Gaussian Copula Marginal Regression Modeling for Technology Analysis

Sunghae Jun

***Abstract*—Understanding technological relations between patent technology keywords is an important task for building research and development (R&D) policy of nation and company. Many researches have been actively conducted on this research subject, and various approaches to technology analysis were studied in the field of technology management. Most of the methods of technology analysis were based on patent documents related to target technology, because patent contains diverse information on developed technologies. So the patent keywords extracted from patent documents are valuable sources for technology analysis. The structured patent data become a matrix consisting of patent (row) and keyword (column), and each element of the matrix is frequency value of the keyword occurred in each patent. In this paper, we propose a method of technology analysis using Gaussian copula marginal regression (GCMR) model, and use the R data language for patent analysis by the GCMR. In addition, we carry out a case study to show how this study could be applied to real problem. This research contributes to various R&D planning of nation and company.**

***Index Terms*—Technology analysis, Gaussian copula marginal regression, patent keyword data, technology management, statistical model.**

I. Introduction

Gaussian copula regression is a statistical model to analyze various data sources such as time series, longitudinal, or spatial data [1]. In addition, the likelihood inference has been used for Gaussian copula models [2]. The likelihood function is possible to apply for continuous data, but it is difficult to apply for discrete data [3]. In general, we have to deal with discrete data source in patent technology analysis, because we preprocess the patent documents for statistical analysis. The preprocessed patent document data becomes a structured data type (matrix) consist of patent as row and keyword as column [4]. Each element of the matrix represents the occurred frequency value of keyword in a patent. The matrix has sparsity problem, that is, most frequency values of the matrix are zero values [5]. This sparseness is also a problem of extreme values. To perform patent technology analysis, we have to solve this obstacle. In this paper, we analyze the patent keyword matrix to understand the technological relations between technological keywords. Regression analysis is one of popular methods to find the relations between response and explanatory variables [6]. But this has a limitation to control the sparse and extreme values [7]. To overcome the limitation of linear regression analysis, Masarotto and Varin (2012) proposed the Gaussian copula marginal regression (GCMR) [8]. This method considers Gaussian copula model and marginal regression analysis. In addition, they developed an R package called 'gcmr' [1]. The R is an efficient data language and its diverse packages are based on R data language like 'gcmr' [9], [10]. This package provides diverse functions for GCMR such as model fitting or diagnostics plotting. In the GCMR, the model parameters are estimated the maximum likelihood and maximum simulated likelihood methods for continuous and discrete data types [1]. In general, we use various evaluation measures such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and maximum log-likelihood to evaluate the performance of the GCMR model [1], [6]. Also, the 'gcmr' package provides the function of residual analysis to check the model assumption [1]. In this paper, we use the 'gcmr' package and R data language to analyze the patent data related to target technology. We carry out a case study to illustrate how our model could be applied to real domain. We collect the patent documents related to light-emitting diode (LED) technology for our case study. Using the GCMR, we analyze the extracted keywords of LED technology and make technological structure for understanding LED technology. The remainder of this paper is organized as follows. Section II introduces the patent technology analysis using GCMR. We show a case study in Section III. Lastly the conclusions are presented in Section IV.

Manuscript received January 5, 2018; revised May 1, 2018.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03031152).

S. Jun is with Cheongju University, Chungbuk, 28503 Korea (e-mail: statcs@gmail.com).

II. Technology Analysis using Gaussian Copula Marginal Regression

There are so many results related to developed technology in patent document, because the patent system protects the inventor's exclusive rights to the developed and registered technology over a period of time [11], [12]. So, many existing works with regards to technology analysis have performed by analyzing patent data [11], [13], [14], [15]. In general, a number of factors must be considered in order to analyze patent documents, because patent data consists of various heterogeneous data types such as texts, numbers, dates, and pictures [11]. In order to analyze patent data by statistical methods such as the GCMR, we have to transform the retrieved patent documents into structured data that can be analyzed by statistics. In this paper, we use text mining techniques for transforming patent document data. We use

DOI: 10.7763/IJMO.2018.V8.640

'tm', a text mining package provided by R and R data language to build the structured data [9], [16], [17]. From the result of preprocessing by text mining, the structured data consist of the extracted keywords and their occurred frequency values in patent documents as follow.

	m keywords
n patents	n×m frequency values

Fig. 1. Keyword matrix for GCMR.

Fig. 1 shows our data structure for the GCMR modeling. In the (n×m) frequency values, there are so many zero values. So the matrix has sparsity problem, and we use GCMR model to overcome the sparseness problem of the patent keyword matrix. To perform technology analysis using the GCMR model, we use the patent keywords because the keywords represent technological aspects of given technology fields. In our study the keywords are considered to be variables used in the GCMR model. Also, the whole keywords are divided into response variable and explanatory variables.

In this paper, $(Y_1, Y_2, \ldots, Y_n)$ is a frequency vector of response keywords, and n is the number of collected patents. The frequency matrix of explanatory keywords is as follow [8].

$$X = (x_{i1}, x_{i2}, \ldots x_{ip}) \qquad i = 12, \ldots n, \tag{1}$$

Where p is the number of explanatory keywords. Also the density function of x_i and Y_i is expressed as follow [8].

$$f(Y_i \mid x_i; \beta) \qquad i = 12, \ldots n, \tag{2}$$

The function is denoted by conditional formula of Y_i given x_i, and β is a model parameter vector. In the GCMR, conditional expectation of Y_i given x_i is expressed as follow [8].

$$E(Y_i \mid x_i) = \mu_i \qquad i = 12, \ldots n, \tag{3}$$

We build our GCMR model using various link functions as follow [1], [8].

$$L(\mu_i) = x_i \beta \qquad i = 12, \ldots n, \tag{4}$$

The link function $L(\cdot)$ can be of various distributions as follows [8]; Beta, Binomial, Gamma, Gaussian, Negative Binomial, Poisson, and Weibull. Using this general model, we can find the technological relations between response (Y_i) and explanatory (X) keywords. We express the general equation of Y_i given X as follow [8].

$$Y_i = F_i^{-1}(\Phi(\varepsilon_i) \mid X; \beta) \qquad i = 12, \ldots n, \tag{5}$$

Where error ε_i has a standard normal distribution, and $F_i(\cdot)$ is cumulative distribution of Y_i (response keyword) given X (explanatory keyword vector). $\Phi(\cdot)$ is cumulative standard normal distribution of ε_i. In this paper, we separate the copulas to the marginal of regression from the response component. The proposed process of technology analysis using GCMR is shown in Fig. 2.

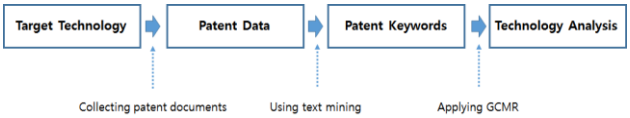

Fig. 2. Technology analysis process using GCMR.

Once the target technology of interest is determined, we first collect patent documents related to the technology from the worldwide patent databases such as WIPS Corporation (WIPSON) and the United States Patent and Trademark Office (USPTO) [19], [19]. Using the text mining techniques, we preprocess the collected patent data and extract patent keywords from the patent data. In this phase, we select and use 'tm' package of R data language for the preprocessing based on text mining techniques [16]. Thus, the preprocessed patent data consists of a matrix in which the patent document and the keyword are row and column, respectively. In the all keywords, we choose response keyword and explanatory keywords. The GCMR model to explain how the explanatory keyword describes the response keyword. Finally, we get the technological relations of the patent technology keywords, because a patent keyword represents a corresponding sub-technology. Using the results of our process, nation and company can build the R&D plan for their technological competition. A more detailed description used in the actual technological field is described in the following section.

III. Case Study

We performed a case study to show how our GCMR modeling could be applied to real problem. We also collected all patent documents related to the LED technology. To collect the patent documents, we used the patent databases of WIPSON [18]. The technology of LED is important area in

emerging technology group [7]. So the analytical researches related to LED technology have been studied in various fields [20]. Using various text mining techniques [17], we extracted technological keywords related to LED technology from collected patent document data. We used the 'LED' keyword as the response variable in the extracted keywords and the remaining keywords as the explanatory variables. The keywords used for explanatory variables are as follow; 'Control', 'Lamp', 'Circuit', 'Power', 'Device', 'Layer', 'Signal', 'Wireless', 'Material', 'Heat', 'Module', 'Display', 'Chip', 'Remote', 'Driving', 'Supply', 'Surface', 'Voltage', 'Board', 'Body', 'Communication', 'Organic', 'Intelligent', 'Illumination', 'Color', 'Plate', 'Switch', 'Screen', 'Energy', and 'Optical'. In addition, we divided the entire patent into three groups according to the region where the patent was applied and filed. The regions are China (CN), Europe (EP), and the United States (US). Table I shows the results of patent analysis by the GCMR model.

TABLE I: PARAMETER ESTIMATION OF GCMR MODEL

Keyword	Estimate	Std. Error	Z Statistic	p-value
(Intercept)	2.9091	0.1034	28.1320	0.0000
Control	0.0897	0.0204	4.3910	0.0000
Lamp	0.3555	0.0165	21.5100	0.0000
Circuit	0.1207	0.0195	6.1960	0.0000
Power	0.0828	0.0315	2.6270	0.0086
Device	0.0459	0.0272	1.6920	0.0906
Layer	-0.1180	0.0224	-5.2580	0.0000
Signal	-0.0296	0.0322	-0.9190	0.3583
Wireless	0.1235	0.0375	3.2970	0.0010
Material	-0.1419	0.0401	-3.5440	0.0004
Heat	0.1173	0.0339	3.4560	0.0005
Module	0.1461	0.0192	7.6050	0.0000
Display	0.0168	0.0338	0.4950	0.6205
Chip	0.4577	0.0331	13.8410	0.0000
Remote	0.0435	0.0432	1.0080	0.3137
Driving	0.1367	0.0427	3.2050	0.0014
Supply	0.0184	0.0570	0.3230	0.7466
Surface	0.0903	0.0470	1.9210	0.0547
Voltage	0.0929	0.0447	2.0760	0.0379
Board	0.1082	0.0407	2.6600	0.0078
Body	-0.0636	0.0484	-1.3140	0.1890
Communication	0.2104	0.0537	3.9210	0.0001
Organic	-0.4674	0.0498	-9.3880	0.0000
Intelligent	0.1529	0.0487	3.1370	0.0017
Illumination	0.1599	0.0540	2.9610	0.0031
Color	0.2536	0.0525	4.8280	0.0000
Plate	0.2124	0.0495	4.2900	0.0000
Switch	-0.0517	0.0560	-0.9240	0.3557
Screen	0.3697	0.0560	6.6060	0.0000
Energy	-0.0107	0.0621	-0.1720	0.8635
Optical	-0.1042	0.0573	-1.8200	0.0688
NationEP	-1.2688	0.2429	-5.2230	0.0000
NationUS	-1.0909	0.1246	-8.7550	0.0000

The AIC and maximum log-likelihood values of this result are 21,713 and 10,822 respectively. We also selected the statistically significant keywords with p-values less than 0.05. This is 95% confidence level. The statistically significant keywords are as follows; 'Control', 'Lamp', 'Circuit', 'Power', 'Layer', 'Wireless', 'Material', 'Heat', 'Module', 'Chip', 'Driving', 'Voltage', 'Board', 'Communication', 'Organic', 'Intelligent', 'Illumination', 'Color', 'Plate', and 'Screen'. Therefore, in order to develop LED technology, it is necessary to develop or secure technologies corresponding to the significant keywords. In addition, according to the regions, the technologies related to LED are different because the p-values of 'NationEP' and 'NationUS' are less than 0.05. We carried out the evaluation of our GCMR model using various diagnostic plots. Fig. 2 shows the plotting results.

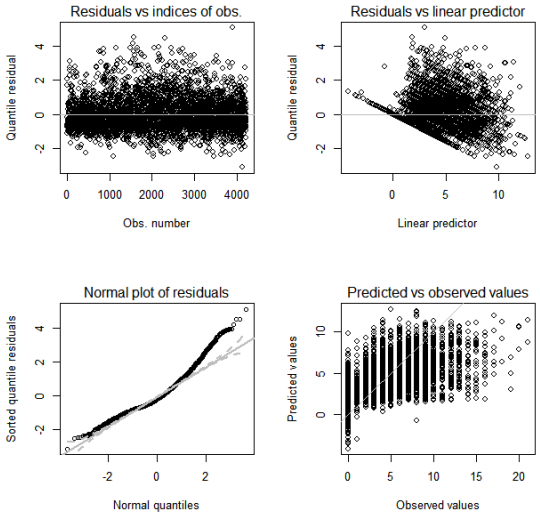

Fig. 2. Diagnostic plots for evaluating GCMR model.

Fig. 2 shows the results of residual analysis and test of normality assumption. From the residual plots of the top 2 graphs in Fig. 2, we confirmed the randomness of residuals. In addition, we illustrate the satisfaction of the normality assumption by Q-Q (quantile-quantile) plots in the Fig. 2 on the below. We therefore show that the GCMR model is an effective approach for analyzing patent keywords with frequency data in this paper. We provided the technological affections between sub-technologies by the GCMR model based on 'gcmr' package of R data language. Using this technological relationship between the patent technology keywords, our research contributes to the R&D planning for companies and nations. Therefore, we conclude that the prior technologies to develop the LED technology are represented as Fig. 3.

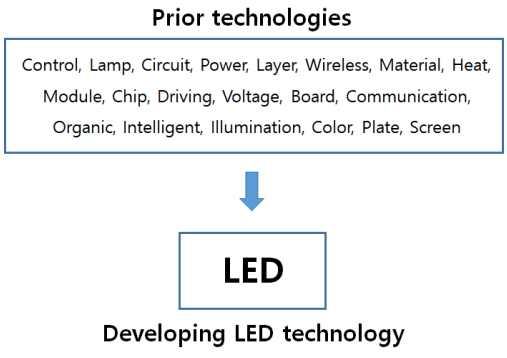

Fig. 3. Diagnostic plots for evaluating GCMR model.

In Fig. 3, we can conclude the technologies based on the patent keywords of 'Control', 'Lamp', 'Circuit', 'Power', 'Layer', 'Wireless', 'Material', 'Heat', 'Module', 'Chip', 'Driving', 'Voltage', 'Board', 'Communication', 'Organic', 'Intelligent', 'Illumination', 'Color', 'Plate', and 'Screen' become antecedent technologies to develop the LED technology. These results can contribute to the development of technology development plans for LED companies.

IV. Conclusions

In this paper, we proposed a method of technology analysis to make discrete data model for patent keyword analysis. This is because most patent data are made up of frequency values of occurred keywords in patent documents after preprocessing the collected patent documents. Moreover, we encounter the sparsity problem in patent technology analysis because most frequency values of patent keyword matrix preprocessed by text mining techniques are zeros. To solve this problem of patent analysis for understanding technology, we considered the GCMR modeling. This modeling deals with marginal regression analysis and Gaussian copula model at the same time. In addition, we carried out a case study to illustrate how our methodology could be applied to real domain. We selected the LED technology as the target technological filed for our case study. This paper contributes to R&D planning in company or nation. In our future work, we will consider more advanced statistical model to overcome various problems in technology analysis.

References

[1] G. Masarotto and C. Varin, "Gaussian copula regression in R," *Journal of Statistical Software*, vol. 77, no. 8, pp. 1-26, 2015.

[2] A. K. Nikoloulopoulos, "Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses," *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 2, pp. 493-505, 2016.

[3] M. Pitt, D. Chan, and R. Kohn, "Efficient Bayesian inference for Gaussian copula regression models," *Biometrika*, vol. 93, pp. 537-554, 2006.

[4] S. Jun, S. Park, and D. Jang, "Technology forecasting using matrix map and patent clustering," *Industrial Management & Data Systems*, vol. 112, pp. 786-807, 2012.

[5] S. Jun, S. Park, and D. Jang, "Document clustering method using dimension reduction and support vector clustering to overcome sparseness," *Expert Systems with Applications*, vol. 41, pp. 3204-3212, 2014.

[6] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists, Fourth Edition*, Seoul, Korea, Elsevier, 2012.

[7] H. K. Sutikno and I. D. Ratih, "Gaussian copula marginal regression for modeling extreme data with application," *Journal of Mathematics and Statistics*, vol. 10, no. 2, pp. 192-200, 2014

[8] G. Masarotto, and C. Varin, "Gaussian copula marginal regression," *Electronic Journal of Statistics*, vol. 6, pp. 1517-1549, 2012.

[9] R Development Core Team. [Online]. Available: http://www.R-project.org

[10] G. Masarotto and C. Varin, *'Gcmr' Package - Gaussian Copula Marginal Regression*, R package version 1.0.0, CRAN, R-Project, 2017.

[11] D. Hunt, L. Nguyen, and M. Rodgers, *Patent Searching Tools & Techniques*, Hoboken, NJ, Wiley, 2007.

[12] A. T. Roper, S. W. Cunningham, A. L. Porter, T. W. Mason, F. A. Rossini, and J. Banks, *Forecasting and Management of Technology*, Hoboken, NJ, John Wiley & Sons, 2011.

[13] S. Jun, D. Jang, and S. Park, "Patent management for technology forecasting: A case study of bio-industry," *Journal of Intellectual Property Rights*, vol. 17, no. 6, pp. 539-546, 2012.

[14] J. Kim and S. Jun, "Graphical causal inference and copula regression model for apple keywords by text mining," *Advanced Engineering Informatics*, vol. 29, pp. 918-929, 2015.

[15] J. Kim, D. Im, and S. Jun, "Factor analysis and structural equation model for patent analysis: A case study of Apple's technology," *Technology Analysis & Strategic Management*, vol. 29, no. 7 pp 717-734, 2017.

[16] I. Feinerer and K. Hornik, *'tm' Package - Text Mining*, R package version 0.7.3, CRAN, R-Project, 2017.

[17] I. Feinerer, K. Hornik, and D. Meyer, "Text mining infrastructure in R," *Journal of Statistical Software*, vol. 25, no. 5, pp. 1-54, 2008.

[18] WIPSON. [Online]. Available: http://www.wipson.com

[19] USPTO. *The United States Patent and Trademark Office*. [Online]. Available: http://www.uspto.gov

[20] S. Park and S. Jun, "Technology analysis of global smart light emitting diode (LED) development using patent data," *Sustainability*, vol. 9, pp. 1363, 2017.

Sunghae Jun is a professor in the Department of Statistics, Cheongju University, Chungbuk, Korea. He received B.S., M.S., and PhD degrees from Department of Statistics, Inha University, Incheon, Korea in 1993, 1996, and 2001, respectively. He also received PhD degree from Department of Computer Science, Sogang University, Seoul, Korea in 2007, and PhD from Information Management Engineering from Korea University, Seoul, Korea in 2013. He was visiting scholar in Department of Statistics, Oklahoma State University, Stillwater, Oklahoma, USA from 2009 to 2010. His current research interests include big data learning and technology forecasting