CANADIAN APPLIED
MATHEMATICS QUARTERLY
Volume 12, Number 1, Spring 2004

PRODUCT-DRIVEN DATA MINING

RITA AGGARWALA, C. SEAN BOHUN, RACHEL KUSKE,
GERRY LABUTE, WEI LU, NILIMA NIGAM,
AND FABIEN M. YOUBISSI

Based on results obtained at the Seventh Annual PIMS Industrial Problem Solving Workshop, May 2003. Original problem submitted by Manifold Data Mining, Toronto, Ontario.

1 Introduction The behaviour of consumers is believed to be influenced by many factors. Some of these factors include the individuals culture, social status, lifestyle and attitudes. Understanding how these complicated and interrelated factors drive the consumer is the primary goal of Manifold Data Mining. The question posed to the group was to 1) find an algorithm that predicts the likelihood of consumers to respond favourably to a given product. In addition, once this prediction is made for a given consumer the group was also asked to 2) develop a second algorithm that infers other statistical information regarding the consumer.

Manifold Data Mining has developed innovative demographic and household spending pattern databases for six-digit postal codes in Canada. Their collection of information consists of both demographic and expenditure variables which are expressed through thousands of individually tracked factors. This large collection of information about consumer behaviour is typically referred to as a *mine*. Although very large in practice, for the purposes of this report, the data mine consisted of m individuals and n factors where $m \simeq 2000$ and $n \simeq 50$. Ideally, the first algorithm would identify a few factors in the data mine which would differentiate customers in terms of a particular product preference. Then the second algorithm would build on this information by looking for patterns in the data mine which would identify related areas of consumer spending.

To test the algorithms two case studies were undertaken. The first study involved differentiating BMW and Honda car owners. The algorithms developed were reasonably successful at both finding questions that differentiate these two populations and identifying common characteristics amongst the

groups of respondents. For the second case study it was hoped that the same algorithms could differentiate between consumers of two brands of beer. In this case the first algorithm was not as successful as differentiating between all groups; it showed some distinctions between beer drinkers and non-beer drinkers, but not as clearly defined as in the first case study. The second algorithm was then used successfully to further identify spending patterns once this distinction was made. In this second case study a deeper factor analysis could be used to identify a combination of factors which could be used in the first algorithm. The case studies are discussed in detail in Section 8.

2 Latent variable models The initial problem proposal suggested that the method of Projected Latent Spaces could prove fruitful in the first task, that is, in identifying a few factors which could differentiate between consumer preferences on a particular product. In essence this means finding the dominant factors that are closely related to a particular difference between customers, while showing that the remaining factors are not significant. Mathematically, we can view this as trying to represent consumer behaviour in a low dimensional space of factors. This idea is common in many different areas of application, with many different names. In this section and the next we give a discussion of latent variables and related methods. Note that in the remainder of the report we show that this method was not useful for the first algorithm for differentiating between customers with a few factors; however, we also discuss how it could be useful in improving the second algorithm.

The underlying task is to model a set of n continuous variables

$$T = (t_1, \ldots, t_n)$$

that have some joint probability density $f(t; \mu, \Sigma)$ where μ and Σ are the mean and covariance of the underlying distribution. The prime here denotes the transpose. If for example we assume that the components of T satisfy a joint Gaussian process then

$$f(t; \mu, \Sigma) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(t-\mu)'\Sigma^{-1}(t-\mu)\right]. \tag{1}$$

Since Σ is an $n \times n$ symmetric matrix and μ is an n component column vector there are

$$n + \sum_{j=1}^{n} j = \frac{1}{2} n(n+3) \tag{2}$$

free parameters in this model. If we denote $\{\tau_j\}_{j=1}^m$ as m observations (columns) of the n variables (rows) then by maximizing the logarithm of the likelihood function corresponding to (1) one obtains the usual maximum likelihood estimates

$$\hat{\mu}_{\text{mle}} = \frac{1}{m}\sum_{j=1}^{m}\tau_j, \qquad \widehat{\Sigma}_{\text{mle}} = \frac{1}{m}\sum_{j=1}^{m}(\tau_j - \hat{\mu}_{\text{mle}})(\tau_j - \hat{\mu}_{\text{mle}})'.$$

Note that the maximum likelihood estimator $\widehat{\Sigma}_{\text{mle}}$ is a biased estimator of the population covariance matrix.

As n, the number of factors one attempts to model increases, expression (2) implies that the number of free parameters grows as n^2. To reduce the number of free parameters one could simply assume that Σ is diagonal. However, this is a very drastic assumption since it is equivalent to assuming that the variables being modelled are independent. On the contrary, it is known from the data mine that some variables are very strongly correlated. One possible way of reducing the number of free parameters while still preserving the main correlations between the various factors is to choose a set of $k < n$ hidden or *latent* variables $x = \{x_1, \ldots, x_k\}$.

For a given latent variable model one specifies a density function $g(x)$ for x and some map from the latent variables into the random variables t as

$$t = y(x, \omega) + \epsilon$$

where ω are the weights that generate t and ϵ is some random variable with zero mean that is independent of x. Typically $h(\epsilon)$, the probability density of ϵ, and $g(x)$ are specified à priori. Knowing these distributions, the density of T is computed by conditioning on the latent variables so that if T is a continuous random variable,

$$u(t) = \int f(t \,|\, x) g(x)\, dx. \tag{3}$$

In summary, a given latent variable method is determined by specifying $g(x)$, $h(\epsilon)$, the map $y(x, \omega)$ and computing $u(t)$ with (3) or its generalization depending on the probability measure involved.

One example of a latent variable method is factor analysis where one specifies that y is a linear map

$$t = y(x, \omega) = \Omega x + \mu + \epsilon$$

from $\mathbb{R}^k$ to $\mathbb{R}^n$. The μ and Ω are parameters, and x, ϵ are assumed to be independent normal random variables with zero mean. For x one assumes unit covariance while for ϵ one assumes that the covariance is a diagonal matrix so that $x \sim N(0, I)$, and $\epsilon \sim N(0, \Gamma)$ where Γ is some diagonal matrix. From the structure of the map one can determine that $T \sim N(\mu, \Gamma + \Omega\Omega')$. As in the case without latent variables, one may estimate μ, Ω and Γ using a maximum likelihood estimate however even in the case of this linear model there is not a closed form for the estimates and they are typically found through an iterative process.

This linear model illustrates the point of using latent variables. In particular for factor analysis, the symmetry of $\Gamma + \Omega\Omega'$ reduces the original kn free parameters of Ω to **[2]**

$$(n+1)(k+1) - \frac{1}{2}k(k+1)$$

which only grows linearly with n. This is accomplished while still preserving some of the underlying correlation structure. The trade-off is the increase in complexity when faced with the determination of the latent variables x. We now turn to another method closely related to the projection onto latent spaces[1] (PLS). Namely principal component analysis.

3 Principal component analysis Principal component analysis (PCA) is the particular latent variable method where the k principal components are the leading eigenvectors of the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{m-1}\sum_{j=1}^{m}(\tau_j - \hat{\mu})(\tau_j - \hat{\mu})'.$$

Rather than using the covariance matrix, an alternative choice (not used here) is to base PCA on the correlation (basically standardized covariance) matrix. In either case, PCA can be viewed as a transformation that diagonalizes $\widehat{\Sigma}$ thereby reducing correlations between various combinations of factors while simultaneously finding directions in which the variance is a maximum.

Another way to view PCA is in the mean squared error sense **[5]**. With this viewpoint, the objective is to find a set of k orthonormal basis vector that span a k dimensional subspace such that the mean squared error between x and its projection onto the subspace is a minimum. As before the $n \times m$ matrix T corresponds to m observations of n random variables. If one denotes the

[1] In much of the statistical literature, the latent space methods are known as partial least squares methods. Fortunately this yields the same PLS mnemonic.

orthonormal basis set as $\{\xi_j\}_{j=1}^k$ and the projection of T onto this set as T_ξ then the expected value of the mean square error is

$$E\left(\|T - T_\xi\|^2\right) = E\left(\left\|T - \sum_{j=1}^{k}(\xi_j' T)\,\xi_j\right\|^2\right) \tag{4}$$

$$= E\left(\|T\|^2\right) - E\left(\sum_{j=1}^{k}(\xi_j' T)^2\right)$$

$$= (m-1)\left(\mathrm{Tr}\,(\Sigma) - \sum_{j=1}^{k}\xi_j'\,\Sigma\,\xi_j\right)$$

where we have assumed $E(T) = 0$ and $\Sigma = TT'/(m-1)$ is the covariance of T. By the spectral theorem every symmetric matrix $\Sigma = \Sigma'$ has a factorization $\Sigma = VDV'$ with D real diagonal and V an orthogonal matrix [**8**]. Consequently, Σ has n eigenvectors that can be chosen to be orthonormal and moreover, all of the corresponding eigenvalues $\{\lambda_j\}$ are real. From the right hand side of expression (4) one can see that the mean squared error is minimized by choosing $\{\xi_j\}$ to be any set of k orthonormal vectors [**3**].

Representation (4) also illustrates the particular advantage of choosing the first k eigenvectors of Σ. In this case the residual of the mean squared error is the sum of the absolute values of the remaining $n-k$ eigenvalues. Because of the relationship, a natural method of choosing the number of latent variables is to fix some acceptable level of error $\delta > 0$ and then choose k so that

$$E\left(\|T - T_\xi\|^2\right) = \sum_{j=k+1}^{n}|\lambda_j| < \delta.$$

It should be emphasized that these results only hold if the error is computed in the mean squared sense.

3.1 Contrasting PCA and SVD The above material shows that PCA corresponds to choosing the k dominant eigenvectors of the covariance matrix Σ. Provided one has $E(T) = 0$ this corresponds to finding the singular value decomposition (SVD) of T. To illuminate the connection, let $T = LSR'$ be the SVD of T where L and R are unitary matrices with columns that span $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. From the decomposition of T and the fact that $E(T) = 0$, one has

$$\Sigma = \frac{TT'}{m-1} = \frac{LSR'RS'L'}{m-1} = \frac{LSS'L'}{m-1}.$$

This should be compared to the eigenvector expansion of $\Sigma = VDV'$ where V is the matrix of orthonormal eigenvectors of Σ and D is the corresponding diagonal matrix of eigenvalues. As a result, one can identify the eigenspace V of Σ with the left singular space L of T. In addition, the matrix of eigenvalues $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ corresponds to the matrix

$$\frac{SS'}{m-1} = \frac{\text{diag}(\sigma_1^2, \ldots, \sigma_n^2)}{m-1}.$$

If one does not ensure that $E(T) = 0$ then at least the first component of PCA and SVD indicate different primary directions. Figure 1 illustrates this behaviour where $T = \langle x, 3 - x + y \rangle$ with x uniformly distributed on the interval $[0, 3]$ and y uniformly distributed on $[-1/2, 1/2]$. In this case the first component of the PCA points in the direction corresponding to the maximum variance of the data cluster, $\langle -1, 1 \rangle / \sqrt{2}$, whereas the first component of the SVD points in the direction of the centroid of the cluster, $\langle 1, 1 \rangle / \sqrt{2}$.

4 Difficulties with latent variables There are two disadvantages of these initial models when one considers them with respect to the data mine. Firstly, these models typically indicate that while there may be only a few principal directions, these directions may have significant weight in many of their components. This corresponds to the situation where one should ask large numbers of questions to determine which cluster to assign to a given individual. In essence, this analysis does not provide a natural way to determine which of the items is the *best* indicator (or the *best few* indicators).

Secondly, one must deal with the diverse collection of data in the data mine. Responses range from binary information about the ethnicity of an individual to continuous data regarding the market value of their dwelling. These two problems suggest that a robust algorithm is needed to gain a foothold on the structure of the mine before a more sophisticated latent variable analysis is attempted.

5 Determining the best question(s)

5.1 Factor analysis as a first look at the mine Factor analysis is the means by which we find the covariance relationship among many variables in terms of a few unobservable (or latent) variables. For example, if someone owns a Porsche, we would suspect that the person also has a high-paying job, lives in an upper class neighbourhood, has a six-figure stock portfolio and dines at high-class restaurants on a regular basis. If we were to label a latent variable that encompasses these four variables, we could label it *quality of life*.

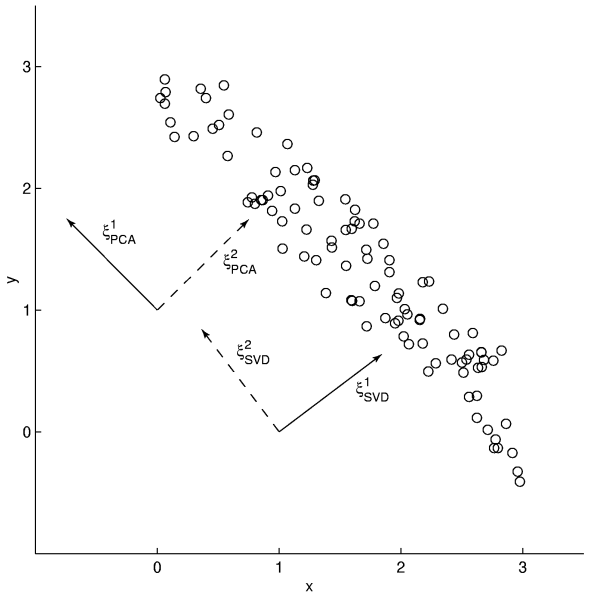


FIGURE 1: Illustrated are the components of a two factor PCA and SVD analysis for a randomly generated set of $m = 100$ points $(x, 3 - x + y)$ where x is uniformly distributed on the interval $[0, 3]$ and y uniformly distributed on $[-0.5, 0.5]$. Since $E(x) = E(y) = 3/2 \neq 0$, the PCA and SVD analysis yield a different principal direction. For this simulation, $\xi^1_{\text{PCA}} = \langle -0.7043, 0.7099 \rangle$ and $\xi^1_{\text{SVD}} = \langle 0.7978, 0.6029 \rangle$. The other complimentary components are $\xi^2_{\text{PCA}} = \langle 0.7099, 0.7043 \rangle$ and $\xi^2_{\text{SVD}} = \langle -0.6029, 0.7978 \rangle$.

In conducting a factor analysis, the basic model as discussed in Section 2 is:

$$t = \Omega x + \mu + \epsilon. \tag{5}$$

t is the observed random vector at n levels with a corresponding mean vector μ so that μ_i is the expected value of t_i. The vector x consists of the common factors at $k < n$ levels and the $n \times k$ matrix Ω is a matrix of coefficients otherwise known as the factor loadings. The element Ω_{ij} is referred to as the loading of the i-th variable on the j-th factor. In the above example, $n = 4$ and $k = 1$.

The model (5) can be rewritten as

$$t - \mu = \Omega x + \epsilon \tag{6}$$

and in order to have an orthogonal factor model, the following assumptions are made:

- x and ϵ are independent so that $\text{Cov}(x, \epsilon) = 0$,
- $E(x) = 0$ and $\text{Cov}(x) = I$,
- $E(\epsilon) = 0$ and $\text{Cov}(\epsilon) = \Gamma$ where $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_n)$.

Based on these assumptions, equation (6) yields an expression for the covariance of t,

$$\text{Cov}(t) = \Omega\Omega' + \Gamma.$$

In particular for the variable t_i one has, $\sigma_{ii}^2 = l_{i1}^2 + \cdots + l_{ik}^2 + \gamma_i$ where $l_{ij} = \text{Cov}(t_i, x_j)$. The sum of l_{i1}^2 through l_{ik}^2 is called the i-th commonality while γ_i is the unique variance.

From here we can determine which common factors contribute the most to the total variability in t_i. The ultimate objective is to be able to group the factor loadings for any one factor and attach some type of label to them as we did with the Porsche example. In particular, we are interested in the loadings which carry a significant amount of the weight.

There are two main methods for estimating the factor loadings: principal component and maximum likelihood. The former uses the eigenvalue/eigenvector pairs of the sample correlation matrix in order to construct Ω. If x and ϵ can be assumed to be normally distributed, then maximum likelihood methods can be used to estimate the covariance matrix of t and thus $\Omega\Omega' + \Gamma$.

In addition, if the initial factor loadings cannot be easily interpreted, various factor rotation methods exist to aid interpretation. The most common method used is the Varimax method which seeks to spread out the squares of the loadings on each factor as much as possible so that the factor loadings can be grouped more easily. It should be noted that in recent years, Bayesian factor analysis has arisen [**6**]. One of the features of the Bayesian approach is the elimination of the need to rotate factors. Bayesian factor analysis was not attempted in the analysis of the data mine.

5.2 Ranked differences of means To deal with the eclectic data in the mine, the most direct method of determining which questions seem to reflect the choice of an individuals product preference is to compute the observed difference in means across the given factors. Moreover, this is easily accomplished when there is a simple choice between two products as in the case studies that follow. The idea can also be generalized to the case when there are multiple products.

If only one product is under consideration the mine Ω is split into two groups, those respondents that have the product and those respondents that do not have the product. Denote these two groups as Ω_1 and Ω_2 consisting

of m_1 and $m_2 = m - m_1$ rows (respondents). For each of the n question responses, one computes the test statistic

$$z_j = \frac{\overline{x}_{1j} - \overline{x}_{2j}}{\sqrt{\frac{s_{1j}^2}{m_1} + \frac{s_{2j}^2}{m_2}}}, \quad j = 1, \ldots, n.$$

Using this statistic to classify the data is based on the need to find factors which are important to all of the respondents, yet at the same time differentiate the two groups. In terms of the solution to the normal equations, those factors with a large value of z_j correspond to coefficients which are large for both groups, but the *peaks* in the graphs of the coefficients occur at significantly different locations. It is precisely this failure to differentiate between the two groups that demonstrates why latent variable methods do not work easily as a first step.

Due to the large number of samples ($m \gg 50$), the z_j are each approximately normally distributed with zero mean and variance one under the null hypothesis that there is no difference in means between the two groups. Ordering the test statistics from the most negative to the most positive induces a reordering of the questions. In this sense, one can rank the indicators as to their ability to differentiate the two populations with respect to a given product. The factor with the largest observed values of z_j define the starting points of the cluster analysis and because of this, these particular questions form the first steps into the data mine when Ω is viewed in the light of a given product.

Typically the number of questions, n, can be large, and one is likely to find some means which will appear significantly different even when no difference in means exists. To analyse this situation let $\mathcal{U}$ be the number of the of questions with a test statistic that lies in the interval $(-s, s)$. If we assume for simplicity that the questions are independent then $\mathcal{U} \sim \text{Bin}(n, p)$ where $p = 2(1 - \Phi(s))$ and $\Phi(s)$ is the normal cdf. Therefore the probability that $\mathcal{U} \geq u$ and the expected value of $\mathcal{U}$ are

$$P(\mathcal{U} \geq u) = 1 - \sum_{j=0}^{u-1} \binom{n}{j} p^j (1-p)^{n-j}, \quad E(\mathcal{U}) = np.$$

The case studies at the end of this report use $n = 53$ data factors compiled from census data. For this many factors and $s = 3$ standard deviations, one finds that $P(\mathcal{U} \geq 1) = 0.1289$ and $E(\mathcal{U}) = 0.1378$. Consequently, to eliminate any *false alarm* differences we have chosen to consider statistically significantly differences at three rather than the common two standard deviations from the mean under the null hypothesis.

6 Consumer based clustering A simple definition of classification or clustering is using a metric or a set of rules which groups the data, and is also used to classify future data. For example, medical diseases may be classified by the manifesting symptoms which in turn describe each class or subclass of a given disease. In data classification one develops a description or model for each class in a database, based on the features present in a set of class-labelled training data. There have been many data classification methods studied, including decision-tree methods, such as C4.5 algorithm, ID3 algorithm, and SLIQ algorithm, statistical methods, neural networks, rough sets, nearest neighbour method, database-oriented methods, parallel algorithms, etc. The method for classification is in general application dependent, based on the goal of mining the data.

In this paper we have chosen a relatively simple metric to determine the clustering of the data, in particular, correlations between data columns corresponding to the different questions. The choice of the metric is based on the underlying goal that the salesperson has the opportunity to learn about a customer's preferences by asking only a few questions. This metric of clustering then indicates which are the most informative data that one would like to infer from these few questions. This metric is most similar to a nearest-neighbour type rule, where two of the census data are *near* when they are strongly positively correlated.

Another reason for looking at this metric is that it is computationally efficient. In order to look for more complicated classification structures, one could consider classification-rule learning which requires finding rules or decision trees that partition the given data into predefined classes. Of course, there many possible such decision trees; for any realistic problem domain of the classification-rule learning, the set of possible decision trees is too large to be searched exhaustively. In fact, the computational complexity of finding an optimal classification decision tree is NP hard.

Therefore, we have not attempted to find an optimal decision tree; rather, we have shown that the correlations give a fast classification of the mine, which can be readily used in designing questions and conversations with customers.

7 Case study A: BMW/Honda The first case study considered BMW and Honda owners. Given the census data on BMW and Honda owners grouped by postal code the goal is twofold:

Select a *few* questions to ask prospective buyers to infer their BMW/Honda preference.
Based upon the indicated preference, infer other information about the consumer.

For the following analysis there are a total of $m = 1995$ respondents which are partitioned into $m_h = 1782$ Honda owners and $m_b = 213$ BMW owners. Corresponding to each of these groups are $n = 53$ census data factors. Those portions of the data mine corresponding to Honda and BMW owners are referred to as Ω_h and Ω_b respectively. As a starting point, we compute a PCA on Ω_h and Ω_b.

7.1 PCA: BMW/Honda The eigenvalue structure of Σ_h and Σ_b, the covariance matrices of Ω_h and Ω_b, are virtually identical, ranging from $\lambda_1 \simeq 2.6 \times 10^{10}$ to $\lambda_{53} \simeq 6.7 \times 10^{-12}$. Figure 2(a) illustrates the logarithm of the magnitude of the $\{\lambda_j\}$. What is apparent from the illustration is that λ_1-λ_4 account for much of the variation in the mine. In fact

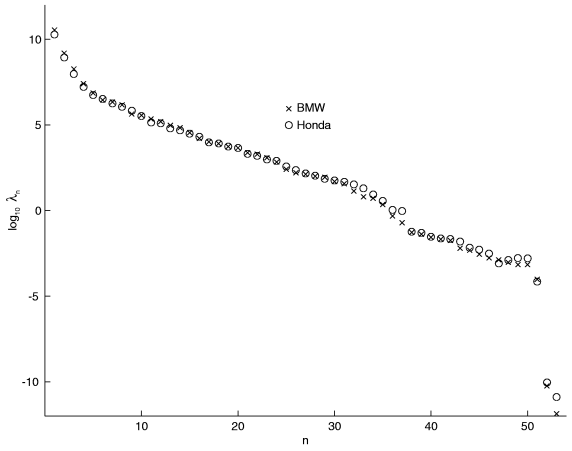
$$\frac{\sum_{j=1}^{4} \lambda_j}{\sum_{j=1}^{53} \lambda_j} = 0.9996.$$

Corresponding to these eigenvalues are eigenvectors focused in the direction of factors 21 to 24. These questions correspond to the average home value, average family income, average household income and total household expenditure. At the other end of the spectrum, λ_{52} and λ_{53} have eigenvectors that identify a strong correlation between factors 32 to 35. These latter indicators correspond to the average amount spent on public transportation, average spent on streetcars and buses, average spent on taxis and average spent on airplanes.

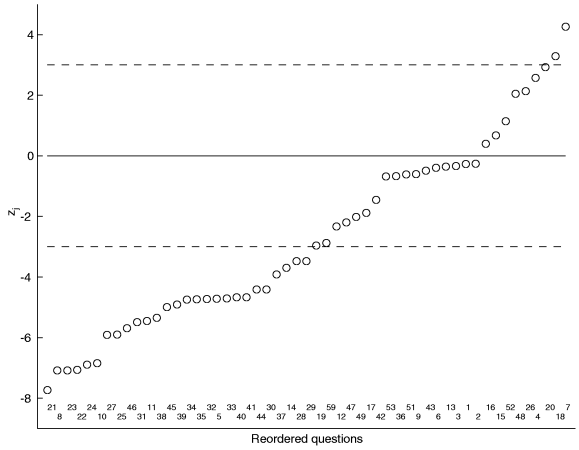
This preliminary analysis indicates that questions that reflect the total income and expenditure of a particular household should be good indicators of whether or not an individual owns a BMW or Honda. In addition, there is a certain amount of redundant information in the data mine with respect to public transportation. The main difficulty with PCA remains in that it does not identify a single question that best identifies BMW owners over Honda owners. Some headway can be made by computing a factor analysis which is attempted next.

7.2 A preliminary factor analysis In order to conduct a factor analysis, the data was again split into Ω_b and Ω_h according to whether the vehicle owned by the respondent was a BMW or Honda. Both principal component and maximum likelihood methods were used with three factors. However, the principal component method accounted for more of the variability than the maximum likelihood method. Varimax rotation was used in both methods. Table 1 summarizes the analysis.

As can be seen from the results, the same variables contributed to whether a person would own a BMW or a Honda with some variations. For example, under Factor 3, average home value contributes more to a person being a BMW



(a)



(b)

FIGURE 2: (a) Depicted are the eigenvalues for Σ_h and Σ_b, the covariance matrices for the Ω_h and Ω_b subsets respectively. The similar spectral structure for the BMW and Honda covariance typifies the difficulty encountered when attempting to find differences between these two groups. (b) Displayed is the ranked test statistic for the difference of means for each of the 53 factors. The dashed lines indicate the level of three standard deviations and the reordering of the factors is indicated at the base of the plot.

Factor 1	Loading values	
	BMW	Honda
Total adult population	0.996	0.997
Total population	0.995	0.997
Total number of households	0.995	0.991
Total adult labour force	0.994	0.995
Total number of families	0.993	0.996
Total number of dwelling units	0.993	0.989

Factor 2	Loading values	
	BMW	Honda
% homeowners	0.922	0.939
% single-detached house	0.861	0.772
Average owners' major payments	0.794	0.859
% home renters	-0.924	-0.928
% apartment with ≥ 5 floors	-0.746	-0.806
Average gross rent	-0.698	-0.759

Factor 3	Loading values	
	BMW	Honda
Ave. home value	0.749	0.579
% self-empl. inc.	0.735	0.580
% univ. degree	0.694	0.714

Factor	Variability explained	
	BMW	Honda
Factor 1	28.8%	28.5%
Factor 2	20.7%	22.3%
Factor 3	11.2%	8.5%
Total	60.7%	59.3%

TABLE 1: Listed are the three factors identified in the BMW/Honda data sample and the corresponding loadings. The final table shows that these three factors account for approximately 60% of the observed variability.

owner than a Honda owner. The percentage of variability in vehicle ownership explained by the complete model is approximately the same for both: 60.7% for BMW and 59.3% for Honda. Further analysis tools, such as discriminant analysis or tree regression can be used to determine which of these variables distinguish between BMW and Honda owners. The main conclusion is that the principal factors are strongly positively correlated and the anti-correlated components are small.

The factor analysis identifies a block of questions that differentiates the

two groups. Further identification is possible by considering the difference of means across the 53 factors.

7.3 Difference of means: BMW/Honda Figure 2(b) shows the ordered test statistics for each of the $n = 53$ factors. Detailed explanations for all of the census data can be found in the appendix at the end of this report. This ordering induces a reordering of the factors to $\{21, 8, 23, 22, 24, 10, \ldots, 18, 7\}$, with factor 21 having the most negative and question 7 having the most positive test statistic. This analysis also indicates that any questions related to 21, 8, 23, 22, 24, 10 are equally efficient at identifying BMW owners while factor 7 can be used to identify Honda owners. Factors 21-24 correspond to average home value, average annual family income, average annual household income, and annual household total expenditure, 8 reflects the percentage of individuals in a dwelling with a university education, 10 indicates self employment, 7 indicates the percentage of those subjects in a dwelling with only up to grade nine education and question 18 identifies those individuals living in dwellings with more than five stories. These initially identified factors can now used as starting points for a cluster analysis. Notice that many of these data items appear in the preliminary factor analysis.

Being able to identify a particular individual as a BMW or Honda owner is an important factor for the cluster analysis that follows. To differentiate we choose question 21, the average home value. We can use the data mine to determine the particular house value that should be used as a cutoff value to correctly identify the maximum number of individuals. That is, determine x such that $P(H < x \text{ and } B > x)$ is a maximum where H and B are the responses to question 21 for the Honda and BMW owners. Figure 3(a) shows that this probability has a maximum of 0.41 for x chosen in the interval ($230K, $240K). This procedure of choosing an optimal cutoff value from the probability structure encoded in the mine can be repeated for other questions to increase the differentiating power.

Detected differences in the mean response can be quite subtle. As an illustration of this, Figure 3(b) contrasts the probability distributions of the response to question 21 and question 7 for the two groups. For the cluster analysis that follows, the first six, 21, 8, 23, 22, 24, 10, and the last six factors, 48, 26, 4, 20, 18, 7, are used to define the initial clusters. By doing this it is hoped that the cluster analysis will be able to identify sequences of questions that link the Honda group to the BMW group. This in turn may help identify characteristics of prospective BMW owners.

7.4 Cluster analysis: BMW/Honda A cluster analysis was performed for two cases with correlations at the 60% and 75% level indicated. The first anal-

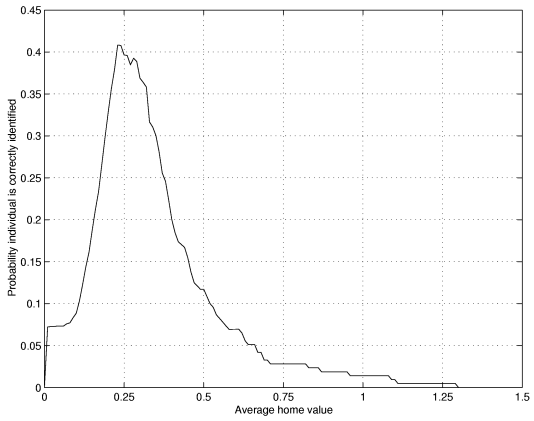
ysis was performed by considering only the BMW owners while the second analysis considered the complete data mine. As the number of Honda owners was much larger than the number of BMW owners, a cluster analysis of only Honda owners matches that obtained when using the complete data mine. Figure 4 summarizes the results.

What is immediately apparent is the greater resolution one can achieve in the data mine with the BMW group. On the left hand side of each cluster diagram are those questions identified with the most negative test statistic (+ BMW) and on the right are those questions corresponding to the most positive test statistic (− BMW). Those traits that identify BMW owners are household value and income, university education and self employment. Characteristics that directly stem from these traits are donations to charity, amount spent on public transportation and amount spent on personal care. From the other end of the data mine, individuals that do not own a BMW are characterized as either renters or having less than a grade nine education. A link between the renters and those with expensive homes is the number of cars per household and the subsequent expenditure on tires, gasoline, food and transportation.

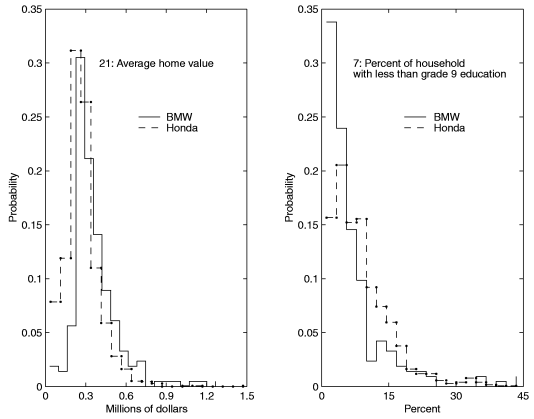
This cluster analysis implies that there are many possible ways to identify a BMW owner. For example, university educated individuals that do not rent and outwardly appear to spend a great deal on personal care. Once identified, the characteristics of this group could be targeted for a broad range of products or services that lie within the identified common interests. Examples of these interests include cosmetics, expensive tires, and perhaps even endowments to universities.

When we consider the entire mine there is a loss of resolution but much of the structure remains. In addition, other characteristics come to the forefront. Two new characteristics are a stronger correlation with the amount spent on computers the loss of the correlation with public transportation. A possible implication here is that Honda owners with an expensive home may be differentiated from BMW owners by the amount that they spend on airlines. Again we point out that being able to accurately classify an individual is an important first step in that the clustering reflects this bias. However, mis-identifying an individual does not have as serious a consequence as one might first expect. The clustering analysis supports this by illustrating that much of the structure is preserved when moving from BMW to Honda owners. To contrast with the BMW/Honda data, the second case study considers consumer preference of two brands of domestic beer.

8 Case study B: beer preference Our second case study addresses beer preferences amongst a sample of 707 individuals. Each individual was asked to



(a)



(b)

FIGURE 3: (a) Probability of correctly identifying Honda and BMW simultaneously for a given known home value. This distribution has an extreme value of 0.41 for the interval ($230K, $240K). At this cutoff value the probability of correctly identifying a Honda owner is 999/1782 (56%) and that of identifying the BMW owner is 156/213 (73%). (b) On the left is the probability distribution of responses to question 21 (average home value). To the right is the probability distribution to factor 7 (percentage of household with less than a grade nine education). These factors yield the most negative and most positive values of z_j respectively.

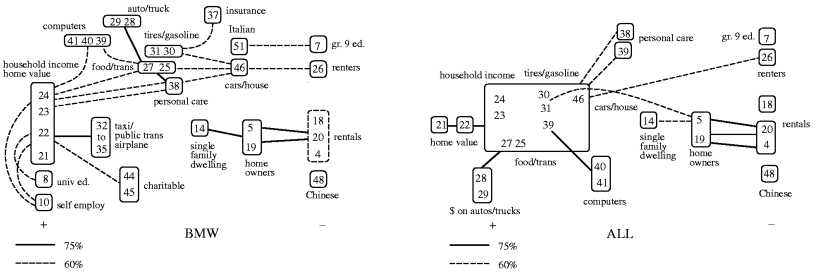


FIGURE 4: To the left are the data clusters considering only the BMW group while to the right is the the same analysis considering the complete data mine. On the $+$ side of each figure, the factors are larger or more likely for BMW owners, while the $-$ side of each figure the factors are smaller or less likely for BMW ownership. Cutoffs at 60% and 75% in the correlation level (either positively or negatively) are indicated. Explanations for all of the data factors can be found at the end of the report.

indicate their preference for two different brands of beer (Brand A and Brand B) according to the four point scale:

0: Don't drink
1: Tried in the past 12 months
2: Becoming usual
3: Usual brand.

As no respondents indicated that either brand was their *usual* brand, the responses broke into nine separate classifications. Figure 5 shows the resulting tree structure and the four groups into which the individuals were placed. Group I essentially consists of non-drinkers, group II and III tend to prefer brands A and B respectively, and group IV respondents strongly prefer both brands.

8.1 Difference of means: beer No significant differences were detected in a direct comparison of groups II and III since all of the z_j statistics were located within two standard deviations. Large scores were detected when comparing groups I and IV but since group IV consisted of only five individuals our underlying assumptions of normality were no longer valid. Since the 53 characteristics do not seem to be able to clearly differentiate the two brands of beer, it was more appropriate with this data set to compare drinkers of both brands versus those individuals that do not drink either brand. As such, groups II, III and IV were consolidated into a single group which was then compared

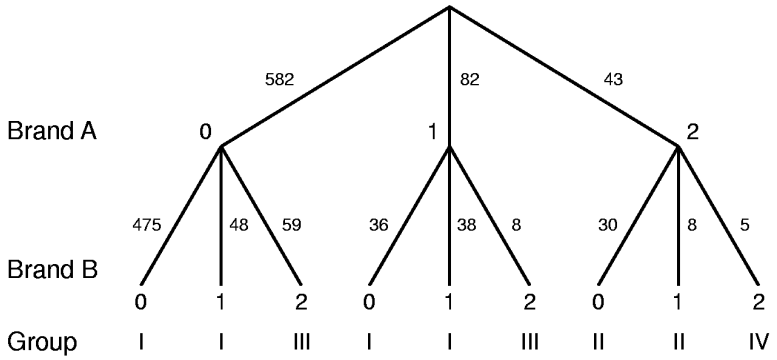


FIGURE 5: The classification tree structure for the beer preference respondents.

to group I. The portions of the data mine concerning these two groups will be referred to as Ω_b and Ω_1. Figure 6 illustrates the spectral structure of these two classifications and distribution of the difference of means.

Comparing Figure 6(a) with Figure 2(a) illustrates a striking similarity with the eigenvalue distribution of the beer data and the car data of the previous study. This similarity is also reflected in the preliminary factor analysis which has been omitted because of its similarity to the BMW/Honda analysis. Even though the eigenvalue structure was similar, none of the test statistics lie outside of the three standard deviations. Despite this, we begin the cluster analysis starting with factors 50, 7, 48 on the *drink beer* side of the data mine and factors 27 and 14 on the *don't drink beer* side of the mine.

8.2 Cluster analysis: beer preference We again remind the reader that a full explanation of each of the factors from the census data can be found in the appendix. Correlations between the starting factors 7, 14, 27, 48, 50 and the remaining questions were detected once the cutoff level was dropped to 50%. This reduction in the cutoff level was expected given the lack of significant differences detected in the previous section. Figure 7 summarizes the results and illustrates that respondents that prefer these brands seem to fall along ethnic lines. The analysis also indicates that beer drinkers are characterized by individuals that rent rather than living in a single detached house. However, with this collection of 53 factors there was no additional product information that could be correlated with these individuals.

Without a clear indication of questions that differentiate between beer drinkers and non beer drinkers, at least for these two brands of beers, we do not attempt to correlate other characteristics. Clearly, some additional analysis is

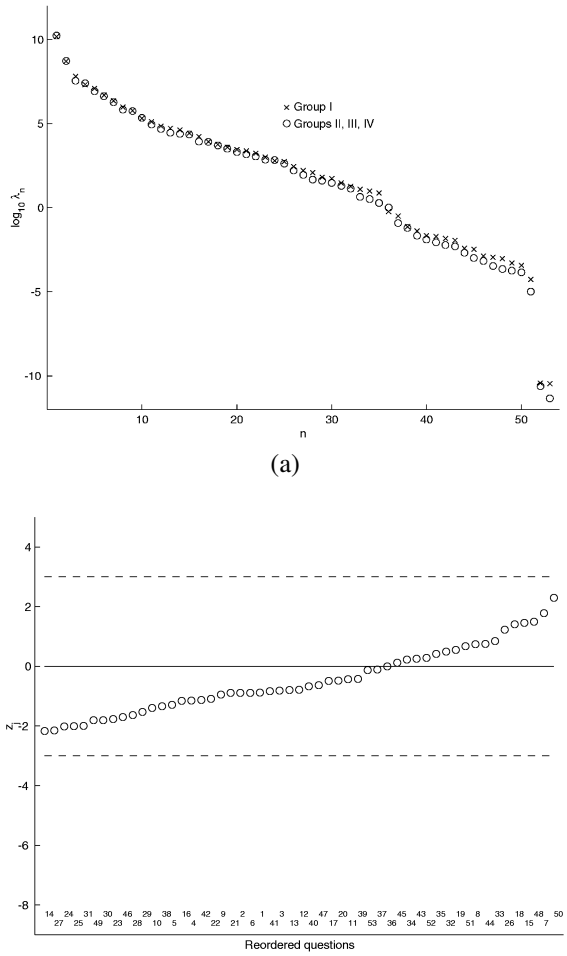


FIGURE 6: (a) Eigenvalues for Σ_b and Σ_1, the covariance matrices for the Ω_b and Ω_1 subsets respectively. (b) Ranked test statistic for the difference of means for each of the 53 factors from the census data . As in Figure 2(b), the dashed lines indicate the level of three standard deviations and the reordering of the factors is indicated at the base of the plot.

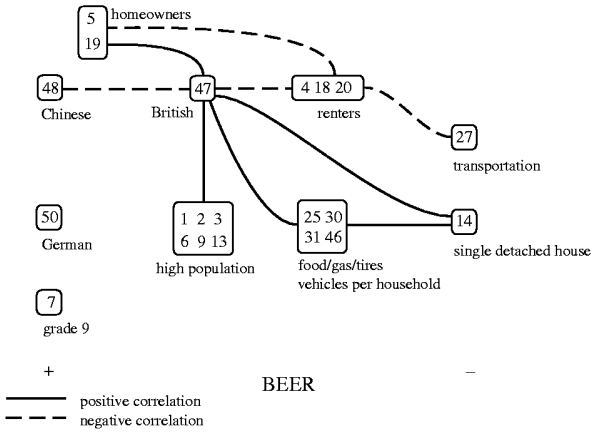


FIGURE 7: Illustrated is the cluster analysis for beer drinkers. The correlation cutoff was set at 50%. Solid lines represent positive correlations and dashed lines represent negative correlations. On the left of the are those indicators whose mean response for beer drinkers was higher than for non beer drinkers.

necessary to improve the first part of the analysis, namely, making distinctions according to a particular product preference. Since the difference in means is not significantly large for any one factor, this suggests that a combination of questions would be necessary to make a significant distinction. As mentioned above, a factor analysis has not been done for this data set; however, one could use this analysis to design a combination of a few questions as a first algorithm for differentiating between consumers.

9 Conclusion For a given product two tasks were required. The first being identification of consumers that would react favourably to product and the second being the inference of other characteristics concerning these consumers.

This was accomplished with a twofold strategy. By ranking the difference of means across all of the factors, those questions that best characterize favourable consumers can be identified. Once identified, the data mine can be used to estimate the power of a given strategy to correctly identify a given individual. For case study A the identification algorithm was simply to use an individuals home value. By optimizing the cutoff level this single variable was able to correctly identify 41% of the individuals in the data mine. By using

a combination of questions this percentage could be increased. Performing a cluster analysis that is rooted at these key identifying questions allows other characteristics of these consumers to be inferred.

The two case studies show that being able to identify questions that significantly differentiate respondents with respect to a given product is a fundamental part of the process. Failure to make this identification decreases the resolution of the subsequent cluster analysis. Case study A exemplifies the situation when there is a clear separation with respect to a product whereas case study B illustrates the decrease in resolution when no clear separation exists. In general, this first step may be dependent on the type of data and the desired differentiations. A combination of factor analysis and the consideration of differences in basic test statistics proved to be superior to methods based on latent variables or principal components, due to the underlying eigenstructure of the data mine.

To increase the capability of this method future advances should include a more sophisticated clustering algorithm. For example, PLS/SVD could be used on the clustering subgroups after the first step of separating with the difference of the means statistic. An addition, automatic determination of the identification power for a given set of identifying questions should also be addressed.

Appendix: Factors from Census Data

Question	Description	Mnemonic
01	Total population	PP-TOT
02	Total number of families	FM-TOT
	Household	
03	Total number of households	HH-TOT
04	Average gross rent	HH-TOTRENT
05	Average owner's major payments	HH-TOTMAPJ
	Education	
06	Total population 15 years old and over	ED-HL
07	Percent education level: less than grade 9	ED-GR-9
08	Percent education level: university with bachelor's degree or higher	ED-UNIDG
	Employment	
09	Total labour force 15 years old and over	EM-TOT
10	Percent employment: self-employed (incorporated)	EM-PSMI
11	Percent employment: self-employed (unincorporated)	EM-PSMU
12	Percent employment: unpaid family workers	EM-UP
	Dwelling	
13	Total number of dwelling units	DM-TOT
14	Percent: dwelling type: single-detached house	DW-SINGLE
15	Percent: dwelling: semi-detached house	DW-SEMI
16	Percent: dwelling type: town house	DW-ROW
17	Percent: dwelling type: apartment, detached duplex	DW-DUP

18	Percent: dwelling: apartment building, five or more storeys	DW-APT5
19	Percent: homeowners	DW-OWNED
20	Percent: home renters	DW-RENTED
21	Average home value	DW-TVALUE
	Income	
22	Annual average family income	IN-AFM
23	Annual average household income	IN-AHH
	Expenditures	
24	Annual household total expenditure	D1000-5230
25	Annual expenditure on food	D1000-1560
26	Annual expenditure on rent	D2000
27	Annual expenditure on transportation	D3000-3260
28	Annual expenditure on purchase of automobiles and trucks	D3000-3004
29	Annual expenditure on automobiles	D3000
30	Annual expenditure on gasoline and other fuels	D3050
31	Annual expenditure on tires, batteries, parts and supplies	D3060
32	Annual expenditure on bus, subway, street car and train	D3200
33	Annual expenditure on public transportation	D3200-3260
34	Annual expenditure on taxi	D3210
35	Annual expenditure on airplane	D3220
36	Annual expenditure on moving, storage and delivery services	D3260
37	Annual expenditure on accident and disability insurance	D3384
38	Annual expenditure on personal care	D3500-3580
39	Annual expenditure on recreation equipment and services	D3700-3830
40	Annual expenditure on computer hardware	D3750-3752
41	Annual expenditure on computer software	D3755
42	Annual expenditure on gifts of money and contributions	D5200-5230
43	Annual expenditure on gifts to persons living outside Canada	D5210
44	Annual expenditure on contributions to charity	D5220-5230
45	Annual expenditure on non-religious charitable organizations	D5230
46	Average number of vehicles owned per household	NMVEHONP
	Ethnicity	
47	Percent ethnicity : British	BRITISH
48	Percent ethnicity : Chinese	CHINESE
49	Percent ethnicity : Dutch	DUTCH
50	Percent ethnicity : German	GERMAN
51	Percent ethnicity : Italian	ITALIAN
52	Percent ethnicity : Polish	POLISH
53	Percent ethnicity : Scandinavian	SCANDINAV

REFERENCES

1. D. W. Aha, *Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms*. International Journal of Man-Machine Studies **36** (1992), 267–287.
2. C. M. Bishop, *Latent variable models*: In Learning and Graphical Models, M.I. Jordan (Ed.). MIT Press, 1999, 371–403.
3. R. L. Burden and J. D. Faires, *Numerical Analysis*, PWF-Kent Publishing, 1989.
4. B. V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, Washington: IEEE Computer Society Press, 1990.
5. A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley, 2001.

6. Z. Ghahramani and M. J. Beal, *Variational inference for Bayesian mixture of factor analysers*: In Advances in Neural Information Processing Systems 12, (S. A. Solla, T. K. Leen and K. R. Müller, eds.) MIT Press, 2000.
7. K. P. Joshi, *Analysis of Data Mining Algorithms*, 1997, `http://userpages.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm`.
8. G. Strang, *Introduction to Linear Algebra*, Wellesey-Cambridge Press, 1993.
9. D. Wettschereck, *A Hybrid Nearest-Neighbor and Nearest-Hyperrectangle Algorithm*: In Lecture Notes in Artificial Intelligence **784**, Springer-Verlag, 1994, 323–335.

CORRESPONDING AUTHOR:

C. SEAN BOHUN

MATHEMATICS AND STATISTICS, PENNSYLVANIA STATE UNIVERSITY, ONE CAMPUS DRIVE,

MONT ALTO, PA, USA 17237

E-mail address: csb15@psu.edu