional means, where we condition the group means on information that is observed at the individual level.

To obtain group-level conditional means, we develop a procedure to infer joint distributions of all the variables for each group using a secondary dataset that is either publicly available or can be relatively easily collected by a survey. For the case when there are few samples for each group in the secondary dataset to efficiently estimate group-level joint distributions, we propose a shrinkage estimation method. We use Gaussian latent variables to deal with typical scenarios involving a combination of continuous and discrete ordinal variables. Hence our method is easily scalable. Our method can also be extended to include purely categorical variables that cannot be represented as latent continuous variables (eg. race, gender). But a large number of categorical variables will create dimensionality challenges just like Putler et al. (1996).

Though we have illustrated our technique in the context of predicting customer profitability, our methods to infer joint distributions have wide applicability across a number of domains in empirical industrial organization, marketing, etc. 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 (2009) 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 our methods valuable in modeling observed customer heterogeneity in a specific market. We hope the techniques discussed in this paper spawn additional research in a variety of settings that require estimating joint distributions of individual characteristics at a disaggregate level.

## Appendix A. Proofs of Proposition 1 and 2

### Proof of Proposition 1:

Let  $\tilde{Z}$  be the matrix of the condition means  $E_j(Z_i | X_i)$ . Treat  $v_j$ 's as fix effects and group-level intercepts that are incorporated into  $X$ . Rewrite model (4) in matrix form

$$Y = [X, \tilde{Z}] \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \varepsilon \quad (\text{A1})$$

The LSE and MLE (assuming  $\varepsilon$  is normal) are unbiased for the  $\alpha$  and  $\beta$  in model (3):

$$E\left(\left\{[X, \tilde{Z}]^T [X, \tilde{Z}]\right\}^{-1} [X, \tilde{Z}]^T Y\right) = E\left(\left\{[X, \tilde{Z}]^T [X, \tilde{Z}]\right\}^{-1} [X, \tilde{Z}]^T [X, Z] \begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right). \quad (\text{A2})$$

$$\text{Note } [X, \tilde{Z}]^T [X, Z] = \begin{bmatrix} X^T X & X^T \tilde{Z} \\ \tilde{Z}^T X & \tilde{Z}^T \tilde{Z} \end{bmatrix} + \begin{bmatrix} 0 & X^T (Z - \tilde{Z}) \\ 0 & \tilde{Z}^T (Z - \tilde{Z}) \end{bmatrix},$$

$$\text{Let } \Omega = [X, \tilde{Z}]^T [X, \tilde{Z}] = \begin{bmatrix} X^T X & X^T \tilde{Z} \\ \tilde{Z}^T X & \tilde{Z}^T \tilde{Z} \end{bmatrix}, \text{ then } [X, \tilde{Z}]^T [X, Z] = \Omega + \begin{bmatrix} 0 & X^T (Z - \tilde{Z}) \\ 0 & \tilde{Z}^T (Z - \tilde{Z}) \end{bmatrix}.$$

Note  $\Omega$  and  $\Omega^{-1}$  are matrix functions of observed  $X$ . Hence,  $E(\Omega^{-1} | X) = \Omega^{-1}$  and

$$(\text{A2}) = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + E\left(\Omega^{-1} \begin{bmatrix} 0 & X^T (Z - \tilde{Z}) \\ 0 & \tilde{Z}^T (Z - \tilde{Z}) \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right) = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

$$\text{because } E\left(\Omega^{-1} \begin{bmatrix} 0 & X^T (Z - \tilde{Z}) \\ 0 & \tilde{Z}^T (Z - \tilde{Z}) \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right) = E\left(\Omega^{-1} E\left(\begin{bmatrix} X^T (Z - \tilde{Z}) \beta \\ \tilde{Z}^T (Z - \tilde{Z}) \beta \end{bmatrix} \middle| X\right)\right) = 0$$

The formula above also shows if

$$\text{plim}_{n_j \rightarrow \infty} \frac{1}{\sum_j n_j} [X, \tilde{Z}]^T [X, \tilde{Z}]$$

is a positive definite matrix, then this least square estimator is also consistent.

## Proof of Proposition 2:

The proof follows exactly the same steps in the proof for Proposition 1. Here we replace the matrix  $\tilde{Z}$  with  $\widehat{E_j(Z_i | X_i)}$  and use the result  $E\left\{\widehat{E_j(Z_i | X_i)} | X_i\right\} = E_j(Z_i | X_i)$ . We hence omit the detailed steps. This estimator is also consistent when the sample sizes of the secondary data and the regression data both go to infinity.

## Appendix B.

The Markov Chain Monte Carlo method for simulating posterior samples to compute the Bayes estimators in Section 3 involves a step to sample the correlation matrix  $R$  with known group means and variances. A Metropolis-Hastings sampler provided here with the non-informative prior for  $R$  proposed in Barnard et al. (2000). The non-informative prior for  $R$  is specified in a way such that each off-diagonal 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  $R$ .

To sample the correlation matrix  $R$  from the posterior, 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:

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

2. 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  $R^{(g-1)}$

3. 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

$$\min \left\{ 1, \frac{\text{Likelihood}(R^{(g)} | \text{data}) \pi(R^{(g)}) \left( \Phi\left(\frac{\xi^{(g)}}{\sqrt{2}\sigma_d}\right) - \Phi\left(\frac{-\xi^{(g)}}{\sqrt{2}\sigma_d}\right) \right)}{\text{Likelihood}(R^{(g-1)} | \text{data}) \pi(R^{(g-1)}) \left( \Phi\left(\frac{\xi^{(g-1)}}{\sqrt{2}\sigma_d}\right) - \Phi\left(\frac{-\xi^{(g-1)}}{\sqrt{2}\sigma_d}\right) \right)} \right\}$$

where  $\Phi$  is the standard normal cumulative distribution function and the Likelihood( $R$ |data) can be one of any likelihood functions in Section 3.

## Appendix C.

### C.1. Estimating Joint Distributions from Ordinal Primary Data and Pooled Secondary Data.

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  $\sum_{m=1}^{M_k} n_{jk}^m = n_{jk} = n_j \forall k$ . 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  $R$  as in Appendix B.

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

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

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

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

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

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

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

7. For sampling the correlation matrix  $R$ , we follow the Metropolis Hit-and-Run algorithm developed in Chen and Dey (1998). The details are in Appendix B.

## C.2. 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  $\tilde{Z}_{ij} = E_j(Z_i | X_i)$ .

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.  $\tilde{\sigma}_j^2 \sim IG(a, b)$ , where  $a, b$  are prior parameters
4.  $\sigma_v^2 \sim IG(a_v, b_v)$ , where  $a_v, b_v$  are prior parameters

Let  $\eta = (\alpha, \beta)$ ,  $\mu_\eta = (\mu_\alpha, \mu_\beta)$  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,  $N\left(H_\eta \left( \sum_{j=1}^J W_j \Omega_j^{-1} \tilde{Y}_j + \Sigma_\eta^{-1} \mu_\eta \right), H_\eta\right)$

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

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

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

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

3) Let  $\tilde{Y}_j = Y_j - X_j^T \alpha - \tilde{Z}_{ij} \beta$ ,  $\nu = (\nu_1, \dots, \nu_J)$  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  $\sigma_\nu^2$  is  $IG \left( a_\nu + \frac{J}{2}, b_\nu + \frac{1}{2} \nu^T \nu \right)$

## References

- Barnard, J., R. McCulloch and X. Meng, 2000, Modeling covariance matrices in terms of standard deviations and correlations with applications to shrinkage. *Statistica Sinica* 10, 1281–311.
- Bouezmarni, T. and J.V.K. Rombouts, 2009, Semiparametric multivariate density estimation for positive data using copulas. *Computational Statistics and Data Analysis*, 53(6), 2040-2054.
- Chen, M. and D.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, S. and E. Greenberg, 1998, Analysis of multivariate probit models. *Biometrika*, 85(2), 347-361.
- Clemen, R.T. and T. Reilly, 1999, Correlation and copulas for decision and risk analysis. *Management Science*, 45(2), 208-224.
- Daniels, M.J. and R.E. Kass, 2001, Shrinkage estimators for covariance matrices. *Biometrics*, 57, 1173-1184.
- van Dijk, B. and R. Paap, 2008, Explaining individual response using aggregate data. *Journal of Econometrics*, 146, 1-9.
- Geronimus, A.T., J. Bound and L.J. Neidert, 1996, On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. *Journal of the American Statistical Association*, 91(434), 529-537.
- Joe, H., 2005, Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94, 401-419.
- Liang, K.Y. and S.L. Zeger, 1986, Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

- Liechty, J.C., M.W. Liechty and P. Müller, 2004, Bayesian correlation estimation. *Biometrika*, 91(1), 1-14.
- Little, R.J.A., 1992, Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420), 1227-1237.
- Oakes, D., 1982, A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*, 44(3), 414-422.
- Pitt, M., D. Chan and R. Kohn, 2006, Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3), 537-554.
- Putler, D.S., K. Kalyanam and J. Hodges, 1996, A bayesian approach to estimating target market potential with limited geo-demographic information. *Journal of Marketing Research*, 33(2), 134-149.
- Romeo, C., 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.
- Rousseeuw, P.J. and G. Molenberghs, 1994, The shape of correlation matrices. *The American Statistician*, 48(4), 276-279.
- Steenburgh, T., A. Ainslie and P.H. Engebretson, 2003, Massively categorical variables: revealing the information in zip codes. *Marketing Science*, 22(1), 40-57.
- Zhu, T. and V. Singh, 2009, Spatial competition with endogenous location choices – an application to discount retailing. *Quantitative Marketing and Economics*, 7(1), 1-35.

**Table 1. Correlation Estimation for Simulated Data**

	True Correlations		Estimated Correlations			
	Mean	St. Dev.	$\hat{R}_M$		$\hat{R}_B$	
			Mean	St. Dev.	Mean	St. Dev.
$\rho_{12}$	0.400	0.080	0.380	0.082	0.400	0.082
$\rho_{13}$	-0.600	0.060	-0.600	0.068	-0.600	0.068
$\rho_{14}$	0.300	0.100	0.290	0.091	0.310	0.900
$\rho_{15}$	0.000	0.100	0.010	0.110	0.010	0.110
$\rho_{23}$	0.300	0.100	0.290	0.095	0.300	0.094
$\rho_{24}$	0.700	0.060	0.690	0.050	0.700	0.050
$\rho_{25}$	0.000	0.100	-0.020	0.100	-0.020	0.100
$\rho_{34}$	0.400	0.080	0.390	0.085	0.400	0.085
$\rho_{35}$	0.000	0.100	-0.040	0.110	-0.030	0.110
$\rho_{45}$	0.000	0.100	-0.010	0.110	-0.010	0.110

Note:  $\rho_{ij}$  is the correlation parameter between  $X_i$  and  $X_j$ . SD is the standard deviation of the shrinkage estimations of the correlations across 100 groups.

**Table 2. Regression Analysis for Simulated Data**

	True Value	Benchmark Model (19)	Model (20) with random effects	Model (21) with conditional means
$\alpha_0$ ( $\beta_0$ )	3.00	1.96 [1.52, 2.40]	1.68 [0.88, 2.49]	2.88 [2.50, 3.25]
$\alpha_1$	1.00	0.87 [0.80, 0.95]	0.73 [0.62, 0.83]	1.03 [0.95, 1.10]
$\alpha_2$	2.00	2.76 [2.68, 2.82]	3.10 [3.03, 3.18]	2.01 [1.94, 2.07]
$\alpha_3$ ( $\beta_1$ )	0.50	0.47 [0.39, 0.57]	0.55 [0.35, 0.75]	0.47 [0.41, 0.54]
$\alpha_4$ ( $\beta_2$ )	1.50	1.65 [1.55, 1.76]	1.68 [1.45, 1.91]	1.52 [1.44, 1.60]
$\alpha_5$	2.40	2.46 [2.39, 2.53]	2.44 [2.36, 2.53]	2.42 [2.36, 2.49]
$\sigma_\varepsilon$	0.80	1.18 [1.14, 1.23]	1.10 [1.07, 1.14]	1.05 [1.02, 1.09]
$\sigma_\eta$			0.49 [0.40, 0.59]	2.10×10 <sup>-4</sup> [1.52×10 <sup>-5</sup> , 0.05]
AIC		7183.27	7059.91	6693.62
BIC		7223.36	7105.73	6739.44

Note: (1)  $\beta_0$ ,  $\beta_1$  and  $\beta_2$  are parameters in the hierarchical model (20) that are equivalent to  $\alpha_0$ ,  $\alpha_3$  and  $\alpha_4$  respectively. (2) The values in parentheses [ ] are 95% confidence intervals

**Table 3. Correlations in simulated data**

	$X_2$	$Z_1$	$Z_2$
	<b>0.10</b>	<b>0.30</b>	<b>0.20</b>
$X_1$	0.06	0.21	0.16
	<i>0.11 [0.09, 0.14]</i>	<i>0.30 [0.26, 0.33]</i>	<i>0.21 [0.18, 0.25]</i>
		<b>0.30</b>	<b>0.60</b>
$X_2$		0.22	0.53
		<i>0.32 [0.28, 0.36]</i>	<i>0.59 [0.52, 0.64]</i>
			<b>0.30</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 correlations. The middle row represents sample correlations of the pooled aggregate sample. The bottom row represents correlations obtained using our approach, with the 95% posterior interval.

**Table 4. Ordinal variables usually reported in zip code demographic data**

Zip Code	Income			Home Value		
	<\$50k	\$50k-\$100k	>\$100k	<\$100k	\$100k-\$250k	>\$250k
06**0	45%	35%	20%	20%	60%	20%
06**0	25%	65%	10%	15%	70%	15%

**Table 5. Posterior mean and predictive intervals of correlation coefficients**

---

	<i>log(Home_Value)</i>	<i>log(Income)</i>
<i>log(Age)</i>	-0.05 [-0.15, 0.03]	<b>-0.33</b> [-0.42, -0.26]
<i>log(Home_Value)</i>		<b>0.34</b> [0.28, 0.44]

---

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

**Table 6. Results and Model Comparison**

	Model (22) with random effects	Model (23) with conditional means
Intercept $\alpha_0$	<b>2.50</b> [2.34, 2.66]	<b>2.78</b> [2.52, 3.04]
$\log(\text{Age}) \alpha_1$	0.02 [-0.06, 0.1]	<b>0.14</b> [0.02, 0.26]
$\log(\text{Home\_Value}) \alpha_2$	<b>0.19</b> [0.15, 0.24]	<b>0.10</b> [0.01, 0.18]
$\log(\text{Income}) \alpha_3 (\beta_1)$	0.02 [-0.04, 0.70]	<b>0.13</b> [0.04, 0.24]
$\sigma_\varepsilon$	<b>0.45</b> [0.43, 0.46]	<b>0.44</b> [0.43, 0.46]
$\sigma_\eta$	<b>0.08</b> [0.05, 0.12]	<b>0.08</b> [0.05, 0.13]
AIC	2064.31	2057.82
BIC	2096.71	2090.23

Note: (1) The values in the parentheses [ ] are 95% confidence intervals. (2) Significant estimates for regression coefficients are in bold font.