이학 석사 석사학위논문

Analysis of Sales data for the Semiconductor using Data mining

데이터마이닝을 이용한 반도체 판매데이터 분석

2014년 8월

서울대학교 대학원

통 계 학 과

최 윤 영

Abstract

Analysis of Sales data for the Semiconductor using data mining

Yoon young Choi

The Department of Statistics

The Graduate School

Seoul National University

Increased demand for products which are Smartphone, tabletPC and other mobile device using Mobile DRAM over the world makes sales increase of Mobile DRAM. This paper suggests valid statistical methods for application of sales data and analyzes the relation and trend among the type of Mobile DRAM, density and sales area. In addition, we could get another new idea via the result. For analysis, Clustering, logistic regression with lasso, decision tree and Partial Correlation Estimation method are introduced. **glasso** (graphic lasso) that is algorithm to estimate a sparse inverse covariance matrix using lasso penalty (L1 penalty) is used for partial correlation estimation and then, hub network graph is made by **space** (Sparse Partial Correlation Estimation) and available to be used for better decision making and developing a strategy in Marketing and Sales.

.....

Keywords : clustering method, inverse covariance, partial correlation estimation, hub network

Student Number : 2012-23015

Contents

Chapter 1
Introduction

Advances of IT technology makes huge data of various type, so that he analysis and prediction of data derive very valuable work via society. That is coming the era of big data. Many companies need to lead high productivity and efficiency with a quick decision making in unpredictable situations using the data occurring in the real-time. They want to have a decision making with high quality that is made by data analysis techniques than experience or intuition. Therefore data analysis is highly significant work in the era of big data.

Statistics is one of the core studies for the analysis of big data. Visualization also helps to understand easily the result of data analysis since complicated and scattered data would enable the visualization intuitively. R language is a good tool for computing the formulas quickly. Handing R program is a greate help to analyze the data since we could use already defined functions in R.

In this paper, Sales data of semiconductor will be handled due to increase of various mobile devices produced in the world. I aim to propose appropriate statistical methods that is able to compute at high speed in high dimension for analysis of the sales data of one semiconductor company, explain the result through graphs and search the pattern or trend. Particularly for searching the relation of attributes, valid proposed statistical methods and how result is produced are explained.

Applied core methods are clustering, logistic regression with lasso, **glasso** by Friedman et al. (2008) and **space** by Peng et al. (2009). From the result and visualized graph of the relation between variables in high dimension, we could get some pattern, trend and other issues more easier. In addition, application plan would be considered lastly.

The next is as the follows. Chapter 2 reviews the proposed methods that are clustering, logistic regression with lasso, decision tree (classification tree), inverse covariance estimation and partial correlation estimation. Chapter 3

describes properties about the Sales data of Mobile DRAM. In Chapter 4, the Sales data application using the methods suggested from Chapter 2 and result is covered. Chapter 5 provides conclusion and discussion. Lastly, reference is written in Chapter 6.

Chapter 2
Statistical methods

This Chapter reviews clustering methods in order to search the relation of data attributions. Described methods are logistic regression with lasso, inverse covariance estimation and partial correlation estimation.

2. 1. Clustering method

Clustering method has several hierarchical clustering methods. These methods are defined by the linkage criterion that determines the distance between sets of observation as a function of the pair-wise distances between all observations: If the minimum distance of two observations between clusters is used, it is defined as single (Sibson, 1973). If the maximum distance of two observations between clusters is used, it is defined as complete (Defays, 1977). If the average of all observations between clusters observation distances is used, it is the average method (Murtagh, 1984).

Euclidean distance is used as distances for clustering analysis. In p dimension, Euclidean distance is defined using two points x_i, y_i on equation (2.1.1).

$$d(x,y) = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{\frac{1}{2}} \tag{2.1.1}$$

In this study, I use the complete-linkage method with Euclidean distance and get the clusters by Dendrogram graph.

2. 2. logistic regression with lasso penalty

The logistic regression model arises from the desire to model the posterior probability of the K classes via linear functions in x, while at the same time ensuring that they sum to one and remain in [0,1]. The logistic regression model is

$$f(x)=\log\left(\frac{p(y=1|X=x)}{p(y=K|X=x)}\right)=\beta_{10}+\beta_1^T x.$$
$$f(x)=\log\left(\frac{p(y=2|X=x)}{p(y=K|X=x)}\right)=\beta_{20}+\beta_2^T x \qquad (2.2.1)$$
$$\cdots$$
$$f(x)=\log\left(\frac{p(y=K-1|X=x)}{p(y=K|X=x)}\right)=\beta_{(K-1)0}+\beta_{K-1}^T x$$

Regression function of logistic loss function is

$$l(\beta)=\sum_{i=1}^{N}(y_i\log p(x_i;\beta)+(1-y_i)\log(1-p(x_i;\beta))$$
$$=\sum_{i=1}^{N}y_i\beta^T x_i-\log(1+\exp(\beta^T x_i)) \qquad (2.2.2)$$

$$\text{where}\quad p(x)=\frac{\exp(f(x))}{1+\exp(f(x))} \qquad (2.2.3)$$

Estimation of $f(x)$ could be computed by MLE (Maximum likelihood Estimation).

Lasso is a shrinkage and selection method for linear regression. The lasso problem in the equivalent Lagrangian form

$$\widehat{\beta^{lasso}}=\arg min_\beta\left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}. \qquad (2.2.4)$$

L_1 lasso penalty $\sum_1^p|\beta_j|$ makes the solutions nonlinear in the y_i.

2. 3. Sparse inverse covariance estimation

I use graphical lasso algorithm for estimating sparse inverse covariance estimation is proposed by Friedman (2008). Suppose N multivariate normal observations of dimension p, with mean μ and covariance Σ. Let $\Theta = \Sigma^{-1}$, and S is the empirical covariance matrix. The problem is to maximize the Gaussian log-likelihood with the respect to the mean μ,

$$\log\det\Theta - tr(S\Theta) - \rho\|\Theta\|_1 \tag{2.3.1}$$

Θ is non-negative matrix and tr denotes the trace and $\|\Theta\|_1$ is the L_1 norm that is the sum of the absolute values of the elements of $\|\Theta\|_1$.

The problem (2.3.1) is convex and can be solved as optimizing over each row and corresponding column of W by Banerjee et al. (2007). Let W be the estimate of Σ.

They show that the solution for w_{12} satisfies

$$w_{12} = \arg min_y y^T W^{-1_{11}} y : \|y - s_{12}\|_\infty \le \rho \tag{2.3.2},$$

partitioning W and S.

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w^{T_{21}} & w_{22} \end{pmatrix},\ S = \begin{pmatrix} S_{11} & s_{12} \\ s^{T_{21}} & s_{22} \end{pmatrix} \tag{2.3.3}$$

Solving (2.3.2) is equivalent to solving the dual problem using convex duality

$$\min_\beta \frac{1}{2}\| W^{\frac{1}{2}_{11}}\beta - W^{\frac{1}{2}_{11}} s_{12}\|^2 + \rho\|\beta\|_1 \tag{2.3.4}$$

If β solves (2.3.4), then $w_{12} = W_{11}\beta$ solves (2.3.2).

(2.3.4) is expressed as a lasso regression and the equivalence between the solutions to (2.3.1) and (2.3.4) is verified as $W\Theta = I$.

Here is the Graphical lasso algorithm in detail :

1. Start with $W = S + \rho I$. The diagonal of W remains unchanged in what follows.
2. For each *j=1,2,...p,1,2,...p,...,* solve the lasso problem (2.3.4), which takes as input the inner products W_{11} and s_{12}. This gives a p-1 vector solution $\hat{\beta}$. Fill in the corresponding row and column of W using $w_{12} = W_{11}\hat{\beta}$.
3. Continue until convergence.

An R language package **glasso** is available to get inverse covariance matrix $\sum^{-1}$.

2. 4. Partial Correlation Estimation

For selecting nonzero partial correlations under the high dimension and low sample size, **space** (Sparse PArtial Correlation Estimation) is proposed as efficient approach by Peng et al. (2009). They show this method performs better than glasso in both model selection and hub identification.

Suppose that, $(y_{1,}\cdots,y_p)^T$ has a joint distribution with mean 0 and covariance Σ, where is a p by p positive definite matrix. Denote the partial correlation between y_i and y_j by ρ^{ij} $(1 \le i < j \le p)$ that equals to $Corr(y_i,y_j|y_{-(i,j)})$ or $Corr(\epsilon_i,\epsilon_j)$, where $y_{-(i,j)} = \{y_k : 1 \le k \ne i,j \le p\}$ and ϵ_i, ϵ_j are the prediction errors of the best linear predictors of y_i and y_j. (Lemma 1) is well-known result that relates the estimation of partial correlations to a regression problem.

Lemma 1 : For $1 \le i \le p$, y_i is expressed by $y_i = \sum_{j \ne i}\beta_{ij}y_j + \epsilon_i$, such that ϵ_i is not correlated with y_{-i} if and only if $\beta_{ij} = -(\frac{\sigma^{ij}}{\sigma^{ii}}) = \rho^{ij}\sqrt{(\frac{\sigma^{ij}}{\sigma^{ii}})}$. Moreover, $var(\epsilon_i) = (\frac{1}{\sigma^{ii}})$, $cov(\epsilon_i,\epsilon_i) = \frac{\sigma^{ij}}{\sigma^{ii}\sigma^{ij}}$.

They propose to estimate the partial correlations θ by minimizing a penalized loss function,

$$L_n(\theta,\sigma,Y) = \ell(\theta,\sigma,Y) + P(\theta) \tag{2.4.1}$$

where the penalty term $P(\theta)$ controls the overall sparsity of the final estimation of θ and the joint loss function $\ell(\theta,\sigma,Y)$

$$P(\theta) = \lambda\|\theta\|_1 = \lambda \sum_{1 \le i < l \le p} |\rho^{ij}|. \tag{2.4.2}$$

$$\ell(\theta,\sigma,Y) = \frac{1}{2}(\sum_{i=1}^{p} w_i \| Y_i - \sum_{j \ne i}\beta_{ij}Y_j\|^2) = \frac{1}{2}(\sum_{i=1}^{p} w_i \| Y_i - \sum_{i \ne j}\rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\ Y_j\|^2)\ , \tag{2.4.3}$$

where $\theta=(\rho^{12},...,\rho^{(p-1)p})^T$, $\sigma=\{\sigma^{ii}\}^{p_{i=1}}$; $Y=\{Y^k\}^{n_{k=1}}$; and $w=\{w_i\}^{p_{i=1}}$ are nonnegative weights.

The implementation of the **space** procedure is minimizing (2.4.1) under the ℓ_1 penalty (2.4.2).

First, they reformulate the problem (2.4.3) corresponds to the ℓ_2 loss of a "regression problem." and then use the active-shooting algorithm.

Active-shooting is motivated by the shooting algorithm (Fu 1998), which solves the lasso regression by updating each coordinate iteratively until convergence.

Suppose to minimize an ℓ_1 penalized loss function with respect to β

$$f(\beta)=\frac{1}{2}\|Y-X\beta\|^{2_2}+\gamma\sum_j|\beta_j|,$$

where $Y=(y_1,...,y_n)^T$, $X=(x_{ij})_{n\times p}=(X_1:...:X_p)$ and $\beta=(\beta_1,...,\beta_p)^T$.

The shooting algorithm procedure is in detail :

1. for *j=1,...,p,*

$$\beta_j^{(0)}=\arg min_{\beta_j}\left\{\frac{1}{2}\|Y-\beta_jX_j\|^2+\gamma|\beta_j|\right\}=sign(Y^TX_j)\frac{(|Y^TX_j|-\gamma)_+}{X_j^TX_j}, \tag{2.4.4}$$

where $(x)_+=xI_{(x>0)}$.

2. For *j=1,...,p,* update $\beta^{(old)}\rightarrow\beta^{(new)}$:

$\beta_i^{(old)}=\beta_i^{(new)}, i\neq j;$

$$\begin{aligned}\beta_i^{(new)}&=aremin_{\beta_j}\frac{1}{2}\|Y-\sum_{i\neq j}\beta_i^{(old)}X_i-\beta_jX_j\|^2+\gamma|\beta_j| \\ &=sign\left(\frac{(\epsilon^{(old)})^TX_j}{X_j^TX_j}+\beta_j^{(old)}\right)\left(\left|\frac{(\epsilon^{(old)})^TX_j}{X_j^TX_j}+\beta_j^{(old)}\right|-\frac{\gamma}{X_j^TX_j}\right)_+\end{aligned} \tag{2.4.5}$$

where $\epsilon^{(old)}=Y-X\beta^{(old)}$

3. Repeat step2 Until convergence

The shooting algorithm procedure is in detail :

1. Same as the 1 step of shooting.
2. Define the current active set A={k:current$\beta_k \neq 0$}.

(2.1) For each $k \in A$, update β_k with all other coefficients fixed at the current value as in Equation (2.4.5).

(2.2) Repeat (2.1) until convergence is achieved on the active set.

3. For j=1 to p, update β_j with all other coefficients fixed at the current value as in Equation (2.4.5). If no β_j changes during this process, return the current β as the final estimate. Otherwise, go back to step 2.

The choice of tuning parameter λ is proposed to use a "BIC-type" criterion because of its simplicity and computational easiness. In Yuan and Lin (2007), a BIC criterion is proposed for the penalized maximum likelihood approach.

$$BIC(\lambda) := \left[-\log|\widehat{\Sigma_\lambda^{-1}}| + trace(\widehat{\Sigma_\lambda^{-1}} S)\right] + \frac{\log n}{n} \times \#\{(i,j) : 1 \leq i \leq j \leq p, \widehat{\sigma_\lambda^{ij}} \neq 0\} \quad (2.4.6)$$

where S is the sample covariance matrix, and $\widehat{\Sigma_\lambda^{-1}} = \left(\widehat{\sigma_\lambda^{ij}}\right)$ is the estimator under λ. We simulate networks consisting of disjointed modules and denote their corresponding networks by Hub network with p nodes.

Chapter 3
Data Description

Sales data of semiconductor between January 2011 and 2013 is extracted from the salse history table. In this study, Mobile DRAM in the lots of semiconductor products is selected because of high sales ratio and in demand in the market. Selected variables representing the properties of the Mobile DRAM are type (catalogue), density and sales area in many attributes. Each variable is classified as 5 or 6 categorical data by appropriate grouping. The quantity of sales result is summarized from week unit to quarter unit, so that total unit of sales result for 3 years is 12. One quarter menas 4 months or 52 weeks. And the difference is large in sales quantity and there are many zero sales quantity.
Data has 74 by 12 under high dimension and low sample size problem. The number of variables is 74 with combination of type, density and area. The label of variables is renamed as abbreviation like T1,...,T5, D1,...,D5, A1,...,A6 by the order by proportion for security.

The followings is a graph describing product attribute's proportion :

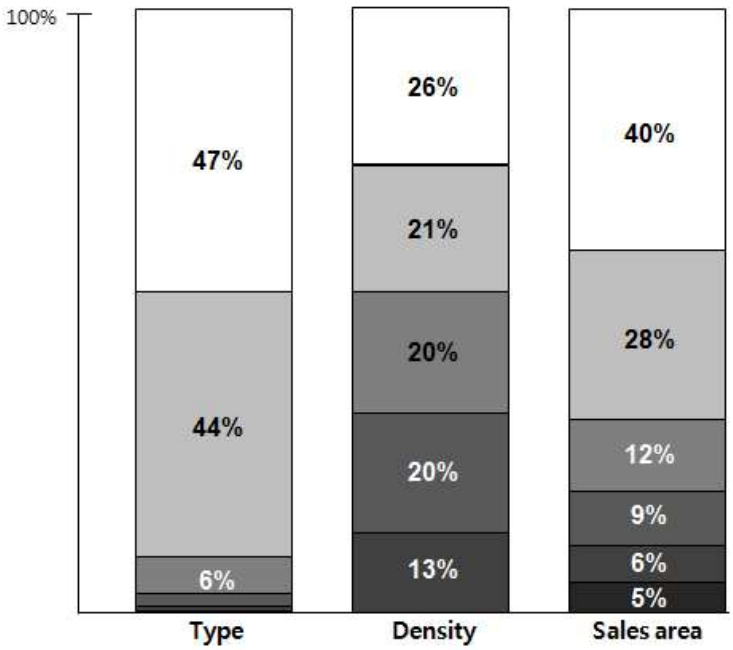

Figure 3.1 : graphs describing Sales data

Chapter 4
Application to Sales data

Chapter 4 shows that statistical methods proposed in Chapter 2 apply to sales data and we search the relation between product attributions and sales area of Mobile DRAM. Data matrix 12×74 is standardized by the mean and standard deviation.

First, I used the Euclidean distance and got the Dendrogram graph using complete-linkage clustering (Figure 4.1). The maximum distance of two observations between clusters makes bigger group than others, so complete-linkage clustering is better because of data issue with large difference in sales quantity between variables.

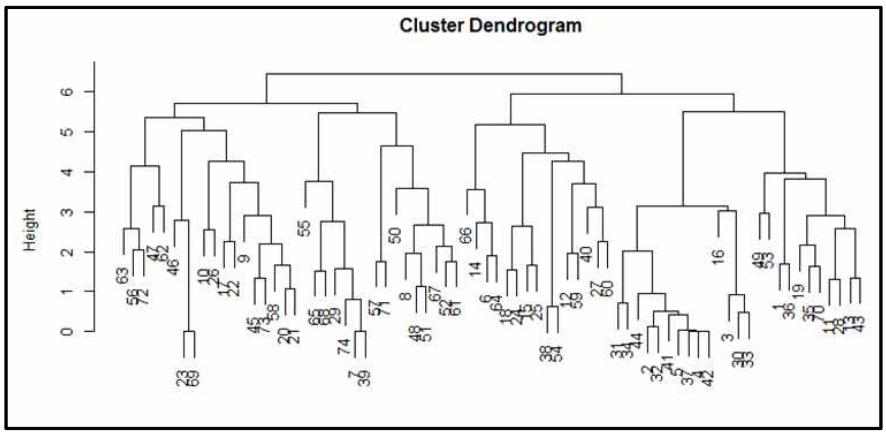

Figure 4.1 Clustering result with complete and Euclidean distance

5 clusters could be made from the above Dendrogram, and then using this cluster result, we apply the logistic regression with $l1$ penalty.
Raw data needs renewed setting for fitting the logistic regression model.: If Cluster is 1, the others is set as 0. If Cluster is 2, the others is set as 0. This process is repeated while fitting the 5 models.

The result by **glmnet** of R language package is the followings.

cluster	tendency of result
1	T5, T3, D5 < > T1, D1, A3, A4
2	T1, T3, D5 < > T4, D1, A2
4	T5, T3 D5 < > T2, A4

Table 4.1 The result of logistic regression with lasso penalty

In case of 3 or 4 cluster, all coefficients except intercept were zero.

In addition, the result between the result fitting classification tree model and logistic regression model is very similar. Figure 4.2 is the result of the classification tree.

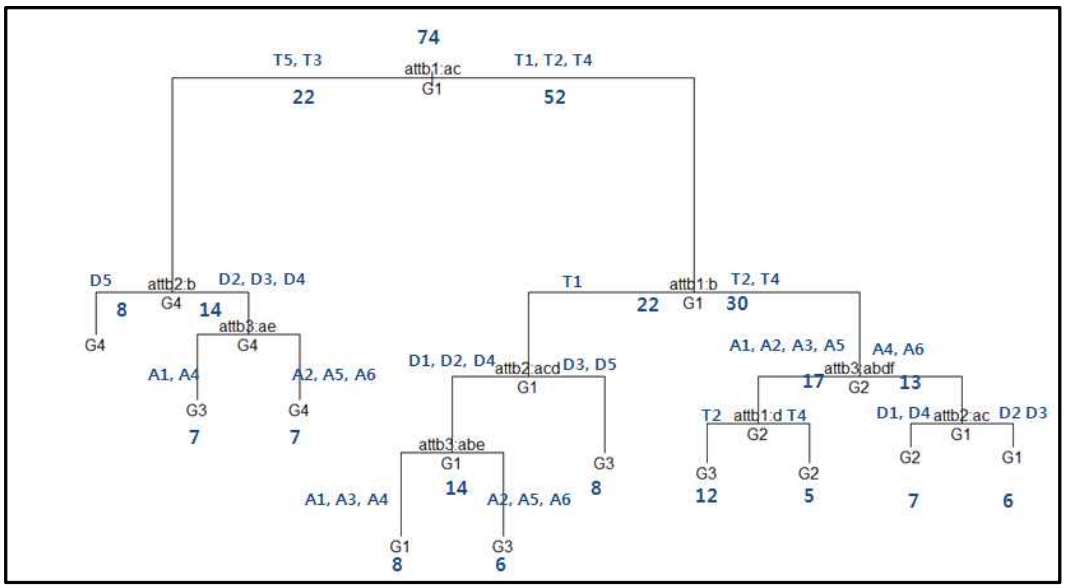

Figure 4.2 Clustering result of classification tree

In figure 4.2 the first node is divided by T5, T3 and T1,T2,T4. For reference the number near nodes is the number of separated variables. If pruning was added to decision tree, the result is more simpler and the number of nodes

decrease.

Another method is the hub network graph for searching the relation of variables. Variables are the combination of type, density, sales area and data matrix has 74 variables while making hub network.

First, I seek to find the inverse covariance matrix using **glasso** (by Friedman et al. (2008)) package in R, and control a λ that is a parameter of penalty term. The method of tunning parameter in this study is the BIC-type (Yuan and Lin (2007)) due to its simplicity and computational easiness.

	λ sequence	Total glasso	BIC glasso
1	0.04000000	922.0	290.2273
2	0.04515352	906.0	291.2080
3	0.05097100	884.0	289.6858
4	0.05753800	856.0	285.6727
5	0.06495107	842.0	287.4350
6	0.07331923	814.0	283.3857
7	0.08276552	788.0	280.1316
8	0.09342886	767.0	278.9206
9	0.10546604	737.0	273.9375
10	0.11905406	716.0	272.6361
11	0.13439273	686.0	267.6010
12	0.15170761	658.0	263.3677
13	0.17125330	621.0	255.5428
14	0.19331721	588.0	248.8712
15	0.21822379	537.0	234.8131
16	0.24633928	520.0	234.7589
17	0.27807712	510.0	237.6897
18	0.31390399	482.0	233.1069
19	0.35434672	472.0	235.9805
20	0.40000000	449.0	233.5949

Figure 4.2 Inverse covariance matrix by tuning parameter method “BIC-type”

In figure 4.2, λ sequence is the parameter value λ in penalty term. Total glasso is the number of the half of nonzero off-diagonal that means total edges. Therefore 0.04515352 in the second row is selected as the value of λ by BIC-type when the value of BIC glasso is a maximum.
Second, the partial correlation matrix of the precision matrix (inverse covariance matrix) is computed with the following condition (4.1) for sparsity. 1.15 is the appropriate value for separating completely connected variables.

$$\text{abs (estimation of precision matrix)} > 1.15 \qquad (4.1)$$

Finally, the Hub network graph is created by the partial correlation matrix using **igraph** and **space** package in R. The sum of each variables in the partial correlation matrix is defined as degree, so that the hub is adjusted by degree. The larger the number of degree is, the sparser the graph is. I controled the value of degree watching the graph shape for better judgment.

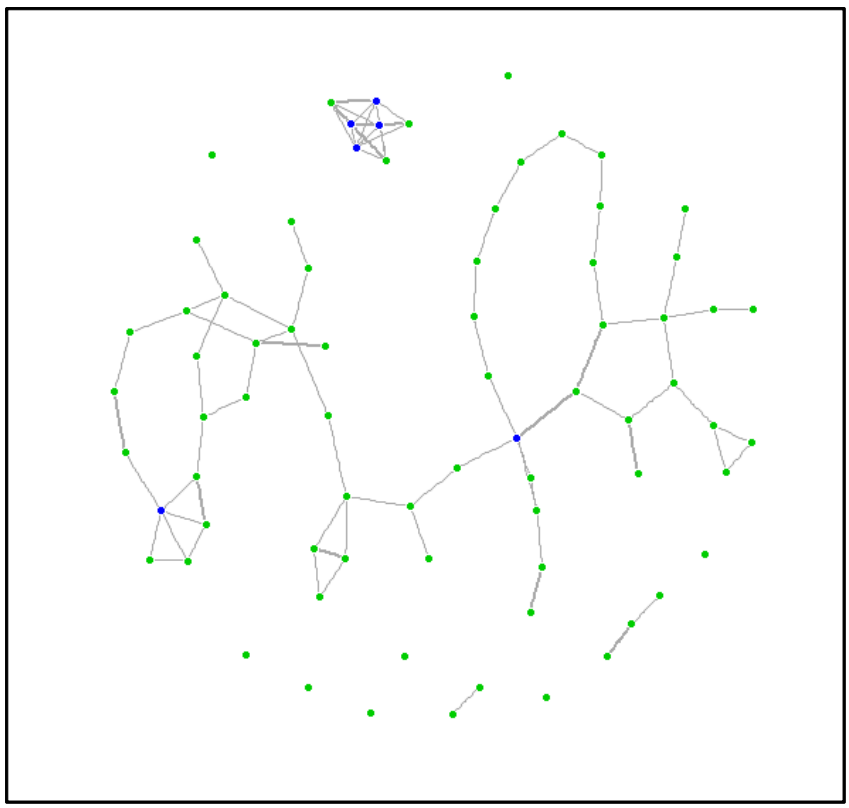

Figure4.3 hub network graphs with degree > 4 using glasso and space

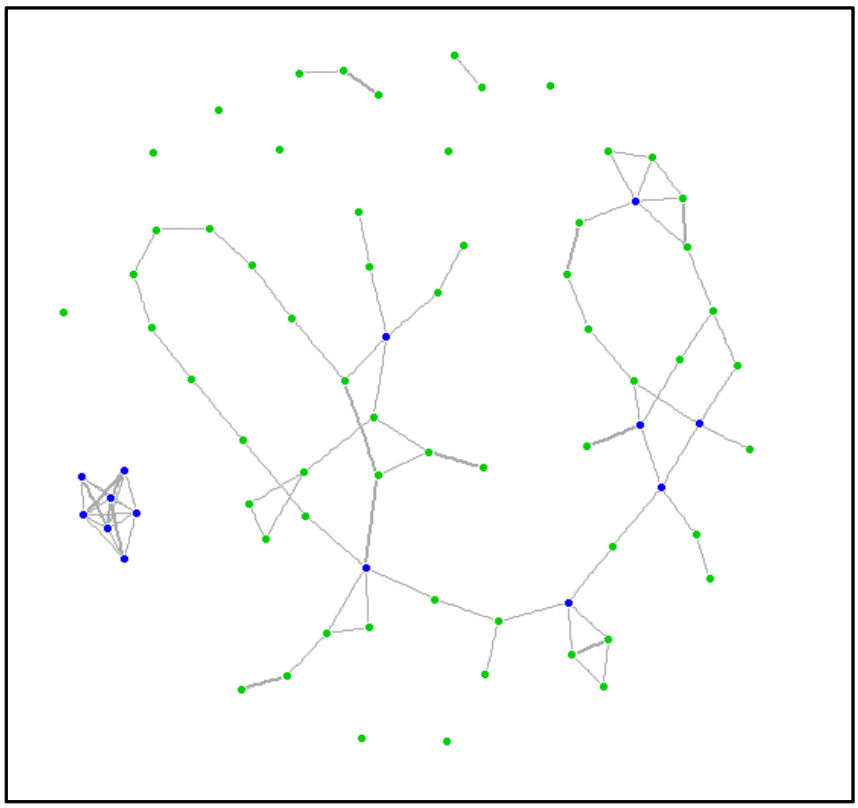

Figure4.4 hub network graphs with degree > 3 using glasso and space

In figure 4.3, the total number of each node is 74 variables. Several dark nodes are called hub site that holds the greatest influence among nodes. Bold lines means high weight between variables. I will seek to search hub site from the hub network with degree>4 due to easy understanding and interpretation of clustering. The difference of two graph in figure 4.3 is the number of hub site and edges between nodes.

The following table 4.2 is the data of hub network in detail. Cluster in title is just random numbering.

cluster	hub site	connected node list
1	T3 D2 A5 T3 D4 A6 T5 D4 A1 T5 D2 A6	T3 D5 A6 T3 D2 A6 T5 D5 A4
2	T2 D3 A4	T2 D3 A3 T2 D2 A2 T1 D2 A1 T1 D2 A5
3	T1 D3 A2	T3 D5 A1 T3 D5 A4 T3 D5 A5 T4 D2 A2 T5 D5 A2
4		T1 D3 A4 T1 D5 A2 T1 D5 A6 T3 D4 A2 T5 D2 A2

Table 4.2. The detail core result of hub network graph

Clusters are distinguished cluster T3, T5, A6, cluster T2, D2 cluster T3, D5 and cluster T1, A2 mainly from hub network. cluster 1 is completely disconnected from other clusters. The distance between cluster 2 and cluster 3 is far, so it means few correlation. Cluster 2 and 3 is connected closely.

We could consider the connected relation of each variable, disconnected group and which node has the largest influence.

The principal trend and pattern among methods given above is under the same. The methods applied Sales data of semiconductor are different but has the similar result.

Chapter 5
Conclusion and Discussion

We analyze sales data of Mobile DRAM in semiconductor using several methods proposed in Chapter 4. To get information about the relation of product attributes, fitted methods were complete-linkage clustering, classification tree, logistic regression with lasso and sparse correlation matrix estimation. The result data was visualized and explained by Dendrogram, tree and hub network graph. In conclusion, the result among different methods was equivalent and has the pattern by type, density and sales area of product.

Next, I will seek to utilize the result that is clustering, connected relation with nodes and edges in hub network. For example, diffusion marketing could be considered in marketing and sales strategy. Diffusion marketing that new products are accepted into a community or market. While releasing new product, company could hold a promotion in hub site in priority and spread sales marketing to the closest connected in hub site sequentially watching a response of core node. In other words we could make some groups that are bound together by the same type or sales area in hub network. New product could be put to a node having weigh heavily or marketing strategy could be.

Later, if we try further analysis about Sales data using new methods or renewed variables and compare the results, then it will help to reduce the risk of sales prediction in marketing and get higher profit.

Chapter 6
Reference

Banerjee, O., Ghaoui, L. E. & d'Aspremont, A. (2007), 'Model selection through sparse maximum likelihood estimation', *To appear, j. Machine learning Research* 101.

D. Defays (1977). An efficient algorithm for a complete link method. The Computer Journal (British Computer Society) 20 (4): 364 – 366.

Friedman, J.,Hastie, T., and Tibshiranik R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432-441.

Fu, W. (1998), "Penalized Regressions: The Bridge vs the Lasso," J*ournal of Computational and Graphical Statistics*, 7(3), 397-416.

Murtagh F (1984). Complexities of Hierarchic Clustering Algorithms: the state of the art. Computational Statistics Quarterly 1: 101 – 113.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735-746.

R. Sibson (1973), SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal (British Computer Society) 16 (1): 30 – 34.

Yuan, M., and Lin, Y. (2007), "ModelSelection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94(1), 19-35

국문초록

전 세계의 모바일 디바이스 (스마트 폰, 테블릿 피씨 등) 제품의 수요 증가는 반도체 모바일 디램의 판매 증가로 연결된다. 이 논문에서는 분석 대상인 반도체의 판매 데이터에 적용할 적절한 통계적 방법론을 제안하고 모바일 디램 제품의 종류와 용량이 판매지역별로 어떤 영향 관계에 있는지 분석하여 그 결과를 이용한 새로운 아이디어를 얻고자 한다. 사용된 통계 방법은 군집분석, 로지스틱 회귀분석, 라쏘(lasso), 의사결정 나무, 부분 상관계수 추정이다. 부분 상관계수 추정에 접근하기 위해서는 공분산 역행렬(= precision matrix)에 lasso penalty가 적용된 **glasso** (graphical lasso)를 이용한다. 그리고 **space** (Sparse Partial Correlation Estimation)을 이용하여 허브 네트워크 (hub network) 그래프를 그리고 제품 판매 트렌드와 패턴을 파악해본다.

.....

주요어 : 군집분석 방법론, 공분산 역행렬, 부분 상관계수 추정, 허브 네트워크
학 번 : 2012-23015