

Cluster Analysis of Imputed Financial Data Using an Augmentation-Based Algorithm.

Halima. Bensmail,
Statistics Department, University of Tennessee
Ramon De Gennaro
Statistics Department, University of Tennessee

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. Part of this work was completed while DeGennaro served as a Visiting Scholar at the Federal Reserve Bank of Atlanta. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Please do not quote without permission. Any errors are the authors'.

1. Introduction

In this paper, we introduce an apply a novel statistical modelling technique based on the on-going work of Bensmail and Bozdogan (2003a,b) to cluster analysis for imputed financial data (DeGennaro 2003). We have two main purposes to achieve: one, to forecast (supervised clustering), and two, to find homogeneous groups within the data (unsupervised clustering). We stress that this is a flexible (or agile) modelling approach in the following sense: we handle large and complex data structures (data mining) with missing observations with mixed structure, that is, data types with both quantitative and qualitative measurements. We achieve this by mapping the data to a new structure which is free of distributional assumptions in choosing homogeneous groups of observations (Bensmail and Bozdogan 2002). 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 those homogeneous clusters (transactions) and to forecast the ability 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.

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. Koppelman and Wen (1997) and Vovsha (1997) have recently developed newer variations. The nested logit model of Ben-Akiva (1973) is designed to handle correlations among alternatives. Other variations of multinomial logic have been developed or used by Koppelman and Wen (1997), 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) is 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.

Though not discrete choice models, methods that produce a continuous variable can 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. For example, applicants with a score below a certain point might be rejected automatically. Applicants above a specified higher point may 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. Some methods can determine the number of classes without providing 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. Data and Preliminary Tests

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, also known as a Super DRIP. Transaction 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 they 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 data base. These data are a subset of those used in the forthcoming work of DeGennaro (2003), and include 36 financial variables. 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 1: Sample Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
<i>Total Assets MM\$</i>	1,812	14,140.7	45,776.1	6.38	575,167
Property, Plant and Equipment MM\$	1,682	2,534.27	6,332.99	0	94,043
Property, Plant, and Equipment (Capital Expenditures) MM\$	1,463	510.6	1,762.14	-5.7	30,549
<i>Research and Development Expense MM\$</i>	698	262.93	799.06	0	7,100
<i>Net Sales MM\$</i>	1,809	5,289.74	13,163.4	0	173,215
<i>Payout Ratio</i>	1,755	36.1	166.41	-3,626.04	3,192.31
<i>Dividend Yield</i>	1,682	2.81	10.15	0	394.45
<i>Common Shares Outstanding MM</i>	1,765	194.88	460.09	0	6,133.4
<i>Common Shares Traded</i>	1,680	165.82	514.06	0	8,129.7
<i>Treasury Stock - Number of Common Shares MM</i>	1,775	9.79	47.32	0	994.8
<i>Common Shareholders M</i>	1,387	36.92	156.99	0	4,206.32
<i>Employees M</i>	1,620	21.33	54.26	0	1,140
Sale of Common and Preferred Stock MM\$	1,469	105.11	468.08	-1.44	10,694.7
Purchase of Common and Preferred Stock MM\$	1,417	123.02	431.38	-0.12	6,645
Pretax Income MM\$	1,810	547.68	1,503.17	-3,889	15,942
Net Income (Loss) MM\$	1,812	358.58	1,016.22	-2,501.6	10,717
S&P Senior Debt Rating Code	1,038	10.21	3.47	2	27
S&P Commercial Paper Rating - Historical Code	500	102.33	0.91	101	107
Risk-Adjusted Capital Ratio (Tier 1)	255	10.4	3.95	5.6	34.7
Risk-Adjusted Capital Ratio (Total)	259	13.75	5.14	6.9	57.06
Non-performing Assets MM\$	261	93.96	313.5	0	3,075
Provision for Loan/Asset Losses MM\$	267	77.58	288.02	-175	2,837
Net Interest Margin (Ratio)	262	4.03	1.06	0.85	12.24
Interest Expense Per Share	1,417	4.74	100.12	0	3,628.12
<i>Net Profit Margin</i>	1,807	4.75	47.96	-1,324.84	726.95
PreTax Interest Coverage	1,470	22.48	474.47	-7,122.29	14,481.7
PreTax Profit Margin	1,807	8.87	56.58	-1,324.84	1,187.23
PreTax Return On Assets	1,810	4.8	11.4	-117.33	157.33
Operating Income Before Depreciation to Total Assets	1,682	10.25	10.02	-97.89	80.2
After Tax Interest Coverage	1,470	16.35	466.84	-8,144.57	14,480.5
<i>After Tax Return On Common Equity</i>	1,801	9.8	270.12	-6,812.12	8,563.59
<i>After Tax Return On Total Assets</i>	1,812	2.78	9.73	-117.33	157.33
<i>Debt Ratio</i>	1,810	0.69	0.23	0	2.74
<i>Market To Book</i>	1,669	2.96	8.82	-238.17	121.53
<i>P/E at Fiscal Year End</i>	1,682	18.39	101.13	-1,693.8	1,437.5
<i>Earnings Per Share</i>	1,727	1.68	9.11	-51.66	276.02

Variable	Number of Observations		Mean		t-statistic
	No plan	DRIP plan	No plan	DRIP plan	
<i>Total Assets MMS</i>	906	906	14,412	13,870	0.25
Property, Plant and Equipment MMS	851	831	2,428	2,644	-0.7
Property, Plant, and Equipment (Capital Expenditures) MMS	751	712	535.05	484.81	0.54
<i>Research and Development Expense MMS</i>	317	381	265.96	260.4	0.09
<i>Net Sales MMS</i>	904	905	4,737	5,842	-1.79
<i>PayoutRatio</i>	875	880	19.34	52.77	-4.23**
<i>Dividend Yield</i>	777	905	1.46	3.96	-5.07**
<i>Common Shares Outstanding MM</i>	862	903	181.43	207.73	-1.2
<i>Common Shares Traded</i>	774	906	191.9	143.56	1.92
<i>Treasury Stock - Number of Common Shares MM</i>	884	891	5.16	14.39	-4.12**
<i>Common Shareholders M</i>	675	712	23.46	49.65	-3.17**
<i>Employees M</i>	800	820	18.49	24.1	-2.08*
Sale of Common and Preferred Stock MMS	757	712	155.93	51.07	4.32**
Purchase of Common and Preferred Stock MMS	706	711	75.49	170.21	-4.16**
Pretax Income MMS	905	905	473.8	621.55	-2.09*
Net Income (Loss) MMS	906	906	313.83	403.33	-1.88
S&P Senior Debt Rating Code	491	547	N/A	N/A	N/A
S&P Commercial Paper Rating - Historical Code	163	336	N/A	N/A	N/A
Risk-Adjusted Capital Ratio (Tier 1)	107	148	10.44	10.37	0.13
Risk-Adjusted Capital Ratio (Total)	110	149	14.22	13.41	1.24
Non-performing Assets MMS	109	152	125.77	71.15	1.39
Provision for Loan/Asset Losses MMS	115	152	107.12	55.23	1.46
Net Interest Margin (Ratio)	111	151	3.91	4.11	-1.47
Interest Expense Per Share	678	739	8.4	1.38	1.32

Table 1 presents sample statistics. Only three variables (Total Assets, Net Income, and After Tax Return on Total Assets) have no missing values. Five variables are only available for financial institutions, so about six of every seven observations is missing for these. Still, we have upwards of 1650 observations for most variables. Note that the two S&P rating codes are categorical.

Because of certain screens to eliminate extreme observations (DeGennaro, forthcoming), 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 2 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 with 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 2 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 generally have higher margins before taxes, and generally, higher earnings. They purchase more shares on the open market (probably to meet the needs of the plan), but sell fewer shares (probably because they tend to grow less rapidly than companies without plans). None of the t-ratios reject the equality of means for variables unique to financial institutions, possibly due to the smaller samples. Economic reasons for these results and further tests are in DeGennaro (forthcoming). For our purposes, the point is that these differences hold promise for partitioning the data in Section IV.

We also conduct our tests on a subset of these data. This smaller dataset uses only 16 of the 36 variables in Table 1 and Table 2. Gathering data is costly and researchers would prefer to collect smaller dataset if the information loss is minimal. Thus, we use this smaller dataset to check the power of our method to partition the data correctly. <The variables in the subset of the data are printed in the tables using italics>.

3. Clustering and Bayesian Data Augmentation

In this section without going into technical details, we briefly discuss Bayesian data augmentation (see ongoing work of Bensmail and Bozdogan 2003a,b) which we will apply to clustering and data mining the imputed financial data as described above.

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. In the past years 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 this, see, e.g., Hartigan (1975), Kaufman and Rousseeuw (1990), Bozdogan (1994), Gelman (1995), Gordon (1999) and Bensmail and Bozdogan (2002).

One assumes that the population of interest consists of K different subpopulations

G_1, \dots, G_K and that the density of a p -dimensional observation \mathbf{x} from the k th subpopulation is $f_k(\mathbf{y}, \boldsymbol{\theta}_k)$ for some unknown vector of parameters $\boldsymbol{\theta}_k$ ($k = 1, \dots, K$). Given observations $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$, we let $\mathbf{v} = (v_1, \dots, v_n)^t$ denote the unknown identifying labels, where $v_i = k$ if \mathbf{y}_i comes from the k th subpopulation. Clustering data using a mixture distribution framework was succesful lately and many authors proposed different approaches. Good sources of references are the paper by Hosmer (1973a, 1973b). In most cases, the data to be classified are viewed as coming from a mixture of probability distributions (McLachlan and Basford 1988; Tan and Chang 1972), each representing a different cluster, so the likelihood is expressed as

$$p(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K; \pi_1, \dots, \pi_K | \mathbf{x}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i | \boldsymbol{\theta}_k) \quad (1)$$

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

Clustering data with missing values has always been difficult. In many cases, sample means were used to replace the missing values and then any clustering method would apply 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.

When \mathbf{y} represents an observation from the sample, we use \mathbf{y}_{obs} for the observed complete part of the data and \mathbf{y}_{miss} the missing one. Here we want to estimate the parameter $\boldsymbol{\theta}$ given in (1) based on $\mathbf{y} = (\mathbf{y}_{miss}, \mathbf{y}_{obs})$. The data (observed and missing) should be clustered in K clusters, the cluster G_1, \dots, G_K are unknown. Each observation y_i belonging to a cluster G_k is supposed to be a random variable drawn from a normal distribution $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance matrix of the cluster G_k such that

$$(\mathbf{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (2)$$

The distribution function of a sample $(\mathbf{y}_1, \dots, \mathbf{y}_{n_k})$ representing a subpopulation G_k is given by

$$\begin{aligned} p(\mathbf{y}_1, \dots, \mathbf{y}_{n_k} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &\propto |\boldsymbol{\Sigma}_k|^{-n_k/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n_k} (\mathbf{y}_i - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k)\right) \\ &= |\boldsymbol{\Sigma}_k|^{-n_k/2} \exp\left(-\frac{1}{2} tr \mathbf{W}_k \boldsymbol{\Sigma}_k^{-1}\right) \end{aligned} \quad (3)$$

where $n_k = \sum_{i \in G_k} I\{v_i = k\}$, $\mathbf{W}_k = \sum_{i: v_i=k} (\mathbf{y}_i - \bar{\mathbf{y}}_k)(\mathbf{y}_i - \bar{\mathbf{y}}_k)^t$, $\bar{\mathbf{y}}_k = \frac{1}{n_k} \sum_{i: v_i=k} \mathbf{y}_i$.

In the Bayesian clustering approach, one needs to estimate the posterior distribution of the parameter $\boldsymbol{\theta}$ involved given its prior distribution. When \mathbf{y}_{miss} denote a subvector of \mathbf{y} containing the missing components, the posterior distribution of the parameter $\boldsymbol{\theta}$ given the observed data \mathbf{y}_{obs} is

$$f(\boldsymbol{\theta} | \mathbf{y}_{obs}) = \int f(\boldsymbol{\theta} | \mathbf{y}_{miss}, \mathbf{y}_{obs}) f(\mathbf{y}_{miss} | \mathbf{y}_{obs}) d\mathbf{y}_{miss} \quad (4)$$

which is a mixture of the posterior distribution of $\boldsymbol{\theta}$ given the data (observed and missing) where the mixing proportion is given by the marginal conditional distribution of \mathbf{y}_{miss} . This is typically very difficult to use, as it often cannot even be expressed in a closed form (without integral).

A very useful method for getting around these difficulties and exploring $f(\boldsymbol{\theta} | \mathbf{y}_{obs})$ is the data augmentation (DA) algorithm. The term 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 Tanner and Wong (1987) Data Augmentation algorithm for posterior sampling, and in the physics literature by Swendsen and Wang (1987) algorithm for sampling from Ising and Potts models (Cipra, 1987) (and its generalizations; in the physics literature, the method of data augmentation is referred to as the method of auxiliary variables. Data augmentation schemes were used by Tanner and Wong to make simulation feasible and simple, while auxiliary variables were adopted by Swendsen and Wang (1987) 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).

In the following, we will describe the DA algorithm for imputing the missing data. The algorithm iterates as the following:

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

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

$$\mathbf{y}_{miss}^{(t+1)} \sim f(\mathbf{y}_{miss} | \mathbf{y}_{obs}, \theta^{(t)}) \quad (5)$$

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

$$\theta^{(t+1)} \sim f(\theta | \mathbf{y}_{obs}, \mathbf{y}_{miss}^{(t)}) \quad (6)$$

2.1 Imputation

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

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

$$\mathbf{y}_1 | \mathbf{y}_2 \sim N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_2^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}) \quad (7)$$

where

$$\mathbf{y}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_1), \text{ and } \mathbf{y}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_2) \quad (8)$$

and

$$\Sigma_{12} = \Sigma_{21} = \text{Cov}(\mathbf{y}_1, \mathbf{y}_2) \quad (9)$$

Case of one missing value and many are observed:

Suppose that an observation $\mathbf{y} = (y_1, y_2, \dots, y_p)$ has one missing value. Let's consider $z_1 = y_1$ the missing value and $\mathbf{z}_2 = (y_2, \dots, y_p)$ the remaining observed values. Then the only information needed are parts of the vector mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_p)$ and the covariance matrix Σ .

Given the mean $\boldsymbol{\mu}$ and given the covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \boxed{\sigma_{12}} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \\ \dots & \dots & \dots & \\ \sigma_{p1} & \dots & \dots & \sigma_{pp} \end{pmatrix}$, the only

input we need are the first row of the covariance matrix except the first variance term, which means the vector $\boldsymbol{\sigma}_{1k(-1)} = (\sigma_{12}, \sigma_{13}, \dots, \sigma_{1p})$ and the covariance matrix minus the first row and

the first column which means the matrix $\boldsymbol{\Sigma}_{-(1,1)} = \begin{pmatrix} \boxed{\sigma_{22}} & \boxed{\sigma_{23}} & \boxed{\sigma_{2p}} \\ \dots & \dots & \\ \sigma_{p2} & & \sigma_{pp} \end{pmatrix}$ and then we can use

those blocks to estimate the missing value $z_1 = y_1$, by generating the data from a normal distribution

$$z_1 | z_2 \sim N \left(\begin{array}{c} \boldsymbol{\mu}_1 + \boldsymbol{\sigma}_{1k(-1)} \boldsymbol{\Sigma}_{-(1,1)} (y_2 - \mu_2, \dots, y_p - \mu_p)^t, \\ \sigma_{11} - \boldsymbol{\sigma}_{1k(-1)} \boldsymbol{\Sigma}_{-(1,1)} \boldsymbol{\sigma}_{1k(-1)}^t \end{array} \right)$$

General case:

For the general case, we have multivariate data $\mathbf{y} = (y_1, y_2, \dots, y_p)$ where two y_j and y_h or more are missing and the others are observed. Using the same schema as before, the only information needed are parts of the vector mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. Using Anderson's Lemma (1984), $(y_j, y_h) | (y_2, \dots, y_p, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is normally distributed with mean vector $\tilde{\boldsymbol{\mu}}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}$ as described in Bensmail and Bozdogan (2003b). Here we will not describe the details of the calculating $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$, see the previous work for more details.

3.2 Posterior Estimation:

To estimate the parameter $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we need to specify priors on those parameters. Here we use conjugate priors for the parameters $\boldsymbol{\pi}$ which is a *Dirichlet* distribution $Dirichlet(\alpha_1, \dots, \alpha_K)$.

Since the log-likelihood is a quadratic form in $\boldsymbol{\mu}_k$, the conjugate prior distribution of $\boldsymbol{\mu}_k$ is given by:

$$\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k \sim N_p(\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k / \tau_k), \quad (10)$$

and a conjugate prior of $\boldsymbol{\Sigma}_k$ is given by

$$\boldsymbol{\Sigma}_k \sim W_p^{-1}(m_k, \boldsymbol{\Psi}_k) \quad (11)$$

The posterior distribution of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ given the missing and observed data are then given by:

$$\boldsymbol{\mu}_k | \mathbf{y}_{miss}, \mathbf{y}_{obs}, \boldsymbol{\Sigma}_k \propto N_p((\boldsymbol{\xi}_k + n_k \bar{\mathbf{y}}_k) / (n_k + \tau_k), \boldsymbol{\Sigma}_k / (n_k + \tau_k)) \quad (12)$$

and

$$\boldsymbol{\Sigma}_k | \mathbf{y}_{miss}, \mathbf{y}_{obs}, \mathbf{v} \sim W_p^{-1} \left(n_k + m_k, \boldsymbol{\Psi}_k + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{y}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{y}}_k - \boldsymbol{\xi}_k)^t \right) \quad (13)$$

3.3 Algorithm:

We estimate the parameters by simulating from the joint posterior distribution of $\mathbf{y}_{miss}, \boldsymbol{\pi}, \boldsymbol{\theta}$ and \mathbf{v} 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 v_i according to their posterior probabilities ($t_{ik}, k = 1, \dots, K$) conditional on $\boldsymbol{\pi}, \mathbf{y}_{miss}$, and $\boldsymbol{\theta}$, namely

$$t_{ik} = \frac{\pi_k f(\mathbf{y} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^K \pi_h f(\mathbf{y} | \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}; i = 1, \dots, n \quad (14)$$

2. Simulate the missing values \mathbf{y}_{miss} given \mathbf{y}_{obs} from

$$\mathbf{y}_{miss} \sim f(\mathbf{y}_{miss} | \mathbf{y}_{obs}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (15)$$

3. Simulate the vector $\boldsymbol{\pi}$ of mixing proportions according to its posterior distribution conditional on the v_i 's. It consists of simulating $\boldsymbol{\pi}$ from its conditional posterior distribution, namely $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_1 + \sum I\{v_i = 1\}, \dots, \alpha_K + \sum I\{v_i = K\})$
4. Simulate the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ according to their posterior distribution

$$\begin{aligned} \boldsymbol{\Sigma}_k^{(t+1)} | \mathbf{y}_{obs}, \mathbf{y}_{miss}^{(t)} &\sim W^{-1}(n_k + m_k, \boldsymbol{\Psi}_k + \mathbf{W}_k + \frac{n_k \boldsymbol{\tau}_k}{n_k + \tau_k} (\bar{\mathbf{y}}_k - \boldsymbol{\xi}_k)(\bar{\mathbf{y}}_k - \boldsymbol{\xi}_k)^t) \\ \boldsymbol{\mu}_k^{(t+1)} | \mathbf{y}_{obs}, \mathbf{y}_{miss}, \boldsymbol{\Sigma}_k^{(t+1)} &\sim N((\boldsymbol{\xi}_k + n_k \bar{\mathbf{y}}_k) / (n_k + \tau_k), \boldsymbol{\Sigma}_k / (n_k + \tau_k)) \end{aligned} \quad (16)$$

where $\bar{\mathbf{y}}_k$ and \mathbf{W}_k are the sample mean and variance matrix of the data, and \mathbf{W}^{-1} denotes the inverse wishart distribution.

4. Bayesian Model Selection for Choosing the Number of Clusters:

The most important component of any cluster analysis algorithm is finding the number of clusters. Different criteria for model selection and number of clusters choice has been investigated in the litterature by many authors e. g., see (Burnham and Anderson (1998)) including, Akaike Information Criteria (AIC) (Akaike 1973), Information Complexity Criteria (ICOMP) (Bozdogan 1987-2002), Normalized Entropy of assessment (NEC) (Celeux and Soromenho 1996), 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-model and Bayesian kernel mixture-model cluster analysis of mixed and imputed data.

For comparative purposes, in this paper, we use Schwarz's criteria, although regularity conditions for this may not hold for mixture models, there is considerable theoretical and practical support for its use (Leroux 1992; Roeder and Wasserman 1997; Mukerjee et al. 1998; Dasgupta and Raftery 1998; and Fraley and Raftery 1998).

SBC, which is an approximation of the Bayes factor is defined by:

$$SBC(M_k) = -2 \log L(\tilde{\boldsymbol{\theta}}_k, \mathbf{M}_k) + m(k) \log(n_k) \quad (17)$$

for choosing the number of component where $L(\boldsymbol{\theta}, M_k)$ is the likelihood of the posterior mode $\tilde{\boldsymbol{\theta}}$ for the model M_k (here the number of component k), and $m(k)$ is the number of parameter to estimate and n_k is the size of the subpopulation G_k

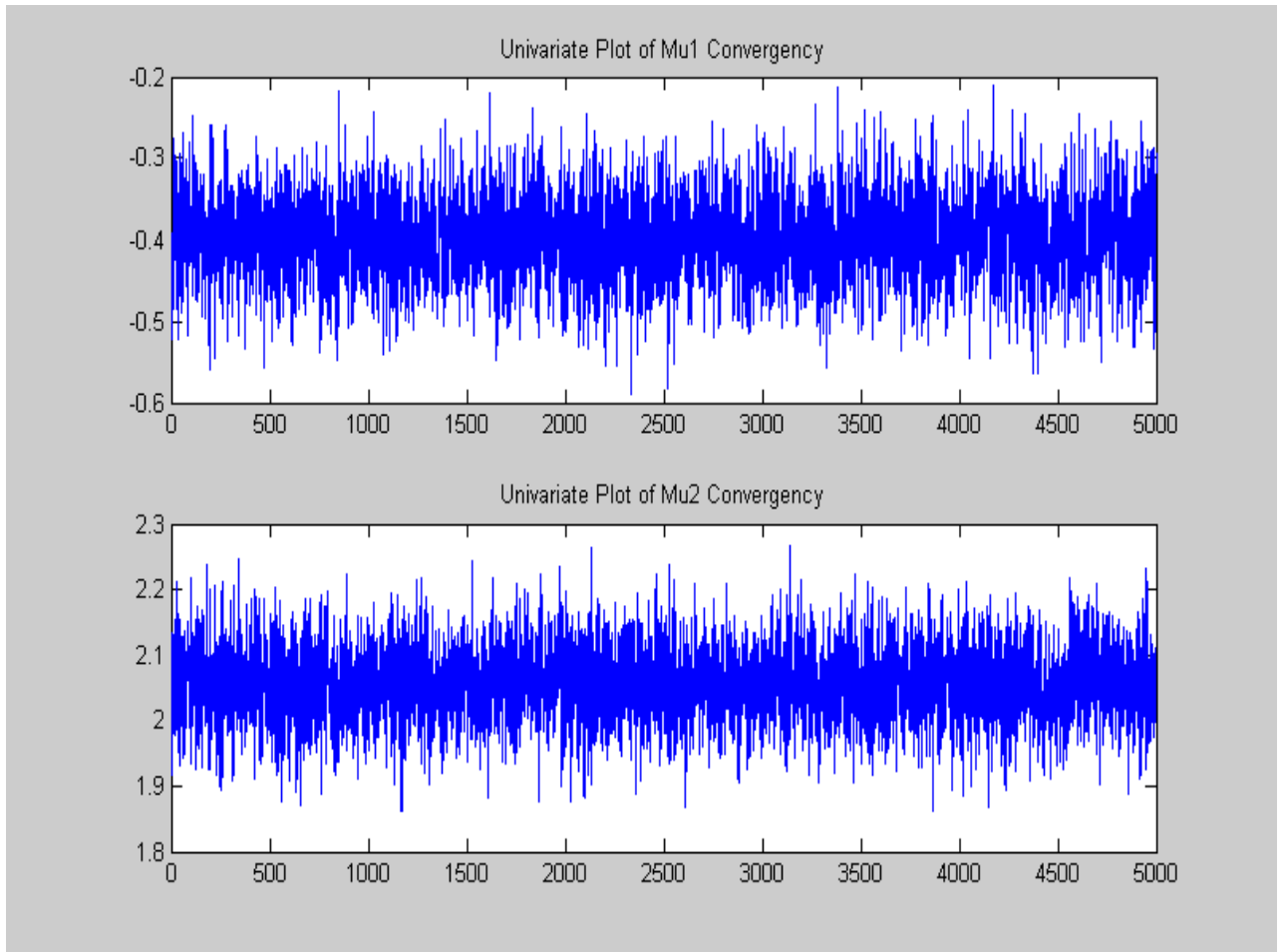
5. Example:

5.1 Simulated data

We simulated 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 $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. We truncated the samples at the following values; $y_{(10,1)}$, $y_{(15,2)}$, $y_{(20,2)}$, $y_{(101,1)}$, and $y_{(120,2)}$.

Using the Data augmentation algorithm, we find in general that the algorithm converges fast and it is stable. The plot of the mean vector μ_1 (variate wide) based on 5000 simulations is given by Figure 1 and its estimation is given by $\hat{\mu}_1 = [-0.3971, 2.0560]$. The mean vector for the second population is given by $\hat{\mu}_2 = (0.001, 0.02)$

Figure 1: convergence of mean vector



The SBC criteria proposes two groups (see Table 3 and Table 4).

k	SBC
1	1811.10
2	1637.03
3	1752.29
4	1748.40

Table 3: SBC values for each

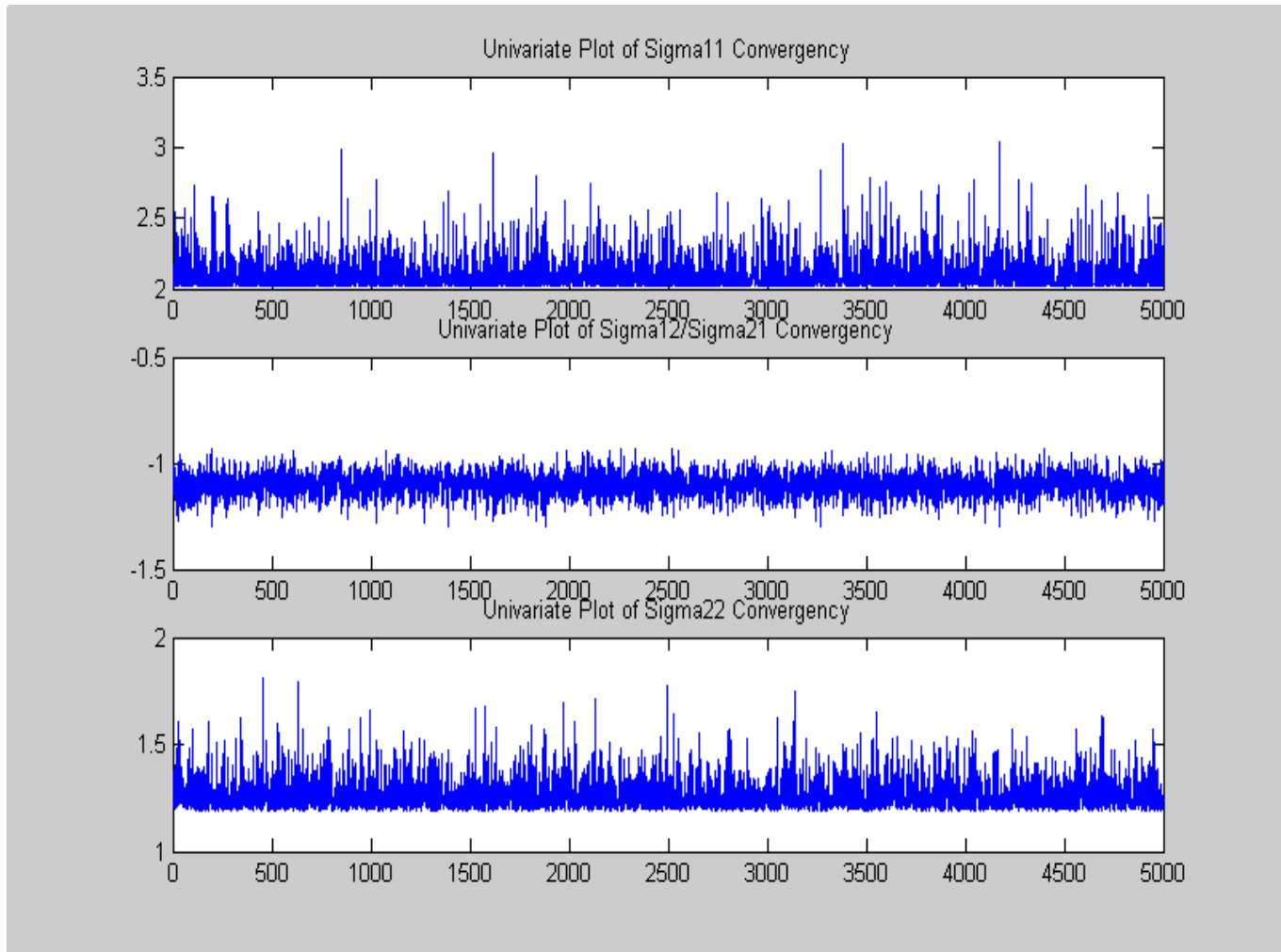
k	1	2	Total
1	58	2	60
2	4	56	60
Total	62(cluster1)	58(cluster2)	$n = 120$

Table 4: Confusion matrix.

The convergence plot for the variance covariance matrix (variate wide) is given by Figure

2 and the Bayesian estimate is $\hat{\Sigma}_1 = \begin{bmatrix} 2.1055 & -1.1016 \\ -1.1016 & 1.2671 \end{bmatrix}$ and the Bayesian estimate of the
 covariance matrix $\hat{\Sigma}_2 = \begin{bmatrix} 1.0095 & -0.0010 \\ -0.0010 & 1.1872 \end{bmatrix}$.

Figure 2: convergence of the Covariance matrix



Both the mean vectors and the variance matrices are close to the true value. For the missing

value estimation, we find the following results:

$$y_{(10,1)} = 2.4569, y_{(15,2)} = 1.7973, y_{(20,2)} = 3.1091, y_{(101,1)} = 0.137, y_{(120,2)} = -0.659.$$

which make sense when we compare them to the neighboring observed data within the column (variable) considered and also when we compare them to the average values of the variable (column) containing the missing values.

5.2. 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, also 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 they 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).

Using the Data augmentation and the Gibbs sampler, we run the algorithm for 1000 iterations. The SBC criteria resulting in Table 3 shows that the two clusters was proposed by SBC which is consistent with the information we have originally from the data. With two clusters, the confusion matrix is given by Table 4 which gives us an error rate of misclassification equal to 0.034. (3.4%).

k	SBC
1	7818.19
2	7680.90
3	8798.92
4	8448.44

Table 5: SBC values for different number of components

k	No Plan	Plan	Total
No Plan	851	55	906
Plan	5	901	906
Total	856(<i>cluster1</i>)	956(<i>cluster2</i>)	$n = 1812$

Table 6: Confusion matrix

In the following, we summarize the estimated parameters using the first four estimates of the vector mean (it will be space consuming to report the whole mean vector and the whole covariance matrix posterior mode). The first four estimates of the posterior mean for the cluster 1 (chosen by SBC) is $\hat{\mu}_1 = (12044.36, 4205.21, 187.64, 204.47)$ and their related posterior covariance matrix estimator is

$$\hat{\Sigma}_1 = \begin{pmatrix} 1774899.9384 & -238289.5187 & 608.8653 & -2883.1619 \\ -238289.5187 & 116291.2842 & 1625.8510 & -646.0837 \\ 608.8653 & 1625.8510 & 142.7815 & -167.2418 \\ -2883.1619 & -646.0837 & -167.2418 & 394.3522 \end{pmatrix}$$

The four terms of the posterior estimated mean for the second cluster is $\hat{\mu}_1 = (13909.44, 5839.01, 206.34, 142.99)$ and their related posterior covariance matrix estimator is

$$\hat{\Sigma}_2 = \begin{pmatrix} 2152933.044 & -416870.6140 & 2459.4332 & 1029.5514 \\ -416870.614 & 238102.3010 & 3120.4234 & 300.1279 \\ 2459.433 & 3120.4234 & 246.0917 & 103.6611 \\ 1029.551 & 300.1279 & 103.6611 & 126.7641 \end{pmatrix}$$

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 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 plan administrator for DRIPS 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 operate a DRIP. 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 with firms that do not operate a DRIP.

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 (readers should 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 continuing to collect it is likely to be a waste of time. In addition, institutions sometimes gather information that other lenders ignore. This also produces missing values. Because this 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?" Forthcoming, *Journal of Banking and Finance*.
- 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.
- Bozdogan, H. (1994): Mixture model cluster analysis using model selection criteria and a new informational measure of complexity. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, 69-113.
- Burnham, K. P. and Anderson, D. R. (1998). *Model selection and inferences, a practical information-theoretic approach*. Springer-Verlag, New York.
- 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.
- Cipra, B. A. (1987). "An Introduction to the Ising Model. *Amer. Math. Monthly* 94, 937-959.
- 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.
- DeGennaro, R. P. (2003): "Direct Investments: A Primer." Forthcoming, Federal Reserve Bank of Atlanta Economic Review.

- 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.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. (1995). *Bayesian data analysis*. Chapman and Hall, London, UK.
- Gifi, A. (1990): *Non linear Multivariate Analysis*. Wiley Series in Probability and Mathematical Statistics, England.
- Gordon, A. D. (1999), *Classification: Methods for the Exploratory Analysis of Multivariate Data*, Chapman and Hall, 2nd Eds. New York.
- Guide to Dividend Reinvestment Plans (1999). *Temper of the Times Communications, Inc.*
- Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley, New York.
- Heiser, W.J., and Meulman, J.J. (1995): *Nonlinear methods for the analysis of homogeneity and heterogeneity*. In: *W.J*
- Hosmer, D.W. (1973a). On Maximum Likelihood Estimator of the parameters of a mixture of two normal distributions when the sample size is small. *Commun. Statist.*, 1, 217-227.
- Hosmer, D.W. (1973b). 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.
- Martin A. Tanner and Wing Hung Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Society*, 82(398):528-550.
- 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.
- 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., 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.
- Krzanowski(Ed.). *Recent Advances in Descriptive Multivariate Analysis*. Clarendon

Press, Oxford.

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

Tan, W.Y. and Chang, W.C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture to two normal densities. *J. Amer. Statist. Assoc.*, 67, 702-708.

Tanner and Wong (1987) "The calculation of posterior distributions by data augmentation", *Journal of American Statistical Association*, 82, 528-550.

Tanner, M.A., and Wei, G.C.G. (1990), "A Monte Carlo Implementation of the EM-Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699– 704.

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.

Vovsha, Peter (1997). Cross-nested logit model: an application to mode choice in the Tel-Aviv metropolitan area. Transportation Research Board, 76th Annual Meeting, Washington DC.