Integration of Graphical Modeling with Fuzzy Clustering for Casual Relationship of Electric Load Forecasting

Hiroyuki Mori, Wenjun Jiang
Dept. of Electronics & Bioinformatics
Meiji University
hmori@isc.meiji.ac.jp

Abstract

This paper proposes a new method for selecting input variables in short-term electric load forecasting. It is known that input and output variables do not follow the Gaussian distribution in load forecasting. In this paper, a hybrid method of Graphical Modeling (GM) and Deterministic Annealing Expectation Maximization (DAEM) clustering is presented to clarify causal relationship between the explained one-step-ahead electric load and the explanatory variables. GM is effective for estimating the relationship between variables with the Gaussian distribution. The DAEM algorithm is used to decompose non-Gaussian data into clusters of Gaussian data so that GM is applied to Gaussian data in clusters. The proposed method is successfully applied to the real data.

1. Introduction

This paper proposes a method for evaluating the causal relationship between the explained and the explanatory variables in short-term electric load forecasting. A hybrid method of graphical modeling (GM) and Deterministic Annealing Expectation Maximization (DAEM) clustering is used to provide more realistic relationship. In power system operation, short-term electric load forecasting is very important to smooth Economic Load Dispatching (ELD), unit commitment, *etc.*[1]. Recently, the degree of uncertainty increases due to the emergence of deregulated and competitive power market. As a result, the power system players are interested in maximizing a profit while minimizing a risk in power systems. However, it is more difficult to understand the variation factor of the electric load due to the complexity of power markets and networks[2-4]. Therefore, it is necessary to clarify the causal relationship between the explained and explanatory variables and to enhance the model accuracy in load forecasting.

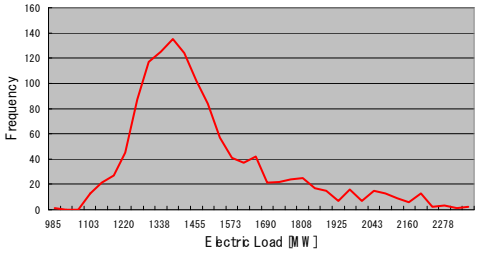

Fig. 1. Frequency Distribution of Electric Hourly Load of Capital (NYISO) at 2 p.m.

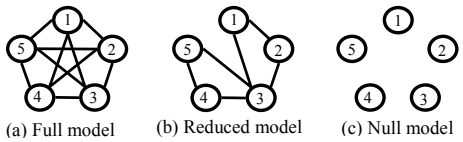

Fig. 2. Network Models

In this paper, a GM-based method with DAEM clustering is presented to deal with the casual relationship of non-Gaussian data. GM is one of statistical multivariate analysis techniques that play a role to clarify the causal relationship of model variables. GM selects the covariance of variables under the assumption that they follow multivariate normal distribution[5]. As a result, the case of handling non-Gaussian data as shown in Fig. 1 is an open problem. This paper proposed a new method to extend the conventional GM to handle non-Gaussian data. This paper employs a fuzzy clustering technique with the DAEM algorithm to cluster the multi-dimensional data through a multivariate mixture normal distribution (MMND) model. The method is successfully applied to real data of the New York Independent System Operator (NYISO).

2. Graphical modeling

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.135

IEEE computer society

This section describes GM that selects the explanatory variables from the candidates[5]. It is known for one of the statistical multivariate analysis techniques that clarifies the causal relationship of variables.

2.1. Network models

As shown in Fig. 2, there exist three kinds of expressions of a network model. The full network is a model with the maximum number of edges, and the null model is minimum one. The reduced model is a network in the middle of them in a way that the edges between variables without relationship are removed. GM is employed to construct the reduced model in understanding a meaningful relationship between nodes that are regarded as the explained or candidates of explanatory variable while filtering out the negligible variables. A direct graph is given through the possibilities of cause-effect of variable pairs, sequence or temporality.

2.2. Partial correlation coefficient

The partial correlation coefficient is the basic concept in GM of a quantitative variable. Although the correlation coefficient shows the relationship of variable pairs quantitatively, it does not show the realistic correlation in a case where an indirect influences exists with a hidden common cause between them. Suppose that the hidden common exists. Since two variables share the correlation with the third one, there is a possibility that the correlation of two variables is affected by the rest. This is called the spurious correlation or spurious association[6]. The partial correlation coefficient differs from the correlation one. It is important to remove the influence of spurious correlation in evaluating the relationship of a couple of variables. Now, let us define the correlation coefficient matrix and its inverse one as $\boldsymbol{\Pi} = (\rho_{ij})$ and $\boldsymbol{\Pi}^{-1} = (\rho^{ij})$ respectively. The partial correlation coefficient may be written as

$$\rho_{ij \cdot rest} = -\frac{\rho^{ij}}{\sqrt{\rho^{ii}}\sqrt{\rho^{ii}}} \tag{1}$$

where

ρ^{ij} : coefficient of inverse matrix of correlation coefficient matrix

$\rho_{ij \cdot rest}$: coefficient of the partial correlation coefficient

2.3. Covariance selection

The covariance selection proposed by Dempster is useful for the covariance selection model[7]. It makes use of the multi-dimensional data that follows a multivariate normal distribution (MND) and assumes the conditional independence of other variables to be fixed. The partial correlation coefficients of variable pairs with a small absolute value often appear after the partial correlation coefficients were obtained from the sample data. Such tiny coefficients should be removed from a standpoint of the principle of parsimony in statistics. To remove the coefficients, the negligible variables are assumed to be conditional independent, and replaced with zero in the covariance selection.

To fit a covariance selection model systematically, this paper focuses on the cyclic fitting algorithm proposed by Wermuth and Scheidt to estimate the parameters of the covariance structure of a MND[8]. This algorithm settles the partial correlation coefficients with the conditional independence to zero sequentially while estimating other parameters. Speed and Kiiveri proved that the solution of this algorithm converged in maximum likelihood estimate (MLE)[9].

3. DAEM clustering

Deterministic Annealing Expectation Maximization (DAEM) algorithm[10] is the modified version of the EM algorithm[11] that carries out the maximum likelihood estimation for system with unobservable hidden variables. Since the input data of GM obeys a MND, the paper makes use of data clustering with DAEM algorithm as a preprocessing to divide population of input data into the subsets to which GM is applicable.

3.1. EM algorithm

The EM algorithm is widely used for fitting the finite mixture model from incomplete data by the maximum likelihood[12]. It has been successfully employed in statistical learning, clustering and data communication[13-14]. Let $\boldsymbol{y}$ be a d- dimensional data. The mixture model may be written as

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m f_m(\boldsymbol{y} \mid \boldsymbol{\theta}_m) \tag{2}$$

where

$f_m(\boldsymbol{y} \mid \boldsymbol{\theta}_m)$: probability density function (PDF) conditional on $\boldsymbol{\theta}_m$

π_m : mixing probabilities, $\pi_m \geq 0 \ \sum_{m=1}^{M} \pi_m = 1$

$\boldsymbol{\theta}_m$: parameters of m-th component

$\boldsymbol{y}$: incomplete data, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$

According to the multivariate mixture normal distribution, function f may be written as

$$f(\boldsymbol{y}\mid\boldsymbol{\theta}) = f(\boldsymbol{y}\mid\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}\exp[-\frac{(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{2}] \quad (3)$$

where

$\boldsymbol{\mu}$: center vector

$\boldsymbol{\Sigma}$: covariance matrix

The log likelihood for $\boldsymbol{\theta}$ that is obtained from $\boldsymbol{y}$ may be written as

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N}\log\{\sum_{m=1}^{M}\pi_m f_m(\boldsymbol{y}\mid\boldsymbol{\theta}_m)\} \quad (4)$$

MLE of $\boldsymbol{\theta}$ is obtained by maximizing the equation. The EM algorithm was proposed to solve the formulation of (4) with the nonlinearity[11]. In the generalized EM algorithm, $\boldsymbol{y}$ is viewed as an incomplete data associated with latent variables and the EM algorithm maximizes (5) of the conditional expectation of the log likelihood instead of maximizing (4) directly.

$$Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)}) = E\{\log p(\boldsymbol{y},\boldsymbol{x}\mid\boldsymbol{\theta})\mid\boldsymbol{y},\boldsymbol{\theta}^{(t)}\} = \sum_{n=1}^{N}P(\boldsymbol{x}\mid\boldsymbol{y},\boldsymbol{\theta}^{(t)})\log p(\boldsymbol{y},\boldsymbol{x}\mid\boldsymbol{\theta}) \quad (5)$$

where

$\boldsymbol{y}$: observed variables

$\boldsymbol{x}$: latent variables

$Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)})$: conditional expectation of the log likelihood of complete data

$P(\boldsymbol{x}\mid\boldsymbol{y},\boldsymbol{\theta}^{(t)})$: *posterior* probability of $\boldsymbol{x}$ conditional on $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{y}$

$\boldsymbol{\theta}^{(t)}$: current estimate of step t

From Bayes' theorem, $P(\boldsymbol{x}\mid\boldsymbol{y},\boldsymbol{\theta}^{(t)})$ may be written as

$$P(\boldsymbol{x}\mid\boldsymbol{y},\boldsymbol{\theta}^{(t)}) = \frac{P(\boldsymbol{y},\boldsymbol{x}\mid\boldsymbol{\theta}^{(t)})}{\sum_{n=1}^{N}P(\boldsymbol{y},\boldsymbol{x}\mid\boldsymbol{\theta}^{(t)})} \quad (6)$$

The EM algorithm may be summarized as follows:

Step 1: Set the initial conditions of $\boldsymbol{\theta}^{(0)}$ and $t\leftarrow 0$

Step 2: (E step) compute $Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)})$

Step 3: (M step) $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)})$

Step 4: Stop if $\left|\frac{\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)}}{\boldsymbol{\theta}^{(t)}}\right| \le \varepsilon$. Otherwise, $t\leftarrow t+1$ and return to Step 2

The monotone behavior of the log likelihood function is guaranteed in a case where the Q function monotonously increases[11]. If the initialization is not close to a global optimum, the solution of the algorithm may be a local one. Suppose that the EM algorithm is applied to the MMND. The Q function may be written as

$$Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)}) = \sum_{n=1}^{N}\sum_{m=1}^{M}q_{nm}^{(t)}\log\{\pi_m f_m(\boldsymbol{y}_n\mid\boldsymbol{\theta}_m)\} \quad (7)$$

$$q_{nm}^{(t)} = \frac{\pi_m^{(t)} f_m(\boldsymbol{y}_n\mid\boldsymbol{\theta}_{m'}^{(t)})}{\sum_{m'=1}^{M}\pi_{m'}^{(t)} f_{m'}(\boldsymbol{y}_n\mid\boldsymbol{\theta}_{m'}^{(t)})} \quad (8)$$

where

q_{nm}: *posterior* probability that the observed variable $\boldsymbol{y}_n$ is assigned to the m-th component

Under the constraint condition of $\sum_{m=1}^{M}\pi_m = 1$, the Q function is maximized by maximizing (9) with the Lagrange multiplier method.

$$Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)}) + \lambda(\sum_{m}^{M}\pi_m - 1) \quad (9)$$

The mixing probabilities may be obtained as

$$\pi_m^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N}q_{nm}^{(t)} \quad (10)$$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ may be calculated as

$$\boldsymbol{\mu}_m^{(t+1)} = \frac{\sum_{n=1}^{N}q_{nm}^{(t)}\boldsymbol{y}_n}{\sum_{n=1}^{N}q_{nm}^{(t)}} \quad (11)$$

$$\boldsymbol{\Sigma}_m^{(t+1)} = \frac{\sum_{n=1}^{N}q_{nm}^{(t)}(\boldsymbol{y}_n-\boldsymbol{\mu}_m^{(t)})(\boldsymbol{y}_n-\boldsymbol{\mu}_m^{(t)})^T}{\sum_{n=1}^{N}q_{nm}^{(t)}} \quad (12)$$

3.2. DAEM algorithm

The EM algorithm is useful for obtaining a local optimum efficiently. However, it does not converge to global one due to the influence of initial conditions on the final one[10][12-13]. Several improved versions have been proposed to improve the convergence characteristics to a global optimum[12][14-15]. In particular, the performance is significantly improved by deterministic annealing[10]. To avoid local optima, this method smoothes the Q function through introducing the concept of temperature state in iteration. From the law of entropy increase, it is repeated until the algorithm converges to an equilibrium configuration for a fixed temperature state while the temperature state slowly decreases from high to low[6]. In the DAEM algorithm, (8) may be rewritten as

$$q_{nm}^{(t)} = \frac{\{\pi_m^{(t)} f_m(\boldsymbol{y}_n\mid\boldsymbol{\theta}_{m'}^{(t)})\}^{\beta}}{\sum_{m'=1}^{M}\{\pi_{m'}^{(t)} f_{m'}(\boldsymbol{y}_n\mid\boldsymbol{\theta}_{m'}^{(t)})\}^{\beta}} \quad (13)$$

β is temperature parameter that corresponds to reciprocal of temperature. It starts from $\beta_{\min}$ of initial

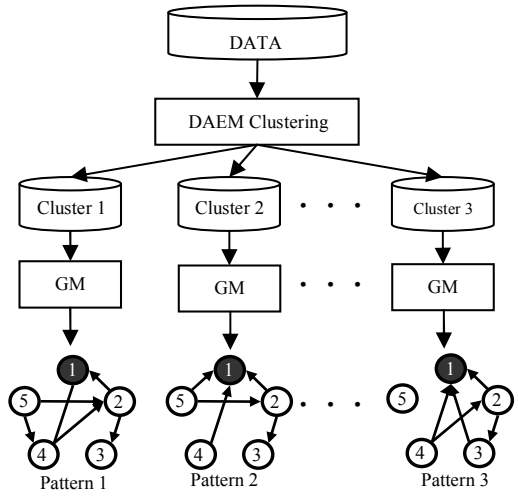

Fig. 3. Concept of Proposed Method

state and ends up to $\beta_{\max}$. If the algorithm converges at each temperate state, β is increased with

$$\beta_{\text{next}} \leftarrow \beta \times \beta_{\text{rate}} \tag{14}$$

4. Proposed method

This paper proposes a new method for clarifying the casual relationship in electric load forecasting. The proposed method is based on the hybrid method that consists of GM and DAEM clustering. GM is very useful for selecting input variables with the partial correlation. DAEM clustering serves as a preconditioned technique that classifies input data of the explained variable and the candidates of explanatory variables into clusters. The reason of using the DAEM clustering gives as follows:

a) Since GM deal with the casual relationship between Gaussian variables, a new technique is required to handle the casual relationship between non-Gaussian variables. The user of DAEM allows the uses to decompose non-Gaussian data into clusters of Gaussian data.
b) The DAEM algorithm has advantage that the obtained results are not affected by initial conditions. Namely, the introduction of Deterministic Annealing into the EM algorithm brings about a robust algorithm for decomposing non-Gaussian data into clusters of Gaussian data.

Fig. 3 shows the conceptual diagram of the proposed method with five variables, where Variable 1 is the explained variable and Variables 2-5 are the

Table 1. Explained Variable and Candidates of Explanatory Variable

	Contents	No.	Zone of NYISO or ISO
Explained Variable	Electric Hourly Load (*day*+1) of 14-hour	Z_1	Capital
Candidates of Explanatory Variable	Electric Hourly Price (*day*) of 14-hour	X_1	Capital
		X_2	Central
		X_3	Dunwoodie
		X_4	Genesee
		X_5	Hydro-Quebec
		X_6	Hudson Valley
		X_7	Long Island
		X_8	Mohawk Valley
		X_9	Millwood
		X_{10}	N.Y.C.
		X_{11}	North
		X_{12}	NEISO
		X_{13}	Ontario Hydro
		X_{14}	PJM
		X_{15}	West
	Electric Hourly Load (*day*) of 14-hour	X_{16}	Capital
		X_{17}	Central
		X_{18}	Dunwoodie
		X_{19}	Genesee
		X_{20}	Hudson Valley
		X_{21}	Long Island
		X_{22}	Mohawk Valley
		X_{23}	Millwood
		X_{24}	N.Y.C.
		X_{25}	North
		X_{26}	West

Note) the zone in gray is Independent System Operators (ISOs) that is adjacent to NYISO

candidates of explanatory variables. First, the distribution of the data is estimated by the DAEM algorithm, and the *posterior* probability that the observed variable $\boldsymbol{y}_n$ is assigned to the m-th component may be calculated by (15).

$$q^*_{nm}(\boldsymbol{y}_n \mid \pi^*_m, \boldsymbol{\mu}^*_m, \boldsymbol{\Sigma}^*_m) = \frac{\{\pi^*_m f_m(\boldsymbol{y}_n \mid \boldsymbol{\mu}^*_m, \boldsymbol{\Sigma}^*_m)\}^\beta}{\sum_{m'=1}^{M} \{\pi^*_{m'} f_{m'}(\boldsymbol{y}_n \mid \boldsymbol{\mu}^*_{m'}, \boldsymbol{\Sigma}^*_{m'})\}^\beta} \tag{15}$$

where

$\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$: MLEs of the parameter of the MMND

Let α to be a threshold value. A rule may be defined as soft clustering that if $\alpha < q^*_{nm} < 1$, data $\boldsymbol{y}_n$ is assigned to the m-th cluster[16]. In the process of GM, the influence of spurious correlation is removed by the partial correlation coefficients and the MLEs of the partial correlation coefficient are estimated by covariance selection. Finally, to evaluate the causal relationship of the explained variable and the candidates of explanatory variable, the cause-effect graphs are constructed with the results.

Table 2. Parameters of Methods

Common Parameters of EM and DAEM		DAEM	
ε	$\|\cdot\| \le 10^{-4}$	$\beta_{\min}$	0.5
M	3	$\beta_{\max}$	1.0368
α	0.2	β_{rate}	1.1

Where, ε is the convergence criterion for EM and DAEM

Table 3. Results of Clustering

		Mixing Probability	Data
Methods A and B	All Data	1	1277
Method C	Cluster 1	0.37	507
	Cluster 2	0.53	714
	Cluster 3	0.1	142
Method D	Cluster 1	0.36	501
	Cluster 2	0.53	709
	Cluster 3	0.11	147

Table 4. Results of Clustering

Methods A and B	Center of All Data	(0.49, 0.55, 0.54,..., 0.47, 0.25, 0.66)
Method C	Center 1	(0.49, 0.56, 0.55..., 0.48, 0.25, 0.67)
	Center 2	(0.47, 0.54, 0.54,…, 0.44, 0.25, 0.64)
	Center 3	(0.55, 0.56, 0.55,..., 0.44, 0.25, 0.64)
Method D	Center 1	(0.50, 0.56, 0.55,…, 0.48, 0.25, 0.67)
	Center 2	(0.47, 0.54, 0.54,…, 0.44, 0.25, 0.64)
	Center 3	(0.55, 0.57, 0.55,…, 0.57, 0.25, 0.72)

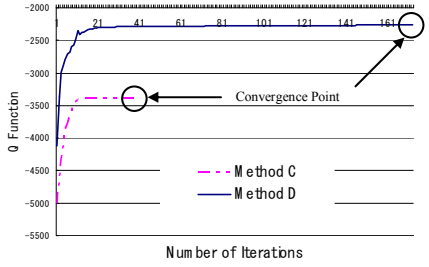

Fig. 4. Conditional Expectation of the Log Likelihood of Method C and D

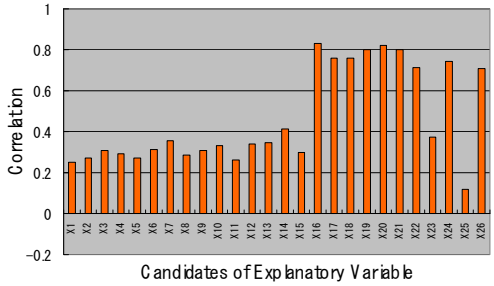

Fig. 5. Results of Method A

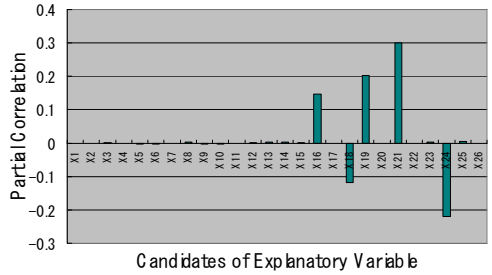

Fig. 6. Results of Method B

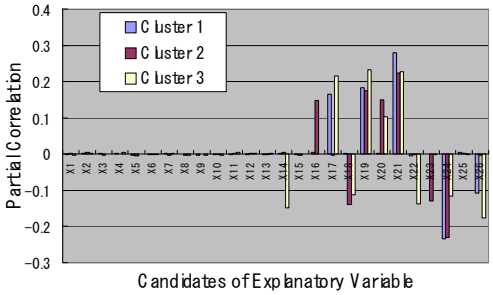

Fig. 7. Results of Method C

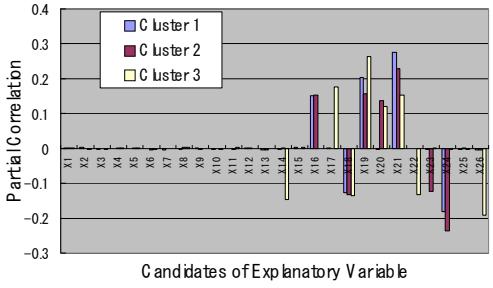

Fig. 8. Results of Method D

5. Simulation

5.1. Simulation conditions

The proposed method was applied to real data in NYISO[17]. This paper focuses on one-day-ahead electric load in Capital at 2 p.m. as the explained variable, where Capital means a zone in NYISO, and the load varies in a nonlinear and non-Gaussian way at 2 p.m. as shown in Fig. 1. To select appropriate explanatory variables, the paper prepares 26 candidates as shown in Table 1. A set of data (26×1276; 2005/2/1 to 2008/7/31) was used as test data. To demonstrate the effectiveness of the proposed method, this paper made a comparison between the proposed and other methods. For convenience, the following methods are defined as

Method A : Correlation Coefficient
Method B : Graphical Modeling (GM)
Method C : EM Clustering and GM (Proposed)
Method D : DAEM Clustering and GM (Proposed)

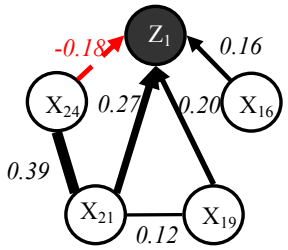

Fig. 9. Cause-effect Graph of Method D of Cluster 1

The parameters of each method were fixed as shown in Table 2.

5.2. Simulation results

Fig. 4 shows the conditional expectation of the log likelihood of EM and DAEM clustering that are used as a data preprocessing of Methods C and D respectively. It can be seen that Method C converges to a local optimum due to the high sensitivity to the initial values. On the other hand, Method D obtains a better solution to avoid local optima with smoothing the Q function. It means that the results of Method D are more accurate than those of Method C. That is because the clustering solution of Method D was more appropriate. Tables 3 and 4 show the results of clustering of Methods C and D. It can be observed that although the center vectors of each cluster estimated by Method D do not overlap mutually, a part of data vectors are shared in two or more clusters for clustering softly. Figs. 5 and 6 show the correlation and the partial correlation coefficients between the explained variable and the candidates of explanatory variables, respectively. It can be seen that the day-ahead load Z_1 of Capital has strongly positive correlation with a lot of candidates. However, the partial correlation coefficient dose not. After the spurious correlation was removed, Z_1 is strongly linked to X_{19} and X_{21} with positive causal relationship and X_{24} with negative one (see Fig. 6). The causal relationship for clustering results was investigated in detail with Methods C and D. It can be seen that the causal relationship of each cluster is different from others. It implies that the optimal explanatory variables may be different on different day. The proposed method makes the directed graph of causal relationship based on temporality. Fig. 9 gives the causal relationship of Z_1, X_{16}, X_{19}, X_{21} and X_{24} for Method D of Cluster 1.

6. Conclusion

This paper has proposed a new method for the variable selection of electric load forecasting. The proposed method makes use of a hybrid method that consists of DAEM clustering and GM. DAEM clustering plays a key role to divide the non-Gaussian data into Gaussian clusters so that the realistic causal relationship of each cluster was estimated by GM. The proposed method was successfully applied to real data of NYISO. The simulation results have shown that the proposed method provides more appropriate variable selection than the conventional methods.

7. References

[1] G. Gross and F. D. Galiana, "Short-term load forecasting", Proc. of the IEEE, Vol. 75, No. 12, Dec. 1987, pp. 1558-1573.
[2] R. C. Garcia, J. Contreras, M. van Akkeran and J. B. C. Garcia, "A GARCH forecasting model to predict day-ahead electricity prices", IEEE Trans. Power Syst., Vol.20, No.2, 2005, pp. 867-874.
[3] J. Bastian, J. Zhu,V. Banunarayanan and R. Mukerji, " Forecasting energy prices in a competitive market", IEEE Comput. Appl. Power, Vol. 12, No. 3, 2001, pp. 44-55.
[4] Julio Usaola, "Probabilistic Load Flow with Wind Production Uncertainty Using Cumulants and Cornish-Fisher Expansion", Electrical Power and Energy Systems, accepted 17 February 2009. Available online 21 March 2009.
[5] Joe Whittaker, Graphical Models in Applied Multivariate Statistics, John Wiley & Sons, 1990.
[6]Christian Borgelt and Rudolf Kruse, Graphical Models: Methods for Data Analysis and Mining, John Wiley and Sons, 2002.
[7] A. P. Dempster, "Covariance Selection", Biometrics, Vol. 28, No. 1, Special Multivariate Issue, Mar 1972, pp. 157-175.
[8] N. Wermuth and E. Scheidt, "Fitting a Covariance Selection Model to a Matrix", Blackwell Publishing for the Royal Statistical Society, Vol. 26, No. 1, 1977, pp. 88-92.
[9] T. P. Speed and H. T. Kiiveri, "Gaussian Markov Distributions over Finite Graphs," The Annals of Statistics, Vol.14, No.1, Mar.1986, pp. 138-150.
[10] N. Ueda and R. nakano, "Deterministic annealing EM algorithm", Neural Networks, Vol. 11, 1998, pp. 271-282.
[11] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", J. Roy. Statist. Soc. Ser. B, Vol. 39, No. 1, 1977, pp. 1-38.
[12] Geoffrey McLachlan and David Peel, Finite Mixture Models, John Wiley & Sons, 2000.
[13] T. K. Moon, "The expectation-maximization algorithm", IEEE Signal Process. Mag., Vol. 13, No. 6, Nov. 1996, pp. 47-60.
[14] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 24, No. 3, Mar. 2002, pp. 381–396.
[15] J. J. Verbeek, N. Vlassis, and B. J. A. Krose, "Efficient greedy learning of Gaussian mixture models", Neural Computation, Vol. 8, No. 2, 2003, pp. 469–485.
[16] S. Z. Selim and M. A. Ismail, "Soft clustering of multidimensional data: A semi-fuzzy approach", Pattern Recogn., Vol. 17, No. 5, 1984, pp. 559–568.
[17]www.nyiso.com