Non-Parametric Method Final Project: Predicting Power Output

Introduction of Data

Power companies are always having to produce electricity for a customer who's demands can change almost instantly. Power production facilities however can't as simply moderate their output in an economical fashion. Reasons vary from internal inputs such as amounts of resources used to external variables like the weather can all have an impact on a facility's power output. Thus, a sought after goal by many producers is a better understanding of the variables controlling power output. Data was collected by Heysem (2012) on variables believed affect output for such efforts.

Specifically, the goal for this data is to produce an accurate prediction of power output using simpler techniques than physical based thermodynamical approaches that involve thousands of nonlinear equations. This way, the power company can generate timely predictions of peak power in a computationally feasible manner. These are then used for deciding how to best, economically speaking, distribute power generation throughout the companies various plants when power demand peaks.

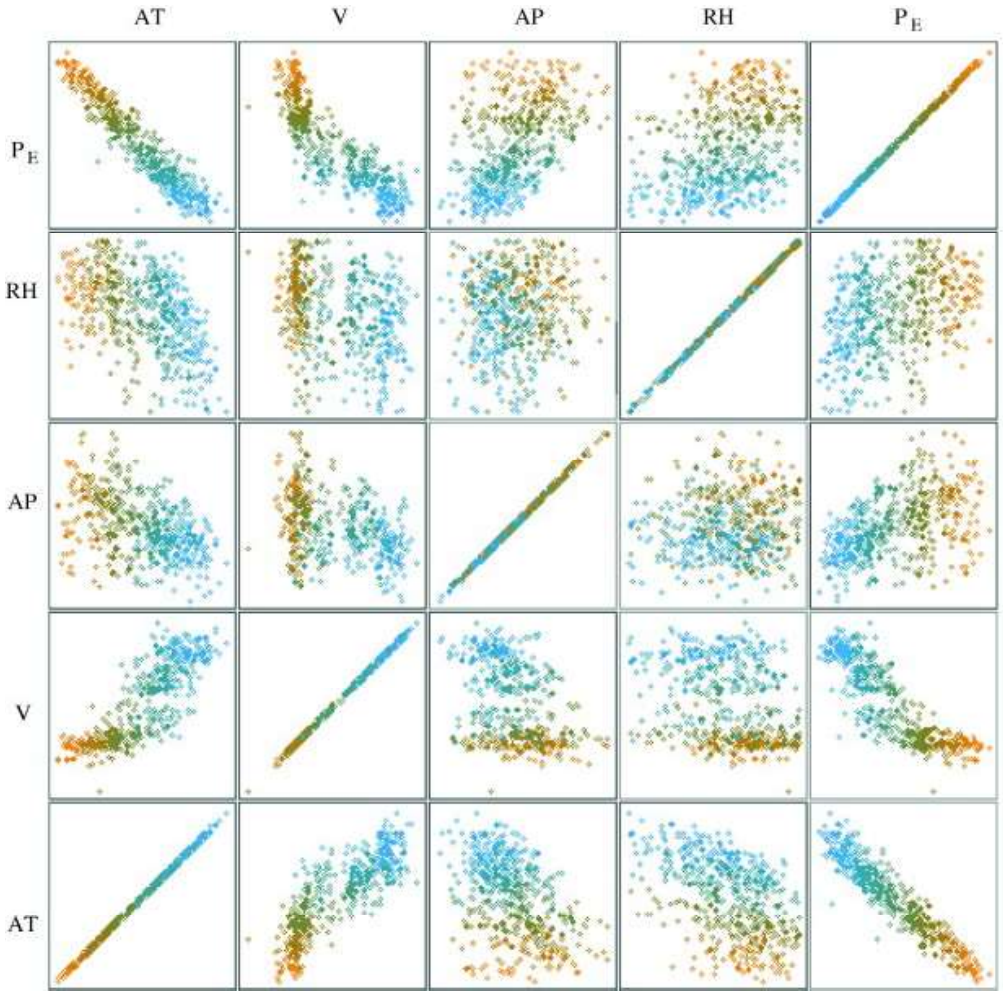
The data set consisting of 9,568 observations of five variables was collected over a six-year period (2006-2011) at a combined cycle power plant (CCPP) in Turkey when operating at full load. Each observation itself is an hourly average of the following five variables:

- **Ambient Temperature (AT)**: The temperature outside of the plant measured in Celsius.
- **Atmospheric Pressure (AP):** The pressure exerted by the weight of air in our atmosphere measured in minibars.
- **Relative Humidity (RH):** The ratio of partial pressure of water vapor to the equilibrium vapor pressure of water (i.e. a percentage). Essentially, the amount of water in the air at a given time.
- **Vacuum Exhaust Steam Pressure (VE**): A measure of pressure inside the condenser that directly impacts steam cycle of combined cycle power plants with units cm Hg. (http://www.nwfpa.org/nwfpa.info/component/content/article/ 365-improve-steam-turbineefficiency?start=3)
- **Full Load Electrical Power Output** (PE): Electricity generation measured in megawatts.

Originally, the data began as a collection of 674 excel data sheets, one for each day that the CCPP operated at full capacity. The data however included occasional observations that were

incompatible with the CCPP's operational capacity. The reason, according to the dataset's author, is that said observations are most likely a result of electrical disturbances causing malfunctions in the measurement equipment. Hence the aberrant observations were removed and the 674 excel data sheets were then combined into the single excel file found that can be found at the University of California Irvine's Machine Learning Repository website.

Since the purpose of this data is to predict the energy output I have included a pairwise scatterplot of the variables that are color coded according to the temperature associate with each observation (yellow for hot, blue for cool, and green for in between).



Parametric Analysis

To analyze the data using a parametric method I will use a standard linear regression model. This method will view the predicted energy output, Y_i, as a linear combination of a subset of p data features, $X_1, \ldots, X_p$. That is

$$Y_i = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

The values of the fitted coefficients, $\hat{\beta}_i$ will be the OLS estimates since the cost function for predicting is the root mean squared error (RMSE). Where the RMSE function of a particular fit, $\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_i$, is

$$\text{RMSE}(\hat{f}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(\mathbf{x}_i))^2}, \text{ where } n \text{ is the number of observations}$$

The assumptions of this model, given the RMSE cost function, are the decomposition of features (as fixed variables) vs. noise (error), independence of the observations, and the distribution of said noise. Note that no assumptions are required to determine the OLS estimates under the RMSE cost function. This is because the $\hat{\beta}$'s can be solved for analytically via minimization of the MSE function. The linear regression model does assume, though, that the decomposition of features are additive with each other, and that the coefficients of each feature must be linear. The predictor and response however can be manipulated via transformation if necessary. Also, the noise (residuals) are assumed to follow an i.i.d. normal distribution, which allows for the development of confidence intervals of the fitted coefficients ($\hat{\beta}$, etc.).

The features that I will use for my model are all four of the predictor variables (Temperature=AT, Vacuum Exhaust Pressure=V, Humidity=RH, Atmospheric Pressure=AP) with a polynomial transformation on Temperature. The resulting model is

$$\hat{Y} = 378.9 - 1349 * \text{AT} + 1495 * \text{AT}^2 - 0.29 * \text{V} + 0.098 * \text{AP} - 0.19 * \text{RH}$$

Note, that colinearity was not an issue because of the large number of observations. Despite the high coorelations between variables. The correlation structure is

	Temp.	Vacuum	Pressure	Humidity	Power
Temp.	1				
Vacuum	0.84	1			
Pressure	-0.51	-0.41	1		
Humidity	0.54	-0.31	0.1	1	
Power	-0.95	-0.87	0.52	0.39	1

Also, the confidence intervals for the model coefficients are given by the following formula (given the model assumptions hold). Note $t_{\alpha/2}$ is the $\alpha/2$ quantile of a t distribution with $n-p = 9,568-3$ degrees of freedom and $\hat{\sigma}_i$ is the standard error of $\hat{\beta}_i$, such that

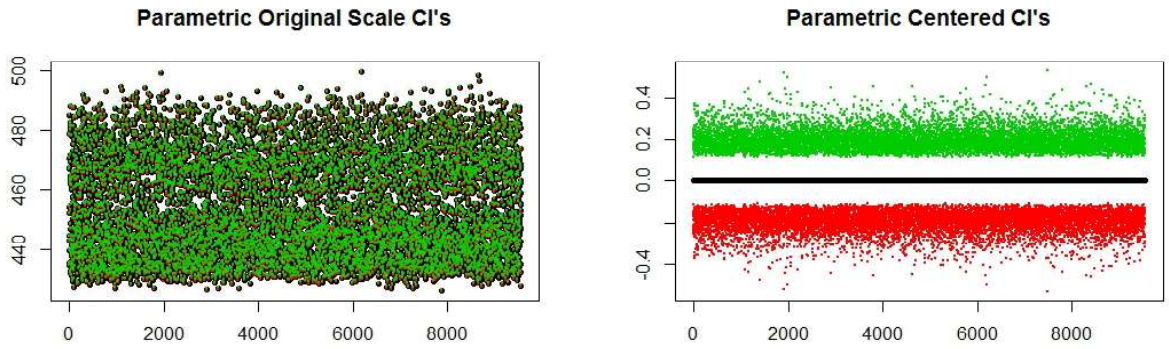
$$\beta_i \in [\hat{\beta} - t_{\alpha/2}\hat{\sigma}_i,\ \hat{\beta} + t_{\alpha/2}\hat{\sigma}_i]$$

The model coefficients have the following confidence intervals (note, all are statistically significant at the 1% level).

	Lower Bound	Upper Bound
Intercept	361.511	396.281
AT	-1369.719	-1328.201
AT^2	140.832	158.113
V	-0.299	-0.272
AP	0.081	0.115
RH	-0.126	-0.111

For the confidence bands the distance between the bounds is very small compared to the scale of the response variables range itself. So for viewing purposes I also graphed the confidence intervals about each point after centering the predicted variable about the mean. The formula used for prediction CI's about the mean for observation $\mathbf{x}_i$ is

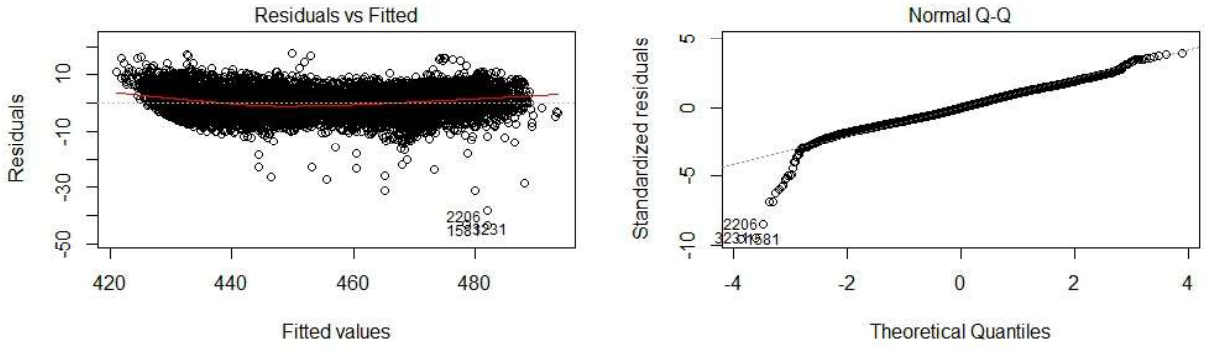
$$[\hat{Y} - t_{n-p,\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_i(X^TX)^{-1}\mathbf{x}_i}, \hat{Y} + t_{n-p,\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_i(X^TX)^{-1}\mathbf{x}_i}]$$



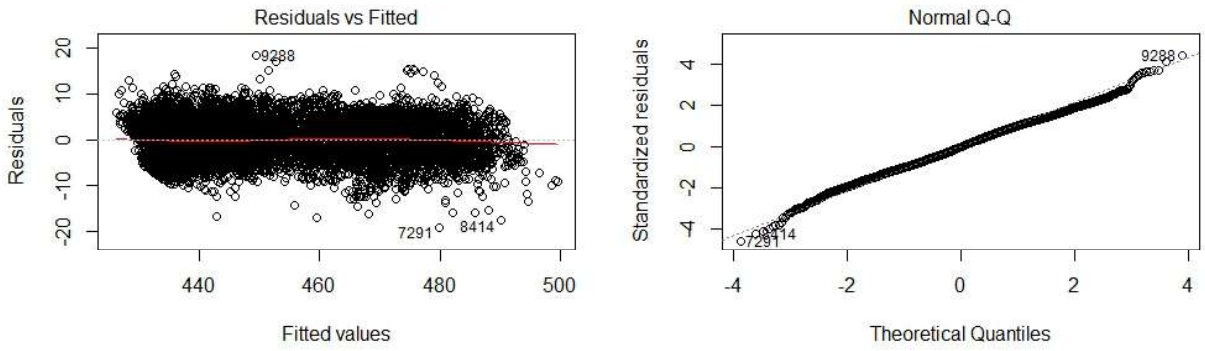
In order to determine the final model described above, I used a two phase process. First I determined the best set of features out of all the possible feature subsets (16-1) using the 2-fold cross validation error rate average as my criterion. The RMSE scores are below

Features	RMSE	Features	RMSE	Features	RMSE
AT	5.43	AT + AP	5.38	AT + V + AP	4.89
AP	8.42	AT + RH	4.8	AT + V + RH	4.57
V	14.6	V + AP	7.88	AT + AP + RH	4.8
RH	15.73	V + RH	8.15	V + AP + RH	7.56
T+V	4.96	AP + RH	13.39	AT + AP + V + RH	4.56

Second, I looked for problems in the residuals and removed/transformed the data as necessary. The model with only the four features and none of the data removed has the following residuals vs. fitted values and QQ-plot below. In them we see that the left tail is a bit thinner than it should be and their appears to be a slight curvature to the residuals indicating a slight degree of non-constant variance.



After removing the problem points to fix the QQ-plot ths also removed some of the outliers in the residuals vs. fitted plots. Also, adding the transformed temperature variable to correct for the non-constant variance results in the following:



Some of the motivation for choosing a linear model was it's direct aim at minimizing my prediction cost function, RMSE. Also, using k-fold crossvalidation as a judge of feature selection should help reduce some of the potential overfitting of the data.

Nonparametric Analysis

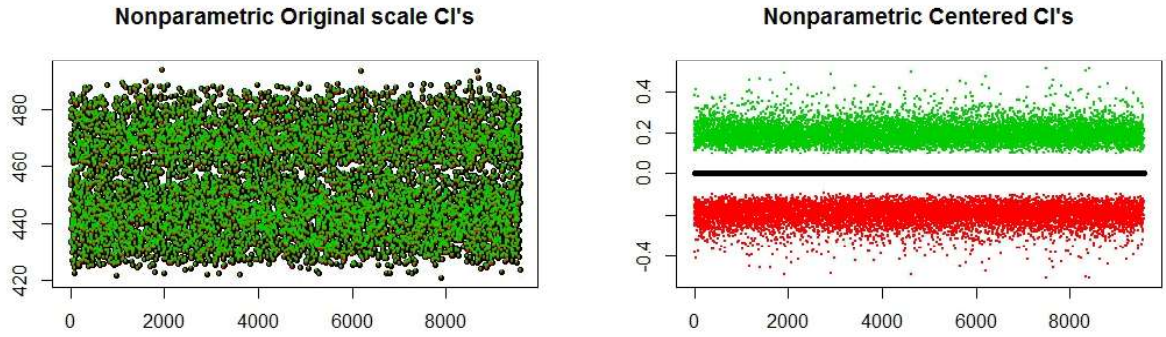
For nonparametric analysis I will apply the generalized additive modeling technique to analyze the CCPP data. This model works by assuming the mean of the response variable, $\mathbb{E}[Y_i|X_i] = f(\mathbf{X}_i)$, is a result of the linear combination of the a constant, μ,and p predictor functions $m_i(x_i)$. Here, each predictor function is a simply a smooth one dimensional nonparametric function that takes the predictor X_i as it's argument. That is,

$$f(\mathbf{X}_i) = \mu + \sum_{i=1}^{p} m_i(X_i)$$

Since my estimation method relies on the R package 'GAM', each of my predictor functions will be a nonparametric local regression function. R's 'GAM' package estimates these by using the backfitting algorithm and is as follows:

1. calculate the mean of the response $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
2. repeat the following steps on each predictor j, until a prespecified level of convergence is reached
   (a) estimate $\hat{m}_j(x_j)$ by fitting it to $\{Y_i - \hat{\mu} - \sum_{k\neq j} \hat{m}_k(x_k)\}_1^n$
   (b) set $\hat{m}_j$ as $\hat{m}_j(x) - \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_{i,j})$, essentially mean centering the data.

The model I estimated includes all four features as inputs since this resulted in the lowest 2 fold crossvalidation prediction RMSE. The resulting CI's for each point again must be transformed such that each CI is centered at zero in order to visualize them, as was done above. The plot appears on the next page. These confidence intervals are pointwise estimates about the estimated response, which is biased.



My choice to use the Generalized Additive model was motivated by the dimension of the predictors. As was noted in class, dimensions of 4 or larger don't always work so well for keneral or local linear estimators because of sparsity. That is, the number of points required for having a dense enough dispersion of points increases exponentially with dimension. This would then have a large impact on boundary bias that the kernel regression suffers from. It would also detract from the predictive quality of both kernel and local linear nonparametric regressions.

Below are the RMSE scores of various GAM fits. Some of the subsets aren't reported because their scores were higher than those reported below. THat is they would be even poorer at predicting as a result of having less information.

Features	RMSE	Features	RMSE
AT + V + AP	15.47	V + AP + RH	22.57
AT + V + RH	13.78	AT + AP + V + RH	27.62
AT + AP + RH	16.41		

Discussion & Conclusion

The parametric model had a 2 fold Crossvalidation score of 4.331 and the nonparametric GAM model had an RMSE of 4.56. This is atypical, since the GAM has fewer assumptions on the way the inputs relate to the response. However, it would seem that because of the high correlation between predictor and response variables, the parametric linear model has a chance to outperform the parametric model. Also, the data residuals are such a good fit to the normal as well. Even after taking the points that were removed in the linear model fitting from the GAM, its' RMSE only reduced to 4.4. So had the data been further from the assumptions of the linear model than I would expect the GAM to have done a better job at prediction.

From the perspective of the electric company however, I would imagine that they would prefer the GAM since it may not require the removal of odd points. Hence it would be easier to implement, especially on the real time continuing basis as the power plant operates. Also since the difference between the two is relatively small.

Another difference between the approaches is interpretability of the models. The linear model preserves the ability of interpreting the relationship between the features and response. That is, we can think of a one unit increase in predictor i as an $\hat{\beta}_i$ increase in the expected response holding all else constant. On the other hand, the GAM looses the ability at interpreting the affect that a

feature has on the response. In prediction neither of these may matter too much unless you have an ability to control the features themselves. And for the electric company the variable vacuum exhaust might be. Also, the Ambient pressure, temperature and relative humidity variables may potentially be influenced.

A benefit of the GAM model though is it's ability to be fit without necessarily considering outliers or points that don't conform to the more strict assumptions of the parametric error distribution. From this perspective the GAM may have a better chance at prediction in the future. That is, a non-technical user wouldn't have to 'clean' the data periodically.

Since it appears to be slightly inconclusive as to which model is better, I would like to suggest some further work that might help resolve the issue. That is, first I would attempt to gather data from other CCPPs and look to see if the normality of data continued to work in favor of the linear model as was the case for this CCPP. I would also like to see which method the electric company would prefer using from an implementation perspective.

Bibliography

- Pinar Tufekci, Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods, International Journal of Electrical Power & Energy Systems, Volume 60, September 2014, Pages 126-140, ISSN 0142-0615.

- Heysem Kaya, PÄśnar TÃijfekci , SadÄśk Fikret GÃijrgen: Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE 2012, pp. 13-18 (Mar. 2012, Dubai)