EasyChair Preprint

№ 135

A Comparison of applying Multiple Methodologies for short-term Load Forecasting

Mingsui Sun, Mahsa Ghorbani, Edwin Chong and Siddharth Suryanarayanan

May 16, 2018

A Comparison of applying Multiple Methodologies for short-term Load Forecasting

1st Mingsui Sun
Electrical and Computer Engineering
Colorado State University
Fort Collins, Colorado, United States
mssun@colostate.edu

2nd Mahsa Ghorbani
Systems Engineering
Colorado State University
Fort Collins, Colorado, United States
mahsa.ghorbani@colostate.edu

3th Edwin K.P. Chong
Electrical and Computer Engineering
Colorado State University
Fort Collins, Colorado, United States
edwin.chong@colostate.edu

4th Siddharth Suryanarayanan
Electrical and Computer Engineering
Colorado State University
Fort Collins, Colorado, United States
sid.suryanarayanan@colostate.edu

***Abstract*—We present five methodologies for probabilistic load forecasting which are a method based on Bayesian estimation, a rank-reduction operation based on principle component analysis, least absolute shrinkage and selection operator (Lasso) estimation, ridge regression, and a supervised learning algorithm called scaled conjugate gradient (SCG) neural network. These five models considered can be regarded as a variety of competitive approaches for analyzing hourly electric load and temperature. The modeling approaches incorporates the load and temperature effects directly, and reflect hourly patterns of the load. We provide empirical studies based on the Global Energy Forecasting Competition 2014 (GEFCom 2014). In this research, we use historical load data only to forecast the future load. The study performs the estimation comparison of the five methodologies, showing that ridge regression has a marginal advantage over the others.**

***Index Terms*—Load Forecasting, methodologies, mean squared error, relative error percentage**

I. Introduction

Point forecasts have been used for several decades to predict energy supply, demand and prices for electrical system and financial planning purposes. An overview of energy forecasting was introduced in Hong (2014), which looks at the forecasting practices of smart grids back upon the inception of the electricity market. Many methodologies have been introduced and extensively applied in the utility market, however, due to the global modernization of smart power grids, the electricity demand becomes more and more volatile and less predictable. To find the consumption pattern of the electricity load and to predict more precisely to meet the economic demand, Global Energy Forecasting Competition (GEFCom) was held by Dr. Tao Hong in 2012, 2014 and 2017 respectively to invite worldwide candidates submit their load estimation results for forecasting energy demand. Hong et al. (2016) summarized the up-to-data research progress on probabilistic energy forecasting, most of them were the submission of GEFCom 2014.

A practical overview of energy forecasting is discussed in a chronological order including short and long term load forecasting for over a century and a summarized computer based method for short term load forecasting is presented in Hong (2014). In this paper, we introduced and discussed several methodologies for point load forecasting using the case study from the forecasting competition GEFCom2014. These methodologies include Gauss-Bayes (GB), Reduced Rank Gauss-Bayes, Lasso, Ridge Regression(RR) and machine learning (scaled conjugate gradient algorithm). Due to the intrinsic conditions/problems that cause the estimation of the future load consumption to be unreliable, we tried to find one method that can minimize the mean square error and achieve the lowest relative error percentage. Thus, the purpose of using these five methods is to compare the load forecasting results that are used to quantify the uncertainty in the electricity demand. The data we used in this paper is GEFCom2014 provided by Hong et al. (2016).

The structure of this paper is as follows: section 2 introduces five methodologies; section 3 presents the load forecasting results in each of the techniques and give a further discussion on ridge regression and principle components analysis; finally, the paper is concluded in section 4.

II. Prediction Methodologies

A. Gauss-Bayes

Assume the multivariate random vector $x_i \in \mathbb{R}_{\mathbb{N}}(i = 1, ..., k)$ includes the hourly load values of k consecutive

consumption months. Two sub-matrices Y and Z can be partitioned from the raw matrix X, such as

$$[X] = [Y \quad Z] \tag{1}$$

where submatrices Y and Z include historical and future load values.

Suppose that random variables y and z are jointly normally distributed. The Bayesian posterior distribution of $(z|y = Y)$ is given by

$$\hat{z}_{z|y} = \Sigma_{zy}\Sigma_{yy}^{-1}y \tag{2}$$

and

$$\hat{\Sigma}_{z|y} = \Sigma_{zz} - \Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz} \tag{3}$$

where Σ_{yy}^{-1} is the generalized inverse of Σ_{yy} and $\Sigma_{zy}\Sigma_{yy}^{-1}$ is the regression coefficients of the matrix.

The Gauss-Bayes technique can result in the optimal values of MSE, however, there is an unavoidable fact that the matrix of Σ_{yy} is not always well conditioned and the numerical calculations of this method cannot be trusted (Ghorbani and Chong 2017).

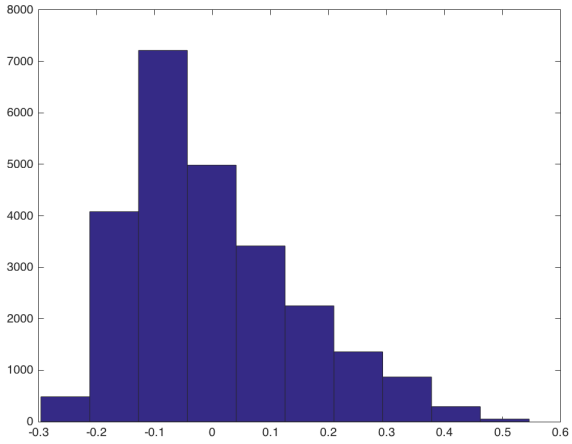


Fig. 1: plot of normalized historical data

B. Reduced Rank Gauss-Bayes

Reduced Rand Gauss-Bayes is defined as a dimension reduction method. Principle component analysis also known as an unsupervised dimension reduction methodology is one of the techniques for multivariate analysis (Maitra and Yan 2008), and the reduced rank normally presents on reducing the number of predictive variable (Bair,Hastie, Paul and Tibshirani 2006). Due to the huge calculating load of computing X^TX to get eigenvector W, there is a computationally efficient way to obtain the W by using singular value decomposition (SVD), which is given by

$$X = U\Sigma\mathbf{V}^{\intercal} \tag{4}$$

However, the smallest singular value from Eq.(4) is big enough to supply determinate information about the condition number of the raw matrix X (Klema and Laub 1980).

Since the first few eigenvalues normally account for the bulk part of the sum of all the eigenvalues, using only a subset of eigenvalues and the corresponding eigenvectors is a reasonable approach. Resolving the noisy observation vector onto a principle subspace which only includes the filtered information can be achieved by

$$\mathbf{w} = Gy \tag{5}$$

where $G = (V_{M,L}^{'}V_{M,L})^{-1}V_{M,L}^{'}$, M is the lenght of random vector y and L is the number of eigenvalues included. (Ghorbani and Chong 2017). By substituting w in Eq.(2) and (3) we have:

$$\hat{z}_{z|w} = \Sigma_{zw}\Sigma_{ww}^{-1}w \tag{6}$$

and

$$\hat{\Sigma}_{z|w} = \Sigma_{zz} - \Sigma_{zw}\Sigma_{ww}^{-1}\Sigma_{wz} \tag{7}$$

Due to the dimensional reduction property of this method, if the posterior distribution of estimation of z using Eq.(6) and (7) has similar results to the estimation results from Eq.(2) and (3), this method can be considered as a good substitute for Gauss-Bayes method.

C. Ridge Regression (RR)

Ridge regression technique is one of the shrinkage methods that incorporate the shrinkage estimator, which is added into the diagonal elements of correlation matrix. Compared to other ordinary least square (OLS) estimation techniques, ridge regression method is more stable as it is a continuous process that shrinks parameters (Tibshirani, 1996). Because of this diagonal of ones considered as a ridge, it is where the ridge regression gets its name. Thus, the RR technique is given by

$$\beta = (R + \lambda I)^{-1}X^{'}Y \tag{8}$$

The value of bias of the estimator is given by

$$\beta = (R + \lambda I)^{-1}X^{'}Y \tag{9}$$

and the equation of the covariance matrix is given by

$$V(\beta^{'}) = (X^{'}X + \lambda I)^{-1}X^{'}X(X^{'}X + \lambda I)^{-1} \tag{10}$$

As a shrinkage method, ridge regression is slightly different as the estimation method which will be mentioned below. RR can only shrink all the coefficient toward zero, but not exactly to zero unless $\lambda = \infty$. However, due to this reason, ridge regression cannot perform variable selection in the linear model, which is one of the advantages of technique. The other difference between the and RR is the ridge penalty term uses the $\|\beta\|_2^2$ where the uses the $\|\beta\|_1$, it represents as

$$argmin\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{11}$$

in which

$$argmin\|y - X\beta\|_2^2, s.t.\|\beta\|_2^2 \leq t \tag{12}$$

where t is the tuning parameter of λ.

D. least Absolute Selection and Shrinkage Operator ()

Least absolute selection and shrinkage operator (Lasso) estimation technique is a penalized ordinary least squares (OLS) regression estimator (Ziel and Liu 2016). One of the reason that Lasso is superior to OLS is that it can determine a smaller subset to give the powerful effects when there are a large number of predictors (Tibshirani,1996). Thus, the Lasso estimation is given by

$$argmin\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{13}$$

where λ is the tuning parameter and X is the regressor matrix. Lasso estimation is another shrinkage method that tunes λ to determine the 'penalty' on the sizes of parameter. The purpose of introducing λ in the OLS is to get a good estimation of y but balancing the coefficient of $|y|$ in a reasonable size. In this research, we use 10-fold cross-validation of estimation that will automatically select a sequence number of different lambdas for the forecasting. During the calculation, Lasso will try to set some coefficient to zero, it represents as

$$argmin\|y - X\beta\|_2^2, s.t.\|\beta\|_1 \leq t \tag{14}$$

t is the 1-to-1 relationship with the λ mentioned in Eq.(14).

E. Scale Conjugated Gradient Algorithm (SCG)

The technique of deep learning used in this research is scaled conjugated gradient algorithm, which is a method of combining the model-trust region approach with the scale conjugate gradient (SCG) approach. Thereinto, the model-trust region approach comes from the Levenberg-Marquardt algorithm which is a variation of the standard Newton algorithm. A Lagrange Multiplier (Fletcher, 1975) λ_k is introduced in SCG algorithm to regulate the indefiniteness of $E^{"}(\tilde{w}_k)$, which is

$$\bar{s}_k = \frac{E^{'}(\tilde{w}_k + \alpha_k\tilde{P}_k) - E^{'}(\tilde{w}_k)}{\sigma_k} + \lambda_k\tilde{P}_k \tag{15}$$

Here, $\tilde{P}_k$ denote as search direction, α_k is the step size, $E(\tilde{w})$ is an error function, in which $\tilde{w}_k + \alpha_k\tilde{P}_k$ is an updated function. If $E^{'}(\tilde{w}_k \neq 0)$ then set $k = k + 1$ and go to 2 else return $\tilde{w}_{k+1}$, as the desired minimum (Moller,1993).

III. Pre-processing Data

In this research, we use five year continually hourly load and corresponding 25 stations temperatures data. To format this data set, we use a Hankel matrix which contains k rows sample of vector data with length N in each row. Hankel matrix stocks K samples each one time shifted from the previous one. Assume L_N represents the load consumption for day N. Then our Hankel matrix is:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} L_1 & L_2 & \cdots & L_N \\ L_2 & L_3 & \cdots & L_{N+1} \\ \vdots & \vdots & \ddots & \vdots \\ L_k & L_{k+1} & \cdots & L_{K+N-1} \end{bmatrix} \tag{16}$$

As the data vectors $x_1, x_2, ..., x_k$ are not drawn from the same distribution with the same units, we should perform a scaling approach to the load and temperatures in different benchmarks to make the normalization meaningful. One approach is introduced here. Suppose that $l_i(M)$ is the maximum value in the vector l_i which is a vector of load and/or temperature over N consecutive measuring 24 hours of days. We can apply the normalization to obtain x_i as

$$x_i = l_i/l_i(M) \tag{17}$$

This normalization in Eq.(26) has the interpretation that vector contains load and/or temperature values as a fraction of the values on the Mth day. Before applying each technique, we also subtract the average vector $\bar{x}$ from each x_i. (Ghorbani and Chong 2017).

IV. Results and Discussion

Figure 1 represent the histogram of the normalized data (first column in matrix X) and as you can see it represents a bell shape.

The discussion based on each of the constructed matrices of the historical data X will be applied in five techniques respectively. The estimation is focused on the Nth hour load value by using the $N-1$ previous hours load consumption value as predictors. Here N is a variable value. Relative error percentage (REP) is used here to compare the performance of each technique. The REP is calculated by comparing the mean squared error with the true future load value, it represents as

$$\begin{aligned} mse &= \frac{1}{N}\sum_{i=1}^{N}(z - \hat{z})^2 \\ base &= \frac{1}{N}\sum_{i=1}^{N} z^2 \\ REP &= 100 * \sqrt{mse/base} \end{aligned} \tag{18}$$

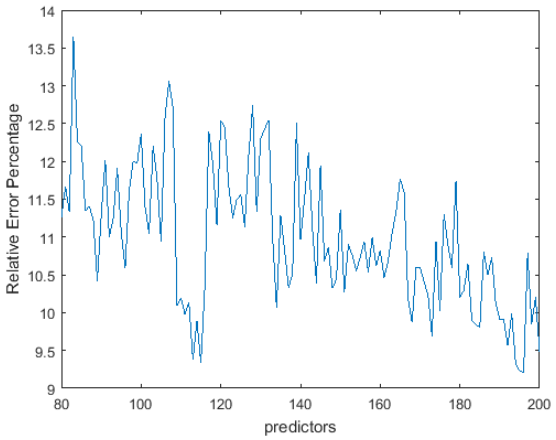
A. Comparison of prediction results

First, we use eight predictor variables to forecast the next-hour load (prediction) value. The eight predictors are as follows: the previous five hours of load values, the load value from 24 hours ago, and the load value from exactly seven days ago. In Table 1, we show the results of the load forecasting for all five techniques.

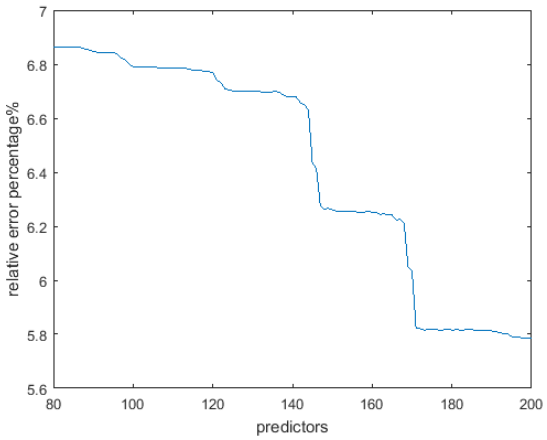
Load data Forecasting					
Techniques	Gauss-Bayes	Reduced Rank Gauss-Bayes	Lasso	Ridge Regression	SCG ANN
Relative Error Percentage	11.5539%	12.4589%	11.5533%	11.6473%	9.3047%

TABLE I: Load forecasting by using only 8 predictors

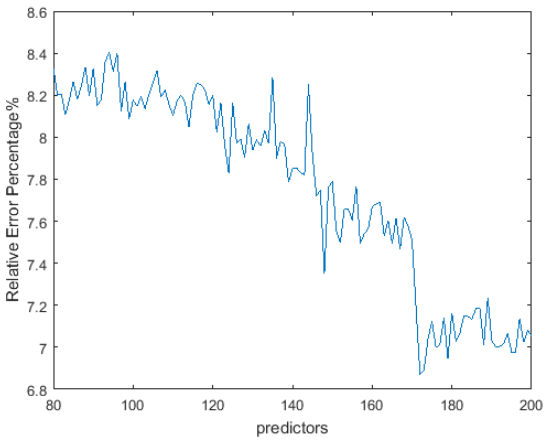
Table 1 shows the percentage of the relative error when only considering eight historical load data as predictors to predict the 6th hourly load consumption. It turns out that SCG gives the best result when the numbers of predictors



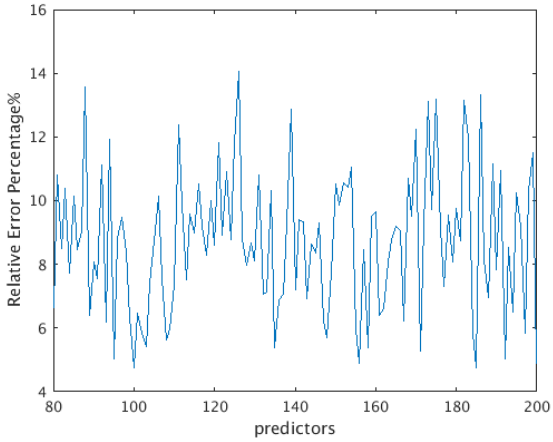
(a) Reduced Rank Gauss-Bayes result



(b) RR result



(c) Lasso Result



(d) SCG ANN result

Fig. 2: Plots of Load Forecasting in Multiple Techniques

are limited. However, except SCG technique shows a smaller percentage that is below 10%, the prediction results of all the other techniques are relatively worse than the expectation. Thus, a further testing was performed to see whether it can provide a more accurate forecasting.

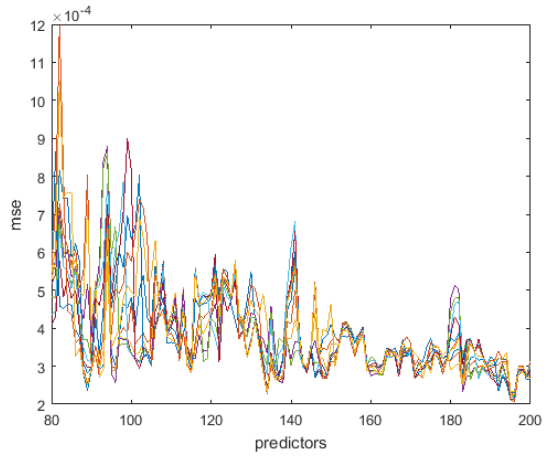
At this point, we change the observation vector $N-1$ from 80 to 200 and the comparison among four techniques are shown in Figure 2. It's worth noting that Gauss-Bayes method cannot be performed when the observation vector is too big. The reason of calculation difficulties is due to ill-conditioning issues associated with Gauss-Bayes. Thus, only Reduced Rank Gauss-Bayes, Lasso, ridge regression and SCG were participated in this competition. The results are shown in Figure 2.

It turns out that the amplitude oscillation of the curve in Reduced Rank Gauss-Bayes and SCG techniques are quite big, and the relative error percentage ranges roughly from 13.5% to 9.2%, and 14% to 5% respectively. In addition, the curve of Reduced Rank Gauss-Bayes method slowly converge when the number of predictors are more than 140 while the curve has no such obvious convergent tendency in SCG ANN method. In ridge regression technique, the value of REP has a dramatic decrease when M is around 165 and then converge to about 6% when M is 180. Although the curve in Lasso technique has the same tendency with ridge regression, the REP is slightly higher than the result in ridge regression and also more oscillated. From now on, ridge regression shows the best prediction results when M changes from 80 to 200. However, the tuning parameter λ is a fixed value which may affect the prediction result although it turns out to have the best forecasting ability. Thus we change the λ value to see how much the load forecasting will be affected.

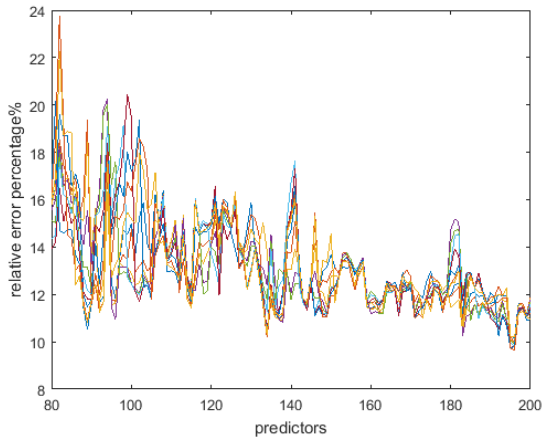
B. Ridge Regression Technique Analysis

We set λ value as a variable value then applied them individually in the ridge regression method simulation. Part of the result of the simulation is shown in Table 2.

The results in Table 2 illustrated an apparently pattern. The load forecasting REP value increased when the Lambda



(a) Reduced Rank Gauss-Bayes mse vs M result



(b) Reduced Rank Gauss-Bayes REP vs predictors result

Fig. 3: Reduced Rank Gauss-Bayes Prediction Results for different L

M \ λ	0.001	0.1	0.3	0.5	0.7	0.9
80	6.865	6.960	7.248	7.506	7.730	7.928
100	6.790	6.876	7.149	7.399	7.619	7.814
120	6.7683	6.849	7.121	7.371	7.590	7.785
150	6.261	6.338	6.601	6.845	7.061	7.253
180	5.816	5.883	6.128	6.359	6.564	6.748
200	5.787	5.855	6.097	6.325	6.527	6.709

TABLE II: M versus Lambda of RR Load Forecasting REP Result

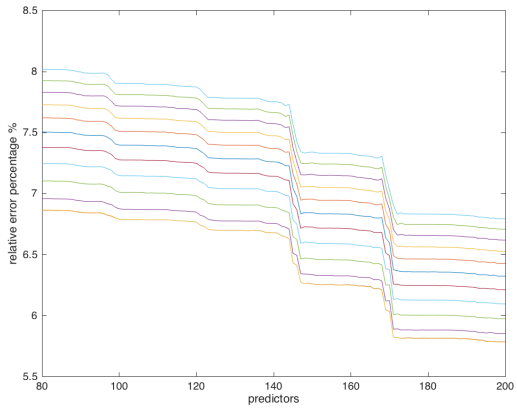


Fig. 4: M versus Lambda of RR Load Forecasting REP Result

value increased, in other words, a higher lambda gave a lower prediction accuracy. However, as the M (predictors) increases, the REP values decreases, which gave a better forecasting. It is easy to understand that the more historical data were used in the technique, the better forecasting result might be. On the other hand, Although the λ had less influence to the load forecasting and the ridge tuning was not that important in this case, ridge regression technique still gave the best forecasting result here. The lowest REP value can achieve as low as 5.787%. The overall simulation result of ridge regression technique with different M and variable λ was shown in Figure 4.

C. Reduced Rank Gauss-Bayes method Analysis

Although we set up data matrices with different lengths of observation vector, in Reduced Rank Gauss-Bayes method, we are more interested in improving values of REP in each of the observation vectors with different lengths.

Thus, instead of just choosing one hypothetical optimal L value by setting only one threshold (95% in the above case), we make the threshold as a variable number ranging from 50% to 95% with 5% increment in each simulation, and the results are shown in Figure 3.

To better understanding the performance of Reduced Rank Gauss-Bayes method, studying of the dimension of the objected subspace is a good way. Here L is the number of eigenvalues that represents part components of the data information. As the setting threshold was a variable value, the obtained L was also changed correspondingly. In Figure 4, the value of MSE and REP for different Ms with variable threshold are shown in Figure .3(a) and Figure .3(b) respectively. As the L value increases to get more information, the value of MSE and REP decreases when M increases.However, after a certain point even by increasing M the value of MSE and REP do not improve much due to the increased noise involved.

Thus, Reduced Rank Gauss-Bayes method by using L represented as the dimension of the principal subspace shows the decreased dimension calculation to find a more accurate prediction result.

D. Discussion

Even though ridge regression methodology outperforms other methods with regard to relative error percentage in load forecasting, we may be able to improve other techniques forecasting result in several ways. In this research, we treat

load data as historical information without considering the seasonal facts. Ziel et al. (2016) mentioned a method to separate the load data in eight groups according to the seasonal daily, weekly and annual patterns. By doing this, a significant improvement in the forecasting result might be achieved. We simply consider historical load values in forecasting future load values, while for the GEFCom2014-L data, the available temperature information might give a different technique winner in load forecasting. The proposed weather station selection method mentioned in Hong et al. (2015) may give a better supporting in the utilization of temperature data. We only take Reduced Rank Gauss-Bayes method and ridge regression technique in further discussion, while Lasso and machine learning technique can also be taken into consideration by applying an iteratively reweighted approach and combined back propagation algorithm respectively in the forecasting progress. Lastly, instead of only predicting the next hour's load value, we can forecast the next several hours' load value in order to give a hint for the peak load pricing in electricity market.

V. Conclusion

We introduced five techniques based on Gauss-Bayes, principle component analysis, Lasso, ridge regression, and scaled conjugate gradient ANN, which were used individually based on historical load and temperature data in load forecasting. Ridge regression ranked first (marginally) among the five methods, and gives a relative error percentage roughly between 5% to 7.5%. Reduced Rank Gauss-Bayes turns out to be second best.

References

[1] Hong, Tao, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond." (2016): 896-913.

[2] Ghorbani, Mahsa and Edwin K.P.Chong.'A new approach for forecasting stock prices using principal components." Paper presented at the 9th Annual American Business Research Conference, New York, NY,July 2017.

[3] Hong, Tao, and Shu Fan. "Probabilistic electric load forecasting: A tutorial review." International Journal of Forecasting 32, no. 3 (2016): 914-938.

[4] Hong, Tao. "Energy forecasting: Past, present, and future." Foresight: The International Journal of Applied Forecasting 32 (2014): 43-48.

[5] Klema, Virginia, and Alan Laub. "The singular value decomposition: Its computation and some applications." IEEE Transactions on automatic control 25, no. 2 (1980): 164-176.

[6] Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, no. 3 (2011): 273-282.

[7] Maitra, Saikat, and Jun Yan. "Principle component analysis and partial least squares: Two dimension reduction techniques for regression." Applying Multivariate Statistical Models 79 (2008): 79-90.

[8] Kleibergen, Frank, and Richard Paap. "Generalized reduced rank tests using the singular value decomposition." Journal of econometrics 133, no. 1 (2006): 97-126.

[9] Ziel, Florian, and Bidong Liu. "Lasso estimation for GEFCom2014 probabilistic electric load forecasting." International Journal of Forecasting 32, no. 3 (2016): 1029-1037.

[10] Bair, Eric, Trevor Hastie, Debashis Paul, and Robert Tibshirani. "Prediction by supervised principal components." Journal of the American Statistical Association 101, no. 473 (2006): 119-137.

[11] Møller, Martin Fodslette. "A scaled conjugate gradient algorithm for fast supervised learning." Neural networks 6, no. 4 (1993): 525-533.

[12] Bakirtzis, A. G., V. Petridis, S. J. Kiartzis, M. C. Alexiadis, and A. H. Maissis. "A neural network short term load forecasting model for the Greek power system." IEEE Transactions on power systems 11, no. 2 (1996): 858-863.

[13] Park, Dong C., M. A. El-Sharkawi, R. J. Marks, L. E. Atlas, and M. J. Damborg. "Electric load forecasting using an artificial neural network." IEEE transactions on Power Systems 6, no. 2 (1991): 442-449.