Bensmail, Halima; DeGennaro, Ramon P.

Working Paper

Analyzing imputed financial data: a new approach to cluster analysis

Working Paper, No. 2004-20

Provided in Cooperation with:
Federal Reserve Bank of Atlanta

Suggested Citation: Bensmail, Halima; DeGennaro, Ramon P. (2004) : Analyzing imputed financial data: a new approach to cluster analysis, Working Paper, No. 2004-20, Federal Reserve Bank of Atlanta, Atlanta, GA

FEDERAL RESERVE BANK of ATLANTA

Analyzing Imputed Financial Data: A New Approach to Cluster Analysis

Halima Bensmail and Ramon P. DeGennaro

Working Paper 2004-20
September 2004

WORKING PAPER SERIES

FEDERAL RESERVE BANK *of* ATLANTA WORKING PAPER SERIES

Analyzing Imputed Financial Data: A New Approach to Cluster Analysis

Halima Bensmail and Ramon P. DeGennaro

Working Paper 2004-20
September 2004

Abstract: The authors introduce a novel statistical modeling technique to cluster analysis and apply it to financial data. Their two main goals are to handle missing data and to find homogeneous groups within the data. Their approach is flexible and handles large and complex data structures with missing observations and with quantitative and qualitative measurements. The authors achieve this result by mapping the data to a new structure that is free of distributional assumptions in choosing homogeneous groups of observations. Their new method also provides insight into the number of different categories needed for classifying the data. The authors use this approach to partition a matched sample of stocks. One group offers dividend reinvestment plans, and the other does not. Their method partitions this sample with almost 97 percent accuracy even when using only easily available financial variables. One interpretation of their result is that the misclassified companies are the best candidates either to adopt a dividend reinvestment plan (if they have none) or to abandon one (if they currently offer one). The authors offer other suggestions for applications in the field of finance.

JEL classification: G20, G29, G35

Key words: dividend reinvestment, Bayesian analysis, Gibbs sampler, clustering

---

The authors gratefully acknowledge the support of a University of Tennessee Finance Department Summer Faculty Research Award and a College of Business Scholarly Research Grant. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to Halima Bensmail, Statistics Department, the University of Tennessee, Knoxville, Tennessee 37996, 865-974-8325, bensmail@utk.edu and Ramon P. DeGennaro, SunTrust Professor of Finance, The University of Tennessee, Knoxville, Tennessee 37996, 865-974-3216, and Visiting Scholar, Federal Reserve Bank of Atlanta, 1000 Peachtree Street, N.E., Atlanta, Georgia 30309-4470, rdegenna@utk.edu.

Analyzing Imputed Financial Data: A New Approach to Cluster Analysis

1. Introduction

We introduce and apply a novel statistical approach to cluster analysis for financial data in this paper. We have two main goals. First, we wish to handle cases in which a subset of variables is missing for some observations. Second, we wish to find homogeneous groups within the data. Put differently, we want to determine the most likely number of categories comprising the data, and to assign observations to those categories optimally. Our approach is flexible in that it handles large and complex data structures with missing observations and with both quantitative and qualitative measurements. We achieve this by mapping the data to a new structure that is free of distributional assumptions in choosing homogeneous groups of observations. For example, when processing credit card transactions of customers, a company may want to explore the possibility of encouraging different or additional transactions by those customers. In this case, the task is to find homogeneous transactions and to forecast the willingness of a new customer to use the credit card to make a different or additional transaction, even if the data are not continuous and even if there are missing data. Our new method also provides the researcher with insight into the number of different categories needed for classifying the data.

Classification methods have a long history of productive uses in business and finance. Perhaps the most common are discrete choice models. Among these, the multinomial logit approach has been used at least as far back as Holman and Marley (in Luce and Suppes, 1965). McFadden (1978) introduced the Generalized Extreme Value model in his study of residential location, and Koppelman and Wen (1997) have recently developed newer variations. The nested logit model of Ben-Akiva (1973) is designed to handle correlations among alternatives. Yet

another variation of multinomial logic has been developed or used by Bierlaire, Lotan and Toint (1997). More recently, Calhoun and Deng (2000) use multinomial logit models to study loan terminations.

Another form of discrete choice model is cluster analysis. Shaffer (1991) offers one example. He studies federal deposit insurance funding and considers its influence on taxpayers. Dalhstedt, Salmi, Luoma, and Laakkonen (1994) use cluster analysis to demonstrate that comparing financial ratios across firms is problematic. They argue that care is necessary even when the firms belong to the same official International Standard Industrial Classification category. von Altrock (1995) explains how fuzzy logic, a variation of cluster analysis, can be useful in practical business applications.

Methods that produce a continuous variable rather than a discrete choice can also be used as classification methods. For example, credit scoring uses information to produce a continuous variable called the credit score. Lending institutions overlay this continuous score with a grid, producing discrete categories. Applicants with a score below a certain point might be rejected automatically. Applicants above a specified higher point might be accepted automatically. Scores falling between these trigger points might be given further investigation. See Mester (1997) for an example. Altman (2000) follows a somewhat similar approach to update the popular method of zeta® analysis.

Related to the problem of classifying data is the issue of determining the number of categories. In addition, some methods that can determine the number of categories provide no evidence on which observations fall within each class. For example, Baillie and Bollerslev (1989) use cointegration methods to study the number of common stochastic trends in a system of exchange rates. In this case, it makes little economic sense to attempt to classify exchange

rates along some dimension. Instead, Baillie and Bollerslev calculate the number of common stochastic trends to gain insight regarding the extent of market efficiency and potentially profitable trading opportunities.

2. Clustering and Bayesian Data Augmentation

Cluster analysis has been developed mainly through the invention of empirical, and lately Bayesian study of ad hoc methods, in isolation from more formal statistical procedures. It has been found that basing cluster analysis on a probability model can be useful both for understanding when existing methods are likely to be successful and for suggesting new methods. For examples, see Hartigan (1975), Kaufman and Rousseeuw (1990), and Bensmail and Bozdogan (2002).

We assume that the population of interest consists of K different subpopulations $G1,\ldots GK$ and that the density of a p-dimensional observation $\boldsymbol{x}$ from the kth subpopulation is $f_k(\mathbf{y},\theta_k)$ for some unknown vector of parameters θ_k ($k = 1 \ldots K$). Given observations $\mathbf{y} = (y_1 \ldots y_k)$, we let $\boldsymbol{\nu} = (\boldsymbol{\nu}_1 \ldots \boldsymbol{\nu}_k)^t$ denote the unknown identifying labels, where $\boldsymbol{\nu}i = k$ if $\mathbf{y}i$ is from the kth subpopulation. Clustering data using a mixture distribution framework has been successful and many authors have proposed different approaches. A good source of references is Hosmer (1973). In most cases, the data to be classified are viewed as coming from a mixture of probability distributions (McLachlan and Basford, 1988), each representing a different cluster. Therefore, the likelihood function is:

$$p(\theta_1 \ldots \theta_K, \pi_1 \ldots \pi_K \mid \mathbf{x}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i \mid \theta_K), \qquad (1)$$

where π_k is the probability that an observation belongs to the k^{th} component ($\pi_k \geq 0; \Sigma^K_{i=1} = 1$).

Clustering data with missing values has always been difficult. In some cases, sample

means were used to replace the missing values and any of several clustering methods would be applied to the complete data. Recently, the EM algorithm (Dempster, Laird, and Rubin, 1977) has been used to overcome the limitations of the average and maximum likelihood estimators. Within the Bayesian framework, similar to the usual EM algorithm, a data-augmentation (DA) algorithm (Tanner and Wong; 1987) has been proposed.

Defining **y** as an observation from the sample, we denote an observed value as **y***obs* and a missing observation as **y***miss*. We want to estimate the parameter θ given in Equation (1) based on **y** = (**y***miss*,**y***obs*). We assume that the data (observed and missing) comprise *K* clusters; these clusters *G*1, . . , *GK* are unknown. Each observation ***yi*** belonging to a cluster *Gk* is a random variable drawn from a normal distribution $Np(\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})$, where $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ are the mean and covariance matrix of the cluster *Gk* such that

$$(\mathbf{y}i|\ \boldsymbol{\mu_k},\boldsymbol{\Sigma_k}) \sim \mathrm{N_p}(\boldsymbol{\mu_k},\boldsymbol{\Sigma_k}) \tag{2}$$

The distribution function of a sample $(y_1, \ldots, y_{nk})$ representing a subpopulation G_k is:

$$p(y_1,\ldots,y_{n_k} \mid \mu_k,\Sigma_k) \propto |\Sigma_k|^{-n_k/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n_k}(y_i-\mu_k)^t\sum_k^{-1}(y_i-\mu_k)\right)$$

$$= |\Sigma_k|^{-n_k/2} \exp\left(-\frac{1}{2} tr W_k \Sigma_k^{-1}\right) \tag{3}$$

where $n_k = \Sigma_{i\in G_k} I\{\upsilon_i = k\}$, $W_k = \Sigma_{i;\upsilon_i=k}(y_i-\bar{y}_k)(y_i-\bar{y}_k)^t$, $\bar{y}_k = \frac{1}{n_k}\Sigma_{i;\upsilon_i=k} y_i$.

In the Bayesian clustering approach, one needs to estimate the posterior distribution of the parameter θ involved given its prior distribution. When **y***miss* denotes a subvector of **y** containing the missing components, the posterior distribution of the parameter θ given the observed data **y***obs* is

$$f(\theta \mid y_{obs}) = \int f(\theta \mid y_{miss}, y_{obs}) f(y_{miss} \mid y_{obs}) dy_{miss} \tag{4}$$

Equation (4) is a mixture of the posterior distribution of θ given the data (observed and

missing) where the mixing proportion is given by the marginal conditional distribution of y_{miss}. This is typically very difficult to use; often it cannot even be expressed in a closed form.

The data augmentation (DA) algorithm is very useful for circumventing these difficulties. Data Augmentation refers to methods for constructing iterative algorithms via the introduction of unobserved data or latent variables. For deterministic algorithms, the method was popularized in the general statistical community by the seminal paper of Dempster, Laird, and Rubin (1977) on the EM algorithm for maximizing a likelihood function, or more generally, a posterior density. For stochastic algorithms, the method was popularized in the statistical literature by the Tanner and Wong (1987) Data Augmentation algorithm for posterior sampling. The Swendsen and Wang (1987) algorithm has been used in the physics literature. Data augmentation schemes were used by Tanner and Wong to make simulation feasible and simple, while Swendsen and Wang (1987) adopted auxiliary variables to improve the speed of iterative simulation. In general, however, constructing data augmentation schemes that result in both simple and fast algorithms is a matter of art, in that successful strategies vary greatly with the observed-data models being considered (Tierney, 1994).

We now describe the DA algorithm for imputing the missing data. The algorithm iterates as follows:

To go from iteration (t) to iteration (t+1) we do the following:

1. **I**-step: *imputation:* generate

$$y_{miss}^{t+1} \sim f(y_{miss} \mid y_{obs}, \theta^t) \tag{5}$$

2. **P**-step: *posterior estimation*: generate

$$\theta^{t+1} \sim f(\theta \mid y_{obs}, y_{miss}^t) \tag{6}$$

2.1 Imputation

To evaluate Equation (5) we use the following Lemma from Anderson (1984):

If **y** *is a random variable having a multivariate normal distribution, then the conditional distribution of any subvector of* **y** *given the remaining elements is once again multivariate normal. If we partition* **y** *into subvectors* $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, *then* $p(\mathbf{y}_1|\mathbf{y}_2)$ *is (multivariate) normal such that*

$$y_1 \mid y_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_2^{-1}(y_2 - \mu_2), \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}) \quad (7)$$

where

$$y_1 \sim N(\mu_1, \Sigma_1) \text{ and } y_2 \sim N(\mu_2, \Sigma_2) \quad (8)$$

and

$$\Sigma_{12} = \Sigma_{21} = \text{cov}(y_1, y_2) \quad (9)$$

Case 1: One missing value and many observed values

Suppose that an observation $\mathbf{y} = (y_1, y_2, \ldots, yp)$ has one missing value. Consider $z_1 = y_1$ the missing value and $\mathbf{z_2} = (y_2, \ldots, yp)$ the remaining observed values. Then the only information needed is part of the vector mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ and the covariance matrix Σ. Given the mean $\boldsymbol{\mu}$ and given the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{p1} & \ldots & \ldots & \sigma_{pp} \end{pmatrix}$$

the only input we need is the first row of the covariance matrix excluding the first variance term; i.e. the vector $\sigma_{1(k-1)} = (\sigma_{12}, \sigma_{13}, \ldots, \sigma_{1p})$ and the covariance matrix minus the first row and the first column. We denote this matrix as $\Sigma_{-(1,1)}$:

$$\Sigma_{-(1,1)} = \begin{pmatrix} \sigma_{22} & \sigma_{23} & \sigma_{2p} \\ \ldots & \ldots & \ldots \\ \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

We can then use these blocks to estimate the missing value $z_1 = y_1$ by generating the data from a normal distribution:

$$z_1 \mid z_2 \sim N\begin{pmatrix} \mu_1 + \sigma_{1k(-1)}\Sigma_{-(1,1)}(y_2 - \mu_2,..., y_p - \mu_p)' \\ \sigma_{11} - \sigma_{1k(-1)}\Sigma_{-(1,1)}\sigma^t_{1k(-1)} \end{pmatrix}$$

Case 2: The General Case

For the general case, we have multivariate data $y = (y_1, y_2, \ldots, y_p)$ where two y_j and y_h or more are missing. Using the same scheme as before, the only information needed is part of the vector mean μ and the covariance matrix Σ. Using Anderson's Lemma (1984), $(y_j, y_h)|(y_2 \ldots, y_p, \mu, \Sigma)$ is normally distributed with mean vector $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma}$. See Bensmail and Bozdogan (2003b) for details.

2.2 Posterior Estimation

To estimate the parameters μ and Σ we need to specify priors on those parameters. Here we use conjugate priors for the parameters π which is a Dirichlet distribution Dirichlet($\alpha_1 \ldots. \alpha_K$). Because the log-likelihood is a quadratic form in μ_k, the conjugate prior distribution of μ_k is given by:

$$\mu_k \mid \Sigma_k \sim N_p(\xi_k, \Sigma_k / \tau_k), \tag{10}$$

and a conjugate prior of Σk is given by:

$$\Sigma_k \sim W_p^{-1}(m_k, \Psi_k) \tag{11}$$

The posterior distribution of μ_k and Σ_k given the missing and observed data are then given by:

$$(\mu_k \mid y_{miss}, y_{obs}, \Sigma_k) \propto N_p[(\xi_k + n_k \overline{y_k})/(n_k + \tau_k), \Sigma_k /(n_k + \tau_k)] \tag{12}$$

and

$$(\Sigma_k \mid y_{miss}, y_{obs}, \upsilon) \sim W_p^{-1}((n_k + m_k, \psi_k + W_k \frac{n_k \tau_k}{n_k + \tau_k}(\overline{y_k} - \xi_k)(\overline{y_k} - \xi_k)') \tag{13}$$

2.3 Algorithm

We estimate the parameters by simulating from the joint posterior distribution of **y***miss*, π, θ, and υ using the Gibbs sampler (Smith and Roberts 1993, Bensmail et. al. 1997, Bensmail and Bozdogan 2003b). In our case this consists of the following steps:

1. Simulate the classification variables υ_i according to their posterior probabilities $t_{ik}, k = 1, \ldots, K)$ conditional on π, **y***miss* and θ, namely

$$t_{ik} = \frac{\pi_k f(y \mid \mu_K, \Sigma_K)}{\sum_{h=1}^{K} \pi_h f(y \mid \mu_h, \Sigma_h)}; \; i = 1, \ldots n \qquad (14)$$

2. Simulate the missing values **y***miss* given **y***obs* from

$$\mathbf{y}miss \sim f(\mathbf{y}miss \mid \mathbf{y}obs, \mu_\kappa, \Sigma_\kappa). \qquad (15)$$

3. Simulate the vector π of mixing proportions according to its posterior distribution conditional on the υ. This consists of simulating π from its conditional posterior distribution, namely

$$\pi \sim Dirichlet(\alpha_1 + \Sigma I\{\upsilon_i = 1\}, \ldots \alpha_k + \Sigma I\{\upsilon_i = K\}).$$

4. Simulate the parameters $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ according to their posterior distribution:

$$\Sigma_k^{t+1} \mid y_{obs}, y_{miss}^t \sim W^{-1}(n_k + m_k, \psi_k + W_k \frac{n_k \tau_k}{n_k + \tau_k}(\overline{y_k} - \xi_k)(\overline{y_k} - \xi_k)') \qquad (16)$$

$$\mu^{t+1} \mid y_{obs}, y_{miss} \Sigma^{t+1} \sim N((\xi_k + n_k \overline{y}_k)/(n_k + \tau_k), \Sigma_k/(n_k + \tau_k))$$

where $\overline{y_k}$ and **w***k* are the sample mean and variance matrix of the data, and w^{-1} denotes the inverse Wishart distribution.

3. Bayesian Model Selection for Choosing the Number of Clusters

Determining the number of clusters is usually the most important component of any cluster analysis. Many authors have investigated different criteria for model selection and choosing the number of clusters. Proposed methods include the Akaike Information Criteria (AIC) (Akaike 1973), the Information Complexity Criteria (ICOMP) (Bozdogan 1987), the Normalized Entropy of assessment (NEC) (Celeux and Soromenho 1996), and the Bayesian Information criterion (BSC) introduced by Schwarz (1978). The new work of Bensmail and

Bozdogan (2003 a,b) develops ICOMP in choosing the number of components in both multivariate kernel mixture-models and Bayesian kernel mixture-model cluster analysis of mixed and imputed data.

We use Schwarz's criterion. Although regularity conditions for this may not hold for mixture models, there is considerable theoretical and practical support for its use (Roeder and Wasserman, 1997; Dasgupta and Raftery, 1998).

BSC is defined as the maximum of:

$$\text{BSC}(\text{M}_\text{k}) = -2\log L(\tilde{\theta}_k, M_k) + m(k)\log(n_k) \qquad (17)$$

where $L(\tilde{\theta}, \text{M}_\text{k})$ is the likelihood of the posterior mode $\tilde{\theta}$ for the model *Mk* (here, the number of components or categories *k*), *m(k)* is the number of parameters to estimate and n_k is the size of the subpopulation G_k. Maximizing the BSC determines the optimal dimension of the model.

4. Example: Simulated data

We simulate a sample of 60 observations from a bivariate normal distribution with mean $\mu_1 = (0,2)$ and variance matrix $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ and 60 observations from a bivariate normal distribution with mean $\mu_2 = (0,0)$ and variance matrix $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. To illustrate the capability of our method to handle missing observations, we next set four values to missing: $y_{(10,1)}$, $y_{(15,2)}$, $y_{(101,1)}$, and $y_{(120,2)}$. We find in general that the Data Augmentation algorithm converges fast and is stable. Table 1 shows that the BSC criterion identifies two groups, which is the correct number for our simulated data.

Table 1

BSC for values of $k = 1 \ldots 4$

k	*SBC*
1	1811.10
2	1637.03
3	1752.29
4	1748.40

Table 2 gives the Confusion matrix. Our method correctly classifies 58 of the 60 observations from the first distribution and 56 of 60 from the second distribution. The overall accuracy rate is 95%.

Table 2

Confusion matrix

k	1	2	Total
1	58	2	60
2	4	56	60
Total	62 (cluster 1)	58 (cluster 2)	n=120

The plot of the mean vector μ_1 (variate wide) based on 5000 simulations is shown in Figure 1. Its estimated value is $\hat{\mu}_1 =$ [-0.3971, 2.0560], compared to the true values of $\mu_1 =$ [0, 2]. The convergence plot for the variance covariance matrix (variate wide) is shown in Figure 2 and the Bayesian estimate is $\hat{\Sigma}_1 = \begin{pmatrix} 2.1 & -1.1 \\ -1.1 & 1.3 \end{pmatrix}$, compared to the true values of $\Sigma_1 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$.

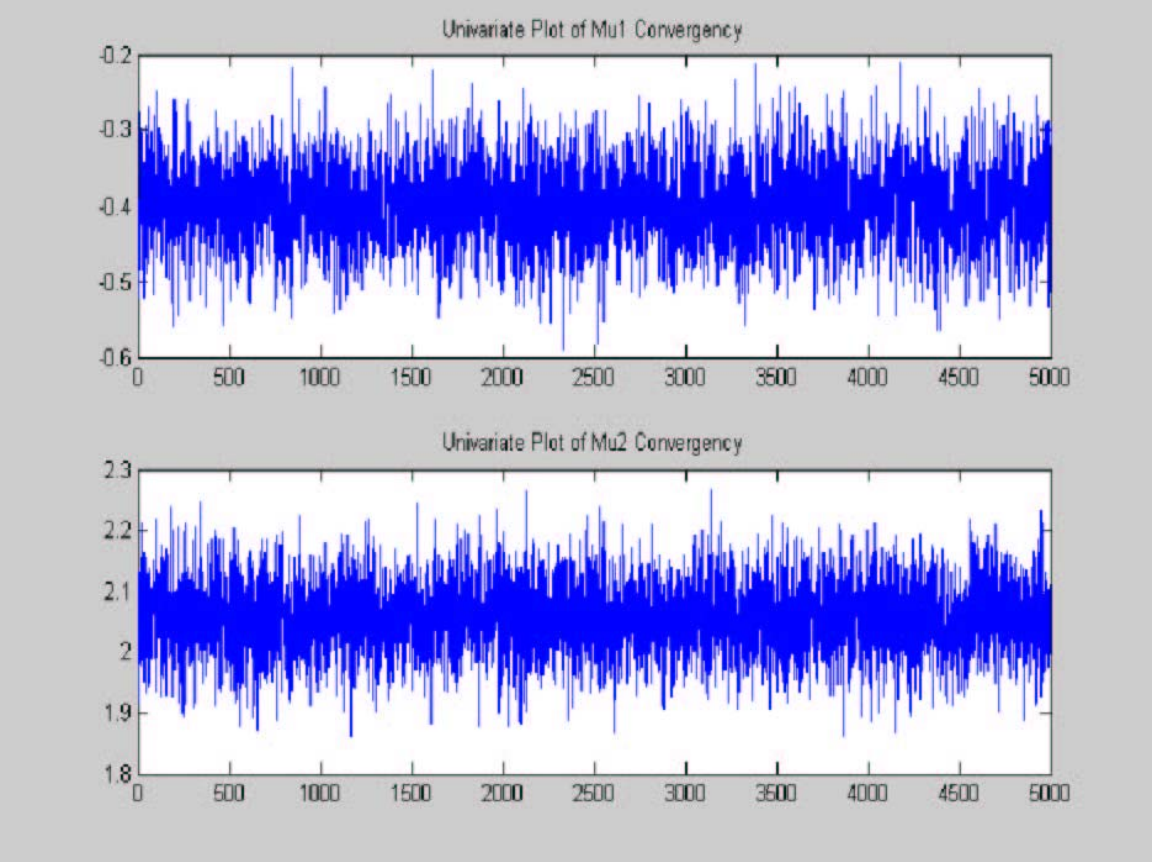
Univariate Plot of Mu1 Convergency
-0.2
-0.3
-0.4
-0.5
-0.6
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Univariate Plot of Mu2 Convergency
2.3
2.2
2.1
2
1.9
1.8
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000

Figure 1

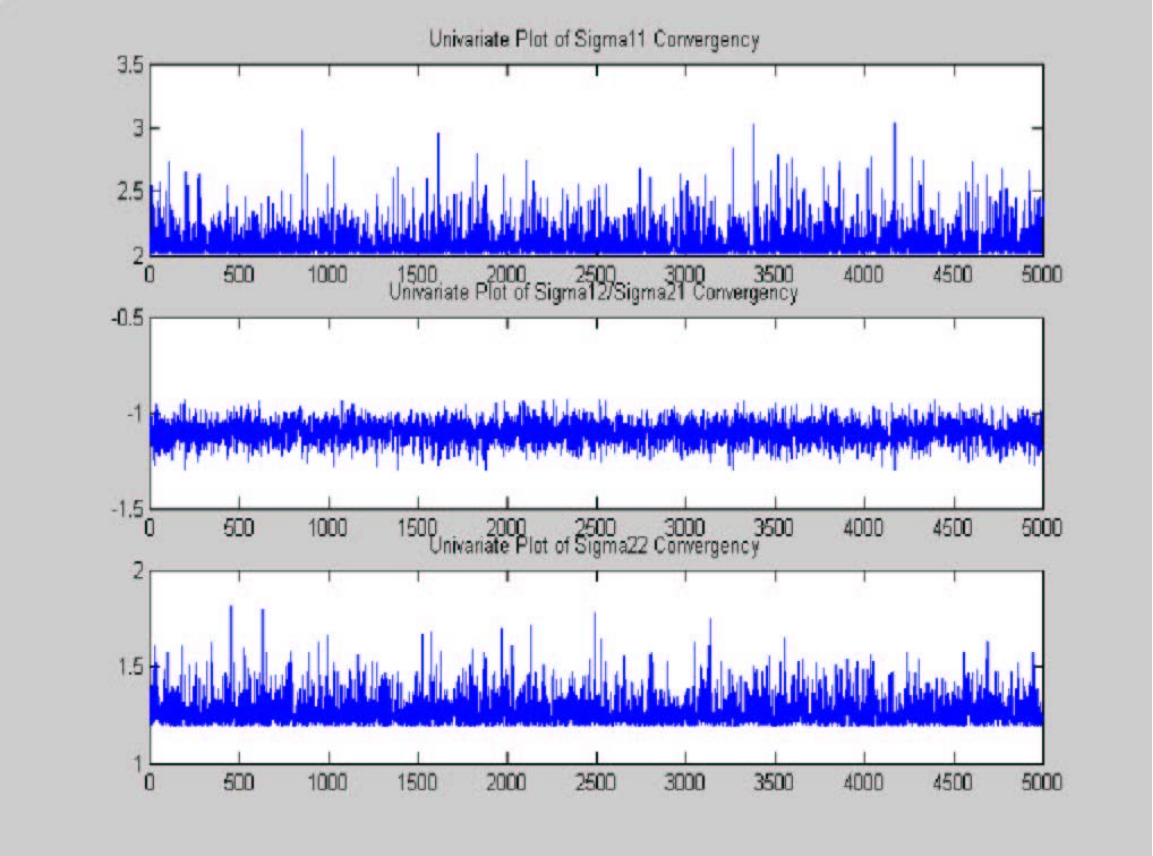

Figure 2

The mean vector for the second population is given by $\hat{\mu}_2 =$ [0.001, 0.02] and the Bayesian estimate of the covariance matrix $\hat{\Sigma}_2 = \begin{pmatrix} 1.0 & -0.0 \\ -0.0 & 1.2 \end{pmatrix}$, compared to the true values of $\mu_2 =$ [0, 0] and $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Both the mean vector and the variance matrix are very close to their true values. The estimated values of the missing observations (with the corresponding actual values in parentheses) are $y_{(10,1)} = 2.4569$ (2.08), $y_{(15,2)} = 1.7973$ (2.013), $y_{(101,1)} = 0.137$ (1.008) and $y_{(120,2)} = -0.659$ (-0.25). These are reasonable compared to the neighboring observed data within the column (variable), and also when compared to the average values of the variable (column) containing the missing values.

5. Analysis of Financial Data

We apply this new approach to a sample of companies that offer direct investment plans and a corresponding, size-matched set of companies without such plans. Dividend Reinvestment Plans and a more general class of investments, Direct Investment Plans, allow investors to avoid investment channels typically used in the past, such as securities brokers. A Dividend Reinvestment Plan is a mechanism that permits shareholders to reinvest their dividends in additional shares automatically. No broker is involved, unless he is the agent of the plan administrator. If the firm does not restrict its plan to current shareholders, then the plan is also what is called a Direct Investment Plan, sometimes known as a Super DRIP. Transactions costs are typically much lower than when using traditional brokerage accounts.

DRIPS are not a different class of security, such as swaps or options. They are simply a new way of selling the traditional equity security. The privileges and obligations of equity ownership are unchanged. For example, DRIP investors receive the usual mailings and retain all voting rights. Tax implications are unaffected, and stock splits are handled exactly as if the investor were using a traditional brokerage account. Readers seeking more detailed information about such plans should see DeGennaro (2003).

Data are from the firms listed in *The Guide to Dividend Reinvestment Plans* (1999) and the Compustat database. These data comprise 15 financial variables and are a subset of those used in DeGennaro (2003). Because DRIP firms tend to be much larger in terms of total assets than companies without such plans (DeGennaro, 2001), we match the 906 DRIP companies with available data to a sample of firms without such plans, for a total of 1812 companies. We use total assets in 1999 as our matching variable. Some companies have missing values for certain variables, but this is not a serious problem given our method; indeed one of the strengths of our approach is that it handles such characteristics. From the perspective of the financial economist,

these data provide information that may let us determine the likelihood that companies without plans will adopt one. Given the results of Dubofsky and Bierman (1988), the ability to predict such an adoption before the marginal investor can do so represents a potentially profitable trading opportunity. In addition, companies that administer direct investment plans that seek new customers can produce a list of firms most likely to be interested in purchasing their services. The reverse is also possible: we can improve our predictions of which companies are likely to abandon their plans, and plan administrators can improve their predictions about which customers are at greatest risk to become former customers. Predicting changes in plan terms may also be possible.

Table 3 presents sample statistics. Only two variables (Total Assets and After Tax Return on Total Assets) have no missing values. Still, we have upwards of 1650 observations but one variable. Because of certain screens to eliminate extreme observations (DeGennaro, 2003), almost all observations on all variables lie within a reasonable range. Exceptions occur for certain ratios with denominators near zero. For example, Compustat defines the *Payout Ratio* as essentially the dollar amount of dividends paid to common shareholders divided by earnings. Because earnings can be near zero, ratios can be large in absolute value. Even these cases, though, are relatively rare.

Table 3

Sample Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
Total Assets (MM$)	1,812	14,140.7	45,776.1	6.38	575,167
Research and Development Expense (MM$)	698	262.93	799.06	0	7,100
Net Sales (MM$)	1,809	5,289.74	13,163.4	0	173,215
Payout Ratio	1,755	36.1	166.41	-3,626.04	3,192.31
Dividend Yield	1,682	2.81	10.15	0	394.45
Common Shares Outstanding (MM)	1,765	194.88	460.09	0	6,133.40
Common Shares Traded	1,680	165.82	514.06	0	8,129.70
Common Shareholders (M)	1,387	36.92	156.99	0	4,206.32
Employees (M)	1,620	21.33	54.26	0	1,140
Net Profit Margin	1,807	4.75	47.96	-1,324.84	726.95
After Tax Return on Common Equity	1,801	9.8	270.12	-6,812.12	8,563.59
After Tax Return On Total Assets	1,812	2.78	9.73	-117.33	157.33
Debt Ratio	1,810	0.69	0.23	0	2.74
Market To Book	1,669	2.96	8.82	-238.17	121.53
P/E at Fiscal Year End	1,682	18.39	101.13	-1,693.8	1,437.50

Source: Authors' calculations.

Table 4 contains the number of observations for the subsets of firms with and without DRIPs, and where meaningful, the means for each group. It also reports t-ratios testing the equality of the means. The first question of interest is the efficacy of the size-matching procedure. Because the number of DRIP firms is a fairly large proportion of the total firms in the size range, there is simply no good match for all companies. In such cases, we match to the closest available company, even though this sometimes means that an individual firm is perhaps 10% larger or smaller than its matched company. This procedure works well under the circumstances, though. Companies without DRIPs are a little bigger than those with plans, but the difference is less than 4% and is insignificant by any conventional standard.

Table 4 shows that several variables do differ significantly. For example, DRIP companies have higher payout ratios and dividend yields (the table is constructed so that a

negative t-ratio means companies with plans have the larger value. They tend to have more common shareholders and more employees. They tend to be more profitable, with higher net profit margins and higher after-tax return on total assets. Economic reasons for these results and further tests are in DeGennaro (2003). For our purposes, the point is that these differences hold promise for partitioning the data into homogeneous clusters.

Table 4

Means and t-tests, 906 Companies with DRIPs Compared to 906 Companies without

Variable	**Number of Observations**	**Mean**	**Maximum**

	No plan	DRIP plan	No plan	DRIP plan	t-statistic
Total Assets (MM$)	906	906	14,412	13,870	0.25
Research and Development Expense (MM$)	317	381	265.96	260.4	0.09
Net Sales (MM$)	904	905	4,737	5,842	-1.79
Payout Ratio	875	880	19.34	52.77	-4.23**
Dividend Yield	777	905	1.46	3.96	-5.07**
Common Shares Outstanding (MM)	862	903	181.43	207.73	-1.2
Common Shares Traded	774	906	191.9	143.56	1.92
Common Shareholders (M)	675	712	23.46	49.65	-3.17**
Employees (M)	800	820	18.49	24.1	-2.08*
Net Profit Margin	902	905	1.3	8.2	-3.06**
After Tax Return on Common Equity	896	905	14.79	4.85	0.78
After Tax Return On Total Assets	906	906	1.53	4.04	-5.53**
Debt Ratio	905	905	0.69	0.69	0.44
Market To Book	766	903	3.07	2.87	0.46
P/E at Fiscal Year End	777	905	20.8	16.31	0.91

Source: Authors' calculations.

Using Data Augmentation and the Gibbs sampler, we run the algorithm for 1000 iterations. The BSC criterion (Table 5) shows that the most likely number of clusters is two. This is consistent with partitioning the companies into those that have DRIPS and those that do not. The confusion matrix for the two clusters is in Table 6. The misclassification error rate is

only 3.4%.

Table 5
BSC values for different number of components

K	*BSC*
1	7818.19
2	7680.90 *
3	8798.92
4	8448.44

Table 6: Confusion matrix

K	No DRIP	DRIP	Total
No Plan	851	55	906
Plan	5	901	906
Total	856 (cluster 1)	956 (cluster 2)	n = 1812

6. Discussion

What economic or managerial implications can we draw from this study? Table 6 is the key. The first row shows that our method correctly classifies 851 of the 906 companies without DRIPS, meaning that 55 companies have been misclassified: According to the model, they should have DRIPS, but in reality, they do not. One interpretation is that the model is simply wrong about 6% of the time when it is used to identify companies with DRIPS. However, another interpretation is that these companies are likely candidates to adopt a plan. A DRIP plan administrator could do far worse than contacting the representatives of these 55 companies to gauge their interest in introducing a DRIP. This is because these companies' financial data show that some aspect of their operations corresponds to firms that typically offer DRIPS. These firms are probably the most likely candidates to start a plan. The second row shows that the procedure does even better for firms that have no DRIP: only *five* companies classified as having no DRIP actually have them, while 901 are correctly classified as having a DRIP. Applying the same reasoning as for the first row, the managerial interpretation is that the plan administrator is most

likely to lose these five companies as customers – the data indicate that some aspect of their financial statements corresponds to firms that do not offer DRIPs.

Other financial applications of this method are easy to find. First, it has obvious value to regulators. Consider the problem of mortgage lending discrimination. Regulators have long been charged with monitoring fairness. Essentially, the problem reduces to determining whether members of one race are equally likely to be denied a mortgage compared to similarly situated member of other races. This problem is extremely difficult for any of several reasons (see Black, Boehm and DeGennaro, 2001 and Black, Boehm and DeGennaro, 2003 for details). Part of the problem is missing data. For example, loan officers often fail to collect all of the usual data for loan applications that are almost sure to be denied, because collecting all of it is likely to be a waste of time. In addition, institutions sometimes gather information that other lenders ignore. This produces missing values when the data are combined across lenders. Because our paper's approach handles missing data well, we conjecture that regulators could identify rejected applicants that, at least according to the method, could easily have been approved. Given that regulatory resources are scarce, it makes sense to concentrate on the cases that are most likely to be problems.

Managers in the private sector, of course, see the matter from the other side. They might use the method to insure compliance with regulations rather than to identify lapses. In addition, this could identify potential profit opportunities. After all, the model identifies a pool of mortgage applications that were denied, yet which had financial characteristics very similar to other applications that were approved. By studying this pool of rejections, management could possibly refine its approval process so that profitable loans are less likely to be missed.

References

Altman, Edward I. (2000). Predicting Financial Distress of Companies: Revisiting the Z-Score and Zeta® Models. Working paper.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis,* 2nd Edition. Wiley.

Baillie, R. and Bollerslev, T. (1989). "Common Stochastic Trends in a System of Exchange Rates." *Journal of Finance 44*, 167-181.

Ben-Akiva, M. E. (1973). Structure of passenger travel demand models. Ph.D. thesis, Department of Civil Engineering, MIT, Cambridge, Ma.

Bensmail, H. and Bozdogan, H (2002). "Regularized Discriminant analysis with Optimally scaled data. In Measurement and Multivariate Analysis, S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji. (Eds), Springer, Tokyo, Japan, 133-144.

Bensmail, H. and Bozdogan. (2003a). Multivariate Kernel Mixture-Model Cluster Analysis for Mixed Data," Working paper.

Bensmail, H. and Bozdogan, H. (2003b). Bayesian kernel clustering mixture-model cluster analysis of mixed and imputed data using Information Complexity. Working Paper.

Bensmail, H., Raftery, A. Celeux, G. and Robert, C. (1997). Inferences for model-based cluster analysis. *Computing and Statistics 7,* 1-10

Bierlaire, M., Lotan, T., and Toint, Ph. (1997). On the overspecification of multinomial and nested logit models due to alternative specific constants. Transportation Science, 1997. (forthcoming).

Black, Harold A., Thomas P. Boehm and Ramon P. DeGennaro (2003). "Is There Discrimination in Overage Pricing?" *Journal of Banking and Finance 27,* Number 6, 1139 - 1165.

Black, Harold A., Thomas P. Boehm and Ramon P. DeGennaro (2001). "Overages, Mortgage Pricing and Race," *International Journal of Finance 13*, 2057-2073.

Bozdogan, H (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370.

Calhoun, Charles A. and Yongheng Deng (2000). A Dynamic Analysis of Fixed and Adjustable-Rate Mortgage Terminations. *The Journal of Real Estate Finance and Economics. 24* # 1 & 2.

Dalhstedt, Roy, Timo Salmi, Martti Luoma, and Arto Laakkonen. (1994). On the Usefulness of Standard Industrial Classifications in Comparative Financial Statement Analysis. *European*

Journal of Operational Research 79, No. 2, 230-238.

Dasgupta, A. and Raftery, A. E. (1998). Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering, *Journal of the American Statistical Association 93*: 294–302.

DeGennaro, Ramon P. (2003). "Direct Investments: A Primer." *Economic Review 88*, Number 1, 1-14.

DeGennaro, Ramon P. (2001). "Direct Investments: A Primer." University of Tennessee Working Paper.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B.,* 39*,* 1-38.

Dubofsky, D. and Bierman, L. (1988), "The Effect of Discount Dividend Reinvestment Plan Announcements on Equity Value," *Akron Business and Economic Review* 19*,* 58-68.

Guide to Dividend Reinvestment Plans (1999). *Temper of the Times Communications, Inc.*

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley, New York.

Hosmer, D.W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics 29,* 761-770.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding Groups in Data*, Wiley, New York.

Koppelman, F. S. and Chieh-Hua Wen. (1997). The paired combinatorial logit model: properties, estimation and application. Transportation Research Board, 76th Annual Meeting, Washington DC.

Luce, R. D. and Suppes, P. (1965). Preference, utility and subjective probability. In R. D. Luce, R. R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, New York, J. Wiley and Sons.

McFadden, D. (1978). Modelling the choice of residential location. In A. Karlquist et al., editor, Spatial interaction theory and residential location, Amsterdam. North-Holland, 75-96.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models Inference and Applications to Clustering*. Marcel Dekker, Inc., New York.

Mester, Loretta J. (1997). What's the Point of Credit Scoring? Federal Reserve Bank of Philadelphia *Business Review*, September/October, 3-16.

Roeder, Kathryn and Larry Wasserman. Practical Bayesian density estimation using mixtures of norms. *Journal of the American Statistical Association.* September.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics 6,* 461-464.

Shaffer, Sherrill (1991). Aggregate Deposit Insurance Funding and Taxpayer Bailouts. *Journal of Banking and Finance*, September.

Smith A.F., and Roberts G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Royal Stat. Soc. B 55,* 3–23.

Swendsen, R. H. and J. S. Wang. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters,* 58*:*86-88.

Tanner, Martin A. and Wing Hung Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Society,* 82 *(*398*):* 528-550.

Tierney, L (1994) "Markov chains for exploring posterior distributions (with discussion)," *Ann. Statist.,* 22*,* 1701-1758.

von Altrock, Constantin (1995). Fuzzy Logic and Neuro Fuzzy Applications Explained, Inform Software Corp., Germany, Prentice Hall.