Mean Variance Mapping Optimization for the Identification of Gaussian Mixture Model: Test Case

Francisco Gonzalez-Longatt
Faculty of Computing and Engineering
Coventry University
Coventry, United Kingdom

fglongatt@ieee.org

José Rueda, István Erlich, Walter Villa
Institute of Electrical Power Systems,
University Duisburg-Essen
Duisburg Germany

jose.rueda@uni-due.de,
istvan.erlich@uni-due.de

Dimitar Bogdanov
Faculty of Electrical Engineering
Technical University of Sofia
Sofia, Bulgaria

dbogdanov@tu-sofia.bg

Abstract— **This paper presents an application of the Mean-Variance Mapping Optimization (MVMO) algorithm to the identification of the parameters of Gaussian Mixture Model (GMM) representing variability of power system loads. The advantage of this approach is that different types of load distributions can be fairly represented as a convex combination of several normal distributions with respective means and standard deviation. The problem of obtaining various mixture components (weight, mean, and standard deviation) is formulated as a problem of identification and MVMO is used to provide an efficient solution in this paper. The performance of the proposed approach is demonstrated using two tests. Results indicate the MVMO approach is efficient to represented load models.**

Gaussian mixture Model; Load Modeling; Mean Variance Mapping Optimization Algorithm; Optimization

I. Introduction

The modeling of power system loads is a complex task due to the stochastic nature of the demand, and it is becoming more and more challenging in future networks where the penetration levels of distributed generation system are expected to be extraordinary high. Accurate model models of power system loads are essential for planning studies and operation. There are several techniques to model the loads. However, the most common is model the loads through *Gaussian distribution*. Studies have demonstrated the single Gaussian assumption is not always justified for all the loads since the statistical distribution of electric load variations may not strictly follow any common probability distribution function [1].

A significant research effort has been devoted to load variation probabilistic modeling using different *probabilistic distributions functions* (PDFs): Normal [2], Log-Normal [3], Gamma [2], Gumbel, Inverse-normal [2], Beta [4], [5], Exponential [6], Rayleigh, and Weibull [7]. An important conclusion that can be derivate from a literature survey is that there is not a unique or generalized technique to model the load PDF [8].

So far the use of Gaussian model of a load profile has been widely used for various reasons: (i) simplicity as it can be described using two parameters: *mean* (μ) and *standard deviation* (σ), (ii) the analysis of this PDF is the most developed and documented in the literature. In recent times, publications in several areas such as: finance, biometric, biology and most recently electrical engineering have used the concept of a parametric probability density function represented as a weighted sum of Gaussian component densities, *Gaussian Mixture Model* (GMM) [9].

R. Singh et al [8] introduced the statistical modeling of the loads in distribution networks through GGM and the *expectation maximization* (EM) algorithm was used to obtain the parameters of the mixture components. Despite of its conceptual simplicity, the EM algorithm may have difficulties in handling problems whit high dimensionality.

The advantage of GMM approach is that different types of load distributions can be fairly represented as a convex combination of several normal distributions with respective means and standard deviation. The problem of obtaining various mixture components (weight, mean, and variance) can be formulated as a problem of identification where optimization methods provide an efficient solution. Hence, based on the success gained in previous applications to different power system optimization problems, this paper presents an application of the *Mean-Variance Mapping Optimization* (MVMO) algorithm to the identification of the parameters of GMMs representing variability of power system loads. MVMO shares certain similarities to other heuristic approaches, but possesses a special mapping function through which a new offspring generated in every update is always inside the respective bound, since it is always inside the range [0,1]. The shape and location of the mapping curve are adjusted according to the progress of the search process, and MVMO updates the candidate solution around the best solution in every iteration step. Thanks to the well-designed balance between search diversification and intensification, MVMO can find the optimum quickly with minimum risk of premature convergence [10], [11].

The remaining sections of the extended abstract are organized as follows: Section II presents the GMM model. Section III describes problem formulation and the adaptation of MVMO to tackle the identification task. Finally, Section IV presents results using load samples from the Venezuelan grid.

II. GAUSSIAN MIXTURE MODEL

A *Gaussian Mixture Model* (GMM) is a parametric PDF represented as a weighted sum of Gaussian probabilistic densities. The GMM is a weighted sum of N_C component Gaussian densities as given by the equation:

$$p\left(\mathbf{x}|\lambda\right)=\sum_{i=1}^{N_C} w_i g\left(\mathbf{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right) \tag{1}$$

where **x** is a D-dimensional continuous-valued data vector (i.e. measurement or features), w_i, i = 1, ..., N_C, are the mixture weights, and $g(\mathbf{x}|\boldsymbol{\mu}_{\mathrm{i}},\boldsymbol{\Sigma}_i)$, i = 1, ..., N_C, are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form:

$$g\left(\mathbf{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)=\frac{1}{(2\pi)^{\frac{D}{2}}\left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}}\mathrm{e}^{\left\{-\frac{1}{2}(x-\mu_i)'\boldsymbol{\Sigma}_i^2(x-\mu_i)\right\}} \tag{2}$$

with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The mixture weights satisfy the constraint that sum of all the weights must equal to one.

$$\sum_{i=1}^{N_C} w_i = 1 \tag{3}$$

This condition is because a PDF must be nonnegative and the integral of a PDF over the sample space of the random quantity it represents must evaluate to unity.

III. PROPOSED MVMO-BASED IDENTIFICATION OF GMM

A. Problem statement

The GMM parameter estimation approach presented in this paper, based on the chi-square goodness-of-fit test, is defined as follows:

Minimize

$$\chi^2=\sum_{i=1}^{\mathrm{n}}\frac{\left(O_{\mathrm{i}}-E_i\right)^2}{E_i} \tag{4}$$

subject to

$$h=0 \tag{5}$$

where χ^2 stands for *Pearson's cumulative test statistic*, which asymptotically approaches a χ^2 distribution. O_{i} and E_{i} denote observed frequency and the expected frequency (asserted by the null hypothesis), respectively; h is a binary variable that indicates whether the null hypothesis can (h=1) or cannot (h=0) be rejected at the 5% significance level.

B. Solution through MVMO

The theoretical background of MVMO has been published in [10]-[12]. MVMO operates on a single solution rather than a set of solutions like in many evolutionary algorithms. The goal is to perform the optimization with a minimum amount of objective function evaluations (e.g. solving differential equations). The procedure of MVMO for solving the GMM identification problem with D parameters to be identified is summarized in Fig. 1.

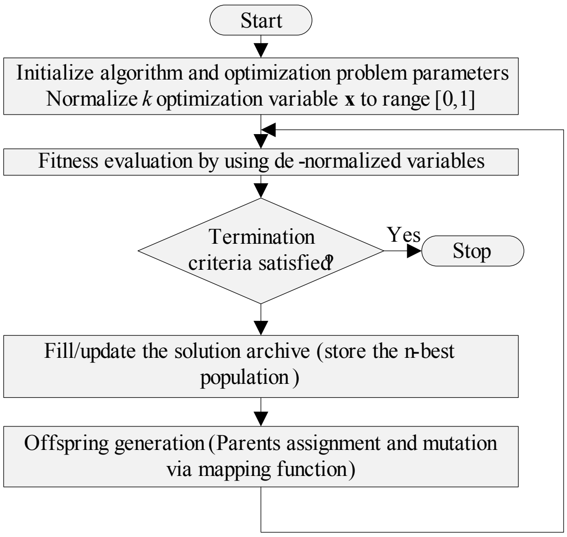
Start
Initialize algorithm and optimization problem parameters
Normalize k optimization variable x to range [0,1]
Fitness evaluation by using de -normalized variables
Termination criteria satisfied?
Yes
Stop
Fill/update the solution archive (store the n-best population)
Offspring generation (Parents assignment and mutation via mapping function)

Figure 1. General flowchart for the MVMO algorithm. MVMO implementation procedure for identification of GMM parameters.

The procedure starts with initialization of the parameters of the algorithm and the definition of the normalized initial guess of the control variable, since the internal searching space of all variables in MVMO is restricted in [0, 1]. Hence, the real min/max boundaries of variables have to be normalized to 0 and 1. Next, the solution archive, which constitutes the knowledge base for search, is filled/ updated at the first/ successive function evaluations. Mean variance and shape factors are also computed for every optimization variable at this stage. Finally, creation of an offspring is performed, which involves selection of m of k dimensions of the optimization problem and mutation operation on the selected variables through a special mapping function. During the iteration it is not possible that any component of the solution vector will violate the corresponding boundaries. This goal is achieved by using the mapping function. The inputs of this function are mean and variance of the best solutions that MVMO has discovered so far. The elegant property of MVMO is the ability to search around the local best-so-far solution with a small chance of being trapped into one of the local optimums. This feature is contributed to the strategy for handling the zero-variance.

Fitness evaluation and constraint handling: For each individual, the chi-square goodness-of-fit test is performed, the feasibility of the solution is checked and a fitness value is assigned. It is considered that an individual is better if the fitness is smaller. The static penalty scheme is used in this paper to handle constraints. Since the control variables in **x** are self-restricted, all dependent variables are constrained by applying the integrated fitness function as follows:

$$\min f'=f+\sum_{i=1}^{n} \upsilon_{\mathrm{i}} \max\left[0, \mathrm{g}_{\mathrm{i}}\right]^{\beta} \tag{6}$$

The where f is the original objective function, n is the number of constraints, β is the order of the penalty term (e.g. 1 or 2), υ_i is the penalty coefficient of the i-th constraint and g stands for inequality constraint. It is worth mentioning that other constraint handling techniques are also applicable to MVMO.

Termination criteria: In this paper, the MVMO search process is terminated based on completion of a pre-specified number of fitness evaluations.

Solution archive: The solution archive constitutes the knowledge base of the algorithm for guiding the searching direction. Hence, the best n individuals that MVMO has found so far are saved in the SA. Additionally, two relevant information parameters, namely fitness and feasibility of each individual are also stored. The following rules are set up to compare the individual generated at each iteration and existing archived solutions in order to avoid losing good solutions [11]: (i) Any feasible solution is preferred to any infeasible solution, (ii) between two feasible solutions, the one having better objective value is preferred, (iii) between two infeasible solutions, the one having smaller fitness value (i.e. smaller constraint violation) is preferred.

An update takes place only if the new individual is better than those in the archive. The archive size is fixed for the entire process. The archived individuals are dynamically sorted so that the first ranked individual is always the best. Feasible solutions are placed in the upper part of the archive. Among these solutions, they are sorted based on their original objective values. Infeasible solutions are sorted according to their fitness values and then placed on the lower part of the archive. Once the archive is filled up by n feasible solutions, any infeasible candidate solution does not have chance to be saved in the archive.

Parent assignment: The first ranked (best-so-far) solution denoted as $\mathbf{x}_{\mathbf{best}}$ is assigned as the parent.

Variable selection: The MVMO searches around the mean saved in the archive for the better solution only in m selected directions. This means that only these dimensions of the offspring will be updated while the remaining D-m dimensions take the corresponding values from $\mathbf{x}_{\mathbf{best}}$. In this paper, a random sequential selection strategy was implemented.

Mutation: For each of the m-selected dimension, mutation is used to assign a new value of that variable. Given a uniform random number $x'_i \in [0,1]$ the new value of the i-th component x_{i} is determined by:

$$x_{\mathrm{i}} = h_{\mathrm{x}} + \left(1 - h_1 + h_0\right) x'_{\mathrm{i}} - h_0 \tag{7}$$

where h_{x}, h_1 and h_0 are the outputs of the transformation mapping function based on different inputs given by:

$$h_{\mathrm{x}} = h\left(u_{\mathrm{i}} = x'_{\mathrm{i}}\right),\quad h_0 = h\left(u_{\mathrm{i}} = 0\right),\quad h_1 = h\left(u_{\mathrm{i}} = 1\right) \tag{8}$$

The mapping function is parameterized as follows:

$$h\left(\bar{x}_{\mathrm{i}}, s_{\mathrm{i}1}, s_{\mathrm{i}2}, u_{\mathrm{i}}\right) = \bar{x}_{\mathrm{i}}\left(1 - e^{-u_{\mathrm{i}} s_{\mathrm{i}1}}\right) + \left(1 - \bar{x}_{\mathrm{i}}\right) e^{-(1-u_{\mathrm{i}}) s_{\mathrm{i}2}} \tag{9}$$

where $s_{\mathrm{i}1}$ and $s_{\mathrm{i}2}$ are shape factors allowing asymmetrical slopes of the mapping function. The slope is calculated by:

$$s_{\mathrm{i}} = -\ln\left(v_{\mathrm{i}}\right) f_{\mathrm{s}} \tag{10}$$

where f_{s} is a scaling factor, which enables the control of the search process during iteration. Interested readers are referred to [2] for further details on how to set the two different shape factors and the scaling factor as well.

IV. Simulations and Results

In this section, the some tests are performed to the MVMO algorithm in order to evaluate the capability to identify the parameters for the GGM. A specific MATLAB®[13] program is developed by the authors for such propose. All simulations are performed using a personal computer based on Intel®, Core™ i7 CPU 2.0GHz, 8 GB RAM with Windows® 7 Home Edition 64-bit operating system.

Two different tests are used in this section: synthetic data and real data. A first test is based on a set of synthetically created data; the MVMO algorithm is used to identify the parameter: weights, means, and standard deviation. The data is created based on a pre-defined full component GMM. A simple comparison of the parameters obtained from the proposed approach and the original values supposed during the synthetic data creation is used to define the goodness-fit. The second test is performed using real data, the aims is demonstrate the suitability of the proposed method to model complex load models.

A. Synthetic data

A probabilistic distribution function for a hypothetical load that consists of a mixture of three components (N_C = 3) is used for test in this section. Table I shows the parameters of GMM components and Fig. 2 shows the original PDF, GGM PDF and the individual components.

TABLE I. Parameters of the GMM components: synthetic data

Gaussian PDF No.	*Weight w_i (p.u.)*	*Mean μ_i (kW)*	*Std σ_i (kW)*
1	0.3500	705.3900	119.6919
2	0.1200	1163.2800	155.5662
3	0.0530	465.2800	80.1255

The performance of the MVMO algorithm is evaluated comparing the parameters identified for the GMM using this approach versus the known parameters shown on Table I, absolute error is used as effective measurement for the performance. Table II shows the errors for the weights (w_i) is the slowest comparing with the results on other parameters. The discrepancies obtained using MVMO are almost negligible for the mean (μ_i) and standards deviation (σ_i)

results.

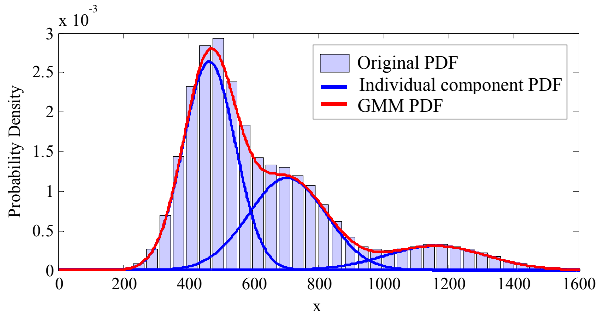

Figure 2. Probabilistic distribution density for a hypothetical load: Data created synthetically.

TABLE II. PERFORMANCE OF MVMO ALGORITHM FOR THE IDENTIFICATION OF GMM

Gaussian PDF No.	*Error* *Weight w_i (p.u.)*	*Error* *Mean μ_i (kW)*	*Error* *Std σ_i (kW)*
1	$0.876{\times}10^{-9}$	$1.350{\times}10^{-2}$	$0.135{\times}10^{-3}$
2	$0.658{\times}10^{-9}$	$0.118{\times}10^{-3}$	$0.227{\times}10^{-3}$
3	$0.797{\times}10^{-9}$	$0.750{\times}10^{-2}$	$0.333{\times}10^{-4}$

B. Real data: Venezuelan test case

Venezuela's power system is an integrated vertical power company, called *Corporación Electrica Nacional* (Corpoelec), which covers most of the country. The Paraguaná Peninsula transmission system is fed from a single circuit (230 kV) transmission line of San Isidro substation as part of the Venezuelan power pool. The average demand is 280 MW and importation by San Isidro tie line is 200 MW. Fig. 3 shows all substations, transmissions lines, static reactive compensators, and generators of the Paraguana's power system.

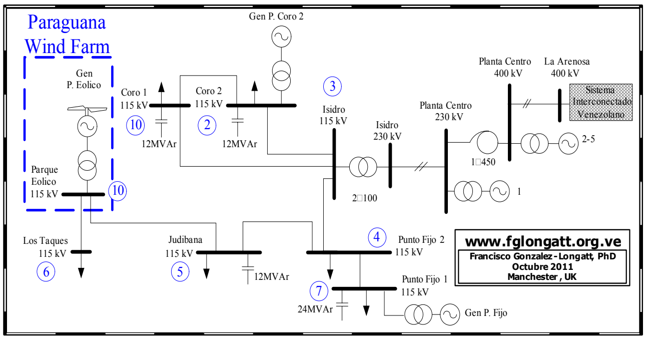

Figure 3. Representative single-line diagram of the Paraguaná transmission system.

Punto Fijo and Judibana substation are very important loads of the Paraguaná power system, for this reason these substations are selected for testing the proposed approach. Real data measurements for two years (hourly basis) of those substations are used.

The load behavior on those substations is chaotic and difficult to be stochastically modeled. Several attempts to create a model using different PDFs reveal such models are not statistically representative for the load behavior. An attempt using *cumulative distribution functions* (CDF) provide better results (see Fig. 4 and Fig. 5), however, *Kolmogorov-Smirnov* (KS) test has been performed for all distribution and results indicate that no single PDF would entail a good fit goodness to explain the variation of the measured load active power at those substations. It is evident these time-series of active power result a good test for the proposed approach.

C. GMM results and conclusions

Figure 6 illustrates the average convergence behavior (after 100 repetitions of the optimization) of the objective function and the number of mixture components needed to define the best GMM fit describing the variability of the loads at Punto Fijo Substation. Note that MVMO is very fast in the global search capability because the lowest χ^2 has been found after 750 objective function evaluations.

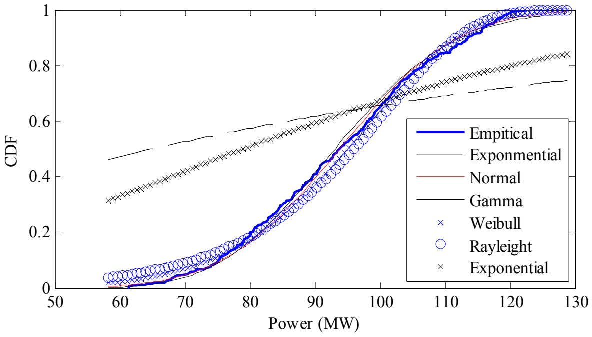

Figure 4. ECDF of the load at Punto Fijo Substation and CDFs of various probability distributions with the same average value and standard deviation.

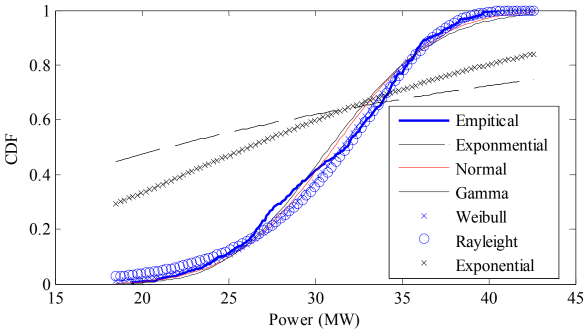

Figure 5. ECDF of the load at Judibana substation and CDFs of various probability distributions with the same average value and standard deviation.

Fig. 7 and Fig. 8 show that the full component merging of the weighted mixture components determined through MVMO optimization provide good approximation of the statistical data collected at two substations of the Paraguaná Transmission System. The GMM fit obtained through the expectation maximization (EM) algorithm is also included in the figure in order to validate the results obtained via MVMO. This estimation can be easily performed in Matlab using the *gmdistribution.fit* command. Note the closeness between the models identified using both MVMO and EM, which indeed highlight the accuracy that can be achieved with the proposed approach. Furthermore, it is worth to mention that the chi-square goodness-of-fit for the MVMO and EM cases was 11.35 and 22.12, respectively for the for the Judibana estimate. Despite of the small difference, it can be concluded that the MVMO estimate provides better estimation accuracy. The parameters of each GMM component (obtained with the proposed approach) are summarized in Table III and Table IV. These results suggest that any load PDF, irrespective of its distribution, can be estimated with a high degree of confidence using the proposed approach.

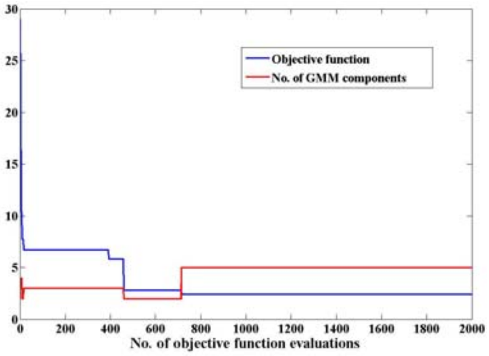

Figure 6. Convergence behavior of the MVMO-based identification of GMM.

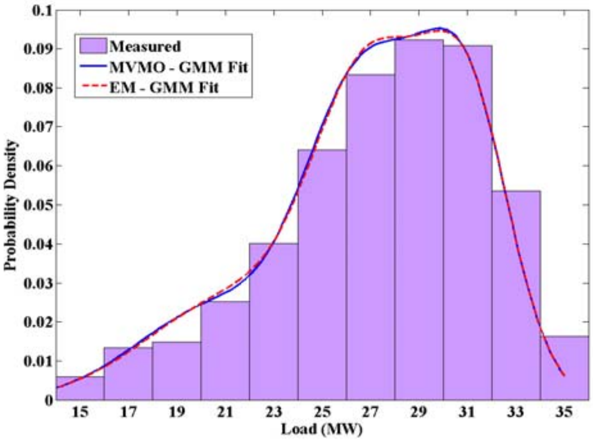

Figure 7. GMM approximation of the load PDF at Punto Fijo substation.

V. Conclusions

An application of the Mean-Variance Mapping Optimization (MVMO) algorithm to the identification of the parameters of Gaussian Mixture Model (GMM) representing variability of power system loads is presented in this paper. This approach provides several advantages; one of them is that different types of load distributions can be fairly represented as a convex combination of several normal distributions with respective means and standard deviation.

In this paper, the problem of obtaining various mixture components (weight, mean, and standard deviation) is formulated as a problem of identification. The novel optimization method, Mean-Variance Mapping Optimization (MVMO) is used to provide an efficient solution. The performance of the proposed approach is demonstrated using two tests. A first test is based on a set of synthetically created data; the MVMO algorithm is used to identify the parameter weights, means, and standard deviation. Results of these tests indicate, and the second test is performed using real data, the aim is demonstrate the suitability of the proposed method to model complex load models. Results indicate the MVMO approach is efficient to represented load models.

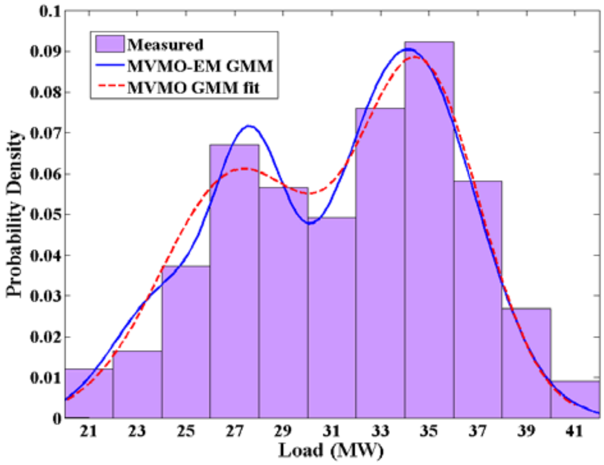

Figure 8. GMM approximation of the load PDF at Judibana substation.

TABLE III. Results of GMM approximation of the load PDF at Punto Fijo substation.

Gaussian PDF No.	*Weight (p.u.)*	*Mean (MW)*	*Std (MW)*
1	0.1790	32.4068	9.2641
2	0.1956	27.6210	2.0460
3	0.4649	34.4864	6.8014
4	0.1605	24.1748	4.5416

TABLE IV. Results of GMM approximation of the load PDF at Judibana substation.

Gaussian PDF No.	*Weight (p.u.)*	*Mean (MW)*	*Std (MW)*
1	0.1665	29.0589	4.8444
2	0.1936	20.6211	10.7520
3	0.1939	26.1117	4.5117
4	0.1922	26.8740	6.3281
5	0.2539	31.1581	3.1041

References

[1] EPRI, "Selected Statistical Methods for Analysis of Load Research Data," EPRI EA-3467, May 1984 1984.

[2] E. Carpaneto and G. Chicco, "Probability distributions of the aggregated residential load," in *Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on*, 2006, pp. 1-6.

[3] A. Seppala, "Statistical distribution of customer load profiles," in *Energy Management and Power Delivery, 1995. Proceedings of EMPD '95., 1995 International Conference on*, 1995, pp. 696-701 vol.2.

[4] R. Herman and J. J. Kritzinger, "The statistical description of grouped domestic electrical load currents," *Electric Power Systems Research,* vol. 27, pp. 43-48, 1993.

[5] S. W. Heunis and R. Herman, "A probabilistic model for residential consumer loads," *Power Systems, IEEE Transactions on,* vol. 17, pp. 621-625, 2002.

[6] L. Hsin-Hui, C. Kaung-Hwa, and W. Rong-Tsorng, "A multivariant exponential shared-load model," *Reliability, IEEE Transactions on,* vol. 42, pp. 165-171, 1993.

[7] G. W. Irwin, W. Monteith, and W. C. Beattie, "Statistical electricity demand modeling from consumer billing data," *Generation, Transmission and Distribution, IEE Proceedings C,* vol. 133, pp. 328-335, 1986.

[8] R. Singh, B. C. Pal, and R. A. Jabr, "Statistical Representation of Distribution System Loads Using Gaussian Mixture Model," *Power Systems, IEEE Transactions on,* vol. 25, pp. 29-37, 2010.

[9] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley and Sons, 2000.

[10] I. Erlich, G. K. Venayagamoorthy, and N. Worawat, "A Mean-Variance Optimization algorithm," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*, 2010, pp. 1-6.

[11] W. Nakawiro, I. Erlich, and J. L. Rueda, "A novel optimization algorithm for optimal reactive power dispatch: A comparative study," in *Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), 2011 4th International Conference on*, 2011, pp. 1555-1561.

[12] I. Erlich, W. Nakawiro, and M. Martinez, "Optimal dispatch of reactive sources in wind farms," in *Power and Energy Society General Meeting, 2011 IEEE*, 2011, pp. 1-7.

[13] MATLAB, *version 7.12.0.635 (R2011a 64-bit)* Natick, Massachusetts: The MathWorks Inc., 2011.