Forecasting Wind Ramps

Erin Summers and Anand Subramanian

Jan 5, 2011

1 Introduction

The recent increase in the number of wind power producers has necessitated changes in the methods power system operators employ to balance load and generation. Furthermore, the number of renewable energy producers is also expected to increase in the US to 20% of total generation by the year 2030 [1]. Wind power, unlike conventional fossil-fuel based generation, is intermittent and uncertain. Therefore, it is crucial that this variability is quantified so that power system operators can schedule appropriate amounts of reserve generation to compensate for fluctuations in wind production.

Ramps are large, sudden increases or decreases in wind power production. These events are of prime concern to power system operators because they threaten the overall stability of the grid. These sudden changes in generation levels effectively limit the operational capabilities of the power grid. Hence, information about ramps and ramp prediction is highly valuable to power system operators.

A variety of methods have been used to provide reliable ramp predictions. Details of state-of-the-art techniques are not readily available as wind forecasting for power systems is a lucrative business. In this study, we attempt to develop plausible models for ramp behavior using techniques from statistical learning theory. We first motivate the intuition behind using specific models to explain wind ramps. Then we create and validate these models with actual wind power production data obtained from the Bonneville Power Authority (BPA). We specifically examine three kinds of models:

1. k-means clustering
2. Gaussian mixture models (GMMs)
3. hidden Markov models (HMMs)

In Section 2, we describe our algorithm to identify wind ramps from time series data. Sections 3, 4, and 5 contain results from models developed to describe wind ramp behavior using k-means clustering, Gaussian mixture models (GMMs) and hidden Markov models (HMMs) respectively. We conclude by outlining possible extensions of this study in Section 6.

2 Ramp Identification

Based on work done to define and categorize wind ramps in [5], we characterize each ramp according to two parameters:

1. Absolute change in wind magnitude over ramp ΔM
2. Duration of ramp ΔT.

This parametrization is useful as it allows for a compact representation of wind ramping events. However, in order to obtain values of (ΔM,ΔT), we must first identify wind ramps from time series data of wind power generation.

Initially, the wind data is normalized with respect to the total capacity of wind production. The data is then smoothed using a moving-average filter to eliminate high frequency noise. Next, a recursive bisection is run on the data, which splits the data into ramp and non-ramp intervals. Given a particular time interval with N points, the bisection algorithm first calculates the mean (μ) and the unnormalized variance (1) of the slope of time series data x_k.

$$\sigma^2 := \sum_{k=1}^{N-1} \|g_k - \mu\|^2, \text{ where } g_k = x_{k+1} - x_k \tag{1}$$

If σ^2 exceeds a specified threshold, the interval is split in half and the process is repeated until all intervals obey this variance threshold.

This recursive bisection method introduces a bias in the way ramp intervals are found, since intervals are always divided in half. In order to remove this bias, we perform a post-processing step a number of times. In this algorithm, each interval is extended as far as possible while the gradient over the interval satisfies the variance condition shown in (1). A brief explanation of this process follows:

Let $\mathbf{I}(X,Y) := X \bigcup Y$ represent the function that combines intervals X and Y, preserving the order. Let $\bar{X}_i$ represent the interval X, excluding the last i data points. Let L_X represent the length of the interval X.

Given a particular interval A and its time consecutive interval B, the unnormalized variance, σ^2 for $\mathbf{I}(A,B)$ is calculated. If σ^2 of $\mathbf{I}(A,B)$ is below a specified threshold, then intervals A and B are combined to form a new interval. If σ^2 of $\mathbf{I}(A,B)$ is above a specified threshold, σ^2 is found for $\mathbf{I}(A,\bar{B_i})$ for $i = 1, \dots L_B$, until either the variance condition is met (in which case, A and B_i are combined), or all of the sub-intervals are tested. The part of B that is not combined is set as A for the next iteration of the algorithm and the algorithm runs until all intervals have been examined.

Ramps are then defined to be intervals over which the gradient of the time series signal exceeds some prespecified value. Identified ramps for a section of the data are shown in Figure 1, where the ramps are identified in blue. (ΔM,ΔT) for ramp intervals are calculated and analysis is performed on this low-dimensional representation of ramping events.

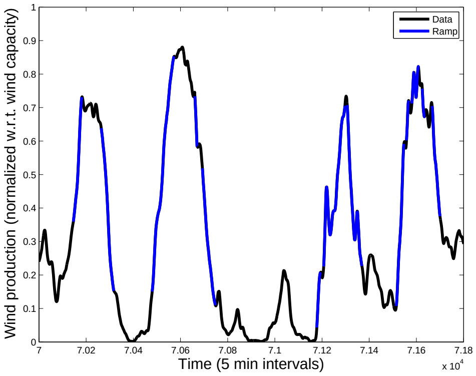

Figure 1: Ramps Identified from Wind Data

3 k-means clustering

As we want to find trends in identified ramps, we first attempt to find a small number (k) of ramp clusters to explain the observations. The k-means clustering algorithm was the first method used to identify these ramp clusters. We tried this algorithm for values of k, the number of clusters, varying from 2 to 20. In order to identify an optimal value of k we associate the following cost to pairs of clusters:

$$\Delta(A,B) := \sum_{i\in A\cup B}(x_i-\mu_{A\cup B})^2-\sum_{i\in A}(x_i-\mu_A)^2-\sum_{i\in B}(x_i-\mu_B)^2=\frac{n_An_B}{n_A+n_B}\|\mu_A-\mu_B\|_2^2 \qquad (2)$$

Here, μ_i is the mean for cluster i while n_i is the number of points assigned to cluster i. This metric (which we call Ward Distance) is the same one used in Ward's method, a well-known hierarchical clustering algorithm [8]. This quantity measures how much the sum of squares of deviation between cluster data points and cluster means associated with two clusters A and B changes if those two clusters are merged. A small Ward distance for a particular pair of clusters indicates that one cluster is sufficient as opposed to two. Therefore, we stop adding clusters if there is a sizable decrease in this metric when a new cluster is added.

For each k, the k-means algorithm is run from 50 different random initial conditions. The cluster means, and the minimum Δ for all pairs of clusters are recorded for the run with the lowest *distortion measure* (3).

$$J := \sum_{n=1}^{N}\sum_{i=1}^{K} z_n^i\|x_n-\mu_i\|^2 \qquad (3)$$

where $z_n^i = 1$ if x_n belongs to cluster i and $z_n^i = 0$ otherwise. Table 1 lists the minimum Ward distance (Δ) for several k values.

Comparing the Δ values for $k = 4$ and $k = 5$, we find that the addition of the 5th cluster does not add much value in terms of explaining the data. We also decided as an alternative threshold to pick the model

k	min Δ
2	29.2521
4	1.7908
5	0.1913
8	0.1280
9	0.0998
10	0.0325
15	0.0020

Table 1: Table of Ward Distances from K-means

with the largest number of clusters for which the minimum Ward distance was greater than 0.1. This situation was reached when $k = 9$. Accordingly, we decided on using $k = 4$, due to the sharp decrease in minimum Ward distance, and $k = 9$, due to the 0.1 threshold on minimum Ward Distance, to build cluster models to represent the data. Figure 2 visualizes the clustering of the data for these two cases.

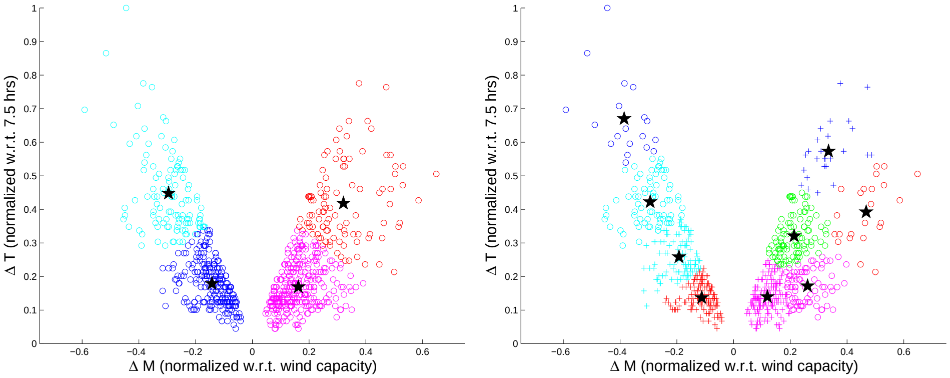

Figure 2: K-means with $k = 4$ (left) and $k = 9$ (right) clusters

4 Gaussian Mixture Model

While k-means clustering was an informative first step, the deterministic nature of allocating observations to clusters is not ideal. In order to overcome this limitation of k-means clustering, we fit a Gaussian Mixture Model (GMM) with a preset number of mixture components, denoted K. We assume a bivariate normal distribution for the emission probabilities $p(y_t|q_t = i)$ with means μ_i and variance Σ_i. Note that y_t refers to the observations while q_t refers to the hidden state. It should be noted that this is perhaps not the best model for the data, since $\Delta T \geq 0$ in our model. Nevertheless, a GMM will provide us an informative starting point for data analysis so long as the probability mass associated with the domain

$\Delta T < 0$ is small.

Our data set is clearly separated into two separate categories: increase ramps and decrease ramps. We are ultimately interested in further classifying these two categories separately. Accordingly, the means and variances of the increase ramps should not affect the update of the means and variances for the decrease ramps and vice versa. Thus, the training data is separated into two data sets and the standard EM update iterations for GMM are performed on the two data sets separately. We initialize the means of the K mixture components of the GMM with the cluster means obtained from k-means clustering.

Once the model is developed for the various values of K, the number of mixture components, we validate the models by evaluating the log likelihood of different data. Figure 3 shows this quantity for different values of K. The values of K which were considered interesting in the k-means clustering analysis correspond to the red triangles. Again, we find that models with $K = 4$ and $K = 9$ have larger likelihoods than those obtained for neighboring values of K.

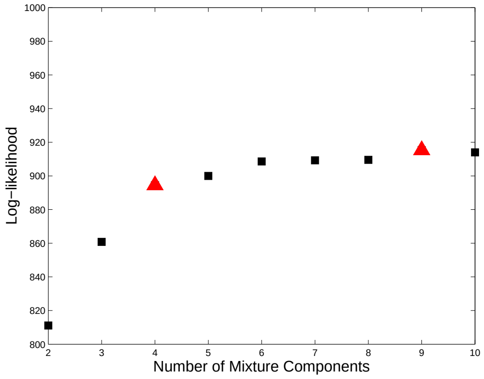

Figure 3: Log-likelihood of GMM on Test Data for different numbers of clusters

Using a method similar to the expectation step of the EM algorithm, we also compute a quantity we call the *expected Ward Distance*, which is an analog to the Ward distance seen in the k-means clustering section.

Here, we outline the derivation of the expected Ward distance. Define $\tau_t^i := \mathbb{P}(q_t^i = 1|y_t)$, to be the probability that given a data point y_t, the *ith* component of q_t is equal to 1. Hence, the expected number of points in a cluster C, denoted as n_C, can be written as

$$\mathbb{E}[n_C|y] = \sum_t \mathbb{E}[\mathbb{I}(q_t^C = 1)|y_t] = \sum_t \mathbb{P}(q_t^C = 1|y_t) = \sum_t \tau_t^C. \tag{4}$$

Then, the expected Ward distance $\bar{\Delta}(A, B)$ between arbitrary clusters A and B easily follows:

$$\bar{\Delta}(A, B) := \mathbb{E}[\Delta(A, B)] = \mathbb{E}\left[\frac{n_A n_B}{n_A + n_B}\|\mu_A - \mu_B\|_2^2\right] = \frac{\sum_t \tau_t^A \sum_t \tau_t^B}{\sum_t \tau_t^A + \sum_t \tau_i^B}\|\mu_A - \mu_B\|_2^2. \tag{5}$$

Figure 4 shows the expected Ward distance for different values of K. Much like the k-means case, there is a sharp decrease between $K = 4$ and $K = 5$ indicating that a 5th cluster would be excessive. If more model complexity is desired we look for the next sharp decrease in expected Ward distance and this occurs between $K = 9$ and $K = 10$. Therefore, this analysis again supports using those two values of $K = 4$ and $K = 9$, to create GMMs to represent the data.

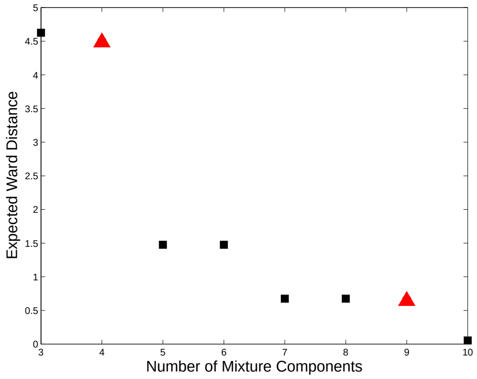

Figure 4: Expected Ward distance for different number of mixture components. Red triangles indicate models with $K = 4$ and $K = 9$.

5 Hidden Markov model with Gaussian emission probabilities

The models considered until this point assume no dynamics between the hidden states q_t representing the underlying nature of the ramps. However, one expects ramp behavior to have some temporal correlations. For instance, a very large increase in wind generation should be followed, some time later, by a decrease of some magnitude. Accordingly, a hidden Markov model (HMM) is developed to analyze any temporal correlations and is validated against a different set of data.

Model Description

The graphical model for an HMM is shown in Figure 5 for completeness. The emission probabilities $p(y_t|q_t = i)$ are assumed to be bivariate normal distributions with means μ_i and Σ_i. This normality assumption was determined to be reasonable in the previous section even though the range of possible values for ΔT is bounded. We also assume that underlying states can represent either increase ramps or decrease ramps only. Similar to the Gaussian mixture model case, only increase ramp data is used to update means and covariances of states assigned to describe increase ramps. Therefore, the number of states corresponding to increase or decrease ramps is determined before parameter estimation via the initialization of means.

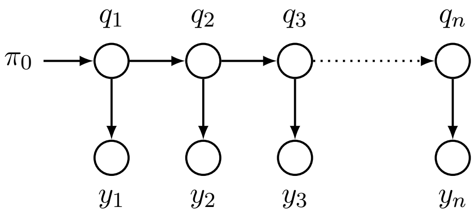

Figure 5: Hidden Markov Model

The number of total HMM states, K, is varied from 3 to 10, and they are equally divided into states representing increase or decrease ramps based on state initialization. The EM algorithm, specifically the alpha-gamma algorithm, is used to estimate the model parameters. The update equations for this algorithm are found in Chapter 12 of the CS 281A course reader while the update equations for the emission means and covariances are similar to those used in the Gaussian mixture model section.

Results

The models obtained for different values of K were validated by computing the log-likelihood of another data set. Figure 6 shows a marked increase in the log likelihood from $K = 3$ to 4 and then a very modest increase from that point onward. This implies that the 4 state HMM is sufficient to explain the variation in ramps.

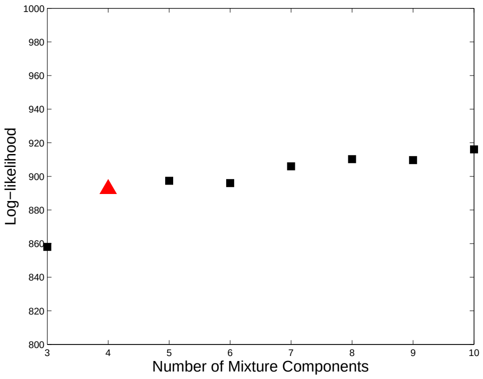

Figure 6: Log-likelihood for HMMs with different numbers of underlying states. The triangle represents the HMM with 4 states which is chosen for further analysis.

The HMM developed with K, the number of states, $=4$ was particularly helpful in analyzing how ramp behavior changes with time. Figure 7 shows the means and Gaussian sub-level sets (in red) for this case. It is interesting to note that these states correspond to 4 distinct types of ramps:

1. Small, short increase ramps
2. Large, long increase ramps
3. Large, long decrease ramps
4. Small, short decrease ramps

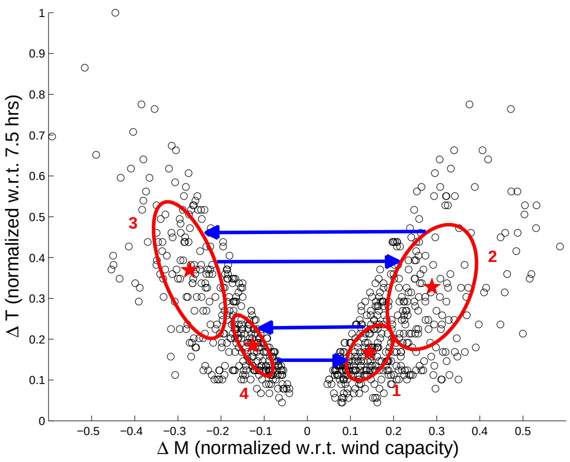

Figure 7: HMM Model with 4 Clusters. Blue arrows denote the 4 largest transition probabilities in the state transition matrix

The state transition matrix for this model is shown in Table 2. First, it is apparent that the large entries are in the upper right and lower left 2x2 blocks. This implies that transitions from increase ramps to decrease ramps and vice versa are more probable events than remaining in an increase or decrease ramp state. Furthermore, as the antidiagonal of the matrix contains the largest transition probabilies, we can infer the following:

1. Large, long increase ramps (2) are followed, with high probability, by large, long decrease ramps (3) and vice versa.
2. Small, short increases (1) are usually followed by small, short decreases (4).

6 Future Work

While GMMs and HMMs have provided a reasonable description of the wind data, they may not be appropriate since our domain is bounded. In the future, we will investigate other bivariate distributions that are bounded, namely the *Weibull distribution* and the *Gamma Distribution*.

	Current State			
Future State	1	2	3	4
1	0.1	0.09	0.26	0.82
2	0.08	0.19	0.6	0.1
3	0.16	0.59	0.08	0.04
4	0.67	0.13	0.06	0.04

Table 2: State transition probability matrix for HMM

The Weibull distribution is parameterized by a shape parameter k and a scale parameter $\lambda > 0$, with pdf

$$p(x, |\lambda, k) = \begin{cases} \frac{k}{\lambda}\frac{x}{\lambda}^{(k-1)} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & else \end{cases}$$

There is ample literature describing models of wind and wave data classification using a Weibull distribution [7],[6], [3], [2].

Likewise, a Gamma Distribution, of which there are many bivariate forms [4], is also a promising distribution and does not violate the domain constraints. Characterized by shape parameter k and scale parameter θ, the probability density function is

$$p(x|k, \theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)} \text{ for } x \geq 0, k, \theta > 0.$$

The histogram of ΔM and ΔT in Figure 8 reveals that the density of the data is similar to that of both the Weibull and Gamma Distributions.

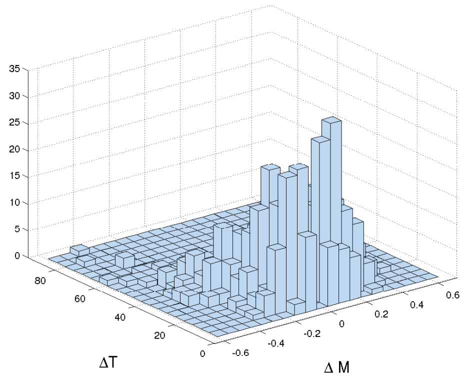

Figure 8: Histogram of ΔM and ΔT

In the future we will plan to enhance our algorithm to improve ramp identification.

References

[1] J. DOE. 20 % wind energy by 2030: Increasing wind energy's contribution to US electricity supply. *Washington, DC*, 2008.

[2] C. Guedes Soares and M. Scotto. Modelling uncertainty in long-term predictions of significant wave height. *Ocean Engineering*, 28(3):329–342, 2001.

[3] Z. Huang and ZS Chalabi. Use of time-series analysis to model and forecast wind speed. *Journal of Wind Engineering and Industrial Aerodynamics*, 56(2-3):311–322, 1995.

[4] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Multivariate Distributions, volume 1, Models and Applications*. New York: John Wiley & Sons,, 2002.

[5] C. Kamath. Understanding wind ramp events through analysis of historical data. In *Transmission and Distribution Conference and Exposition, 2010 IEEE PES*, pages 1–6. IEEE, 2010.

[6] L. Landberg and S.J. Watson. Short-term prediction of local wind conditions. *Boundary-Layer Meteorology*, 70(1):171–195, 1994.

[7] P. Ramirez and J.A. Carta. Influence of the data sampling interval in the estimation of the parameters of the Weibull wind speed probability density distribution: a case study. *Energy conversion and management*, 46(15-16):2419–2438, 2005.

[8] J.H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.