Gaussian-Laplacian Mixture Model for Electricity Market*

Saahil Shenoy[1] and Dimitry Gorinevsky[2]

***Abstract*— This paper develops a statistical modeling and estimation approach combining robust regression and long tail estimation. The approach can be considered as a generalization of Huber regression in robust statistics. A mixture of asymmetric Laplace and Gaussian distributions is estimated using an EM algorithm. The approach estimates the regression model, distribution body, distribution tails, and boundaries between the body and the tails. As an application example, the model is estimated for historical power load data from an electrical utility. Practical usefulness of the model is illustrated by stochastic optimization of electricity order in day-ahead market. The computed optimal policy improves the cost compared to the baseline approach that relies on a normal distribution model.**

I. INTRODUCTION

This paper considers estimation of statistical model that includes the distribution body and long distribution tails. The objective is stochastic optimization using such model.

Estimation of statistical models from data was studied extensively in decision and control, signal processing, statistics, and econometrics. Most common models are based on normal (Gaussian) distributions and lead to versions of least squares estimation. Such models have issues with extreme events that are encountered in many practical problems.

The first issue is that extreme events create outliers that are outside of the normal distribution and can severely bias least squares estimates. There is substantial literature on robust regression methods for handling outliers in the data. One robust regression approach is to change from quadratic loss function in Gaussian model estimation to Huber loss function, which grows linearly for large outliers, e.g., see [16]. Another approach is based on iterative removal of the outliers and update of the solution, e.g., see [4].

The second issue is the need to model long tails distributions for the extreme events. Extreme Value Theory (EVT) predicts that in many cases the distribution tails would asymptotically follow Pareto (power law) or exponential distribution. The last couple of decades saw a number of applications of the EVT methods, e.g., see [10]. The tails can be estimated using peaks-over-threshold method, where tail model is fitted to the data exceeding a threshold. Tail data is usually sparse compared to the distribution body data requiring special procedures for the tail fit. An established approach is Hill's estimator of Pareto tail model [6].

*This work was partially supported by a Seed Grant from TomKat Center for Sustainable Energy at Stanford University

[1]Saahil Shenoy is a PhD student in the Department of Physics, Stanford University, Stanford, CA 94305, USA saahils@stanford.edu

[2]Dimitry Gorinevsky is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA gorin@stanford.edu

The motivating example is stochastic optimization of day-ahead electricity market order. In the example, which is considered in Sections V and VI of this paper, the dependent variable is logarithm of the power load. The exponential (Laplacian) tails in the logarithmic variables correspond to Pareto distribution tails in the original physical variables. Logarithmic variables are commonly used for describing stock returns and power loads [21].

In related recent work [3], a Generalized Pareto model was used for the entire long tail distribution of electricity demand data. Such model cannot be conveniently used for forecasting. Long tail distribution models for electricity markets are considered in [2], [18]; the cited work does not include regression modeling for forecasting. This paper proposes a Laplacian-Gaussian mixture model that allows to estimate the distribution body, the long tails, and the regression model in a Generalized Linear Model (GLM) setting.

There seems to be little prior work that uses Laplacian-Gaussian mixtures. In [14], a linear relationship between the standard deviation of the normal distribution and the scale parameter of the Laplace distribution is assumed for a financial application. Laplacian-Gaussian mixture models were also used for modeling of wind shear data [17] and for speech enhancement [20].

The contributions of this paper are as follows. First, it introduces MALG (mixture of asymmetric Laplace and Gaussian) distribution model as a GLM mixture model, see [15]. MALG model is simple and has few parameters. This makes it useful for on-line decision and control applications.

Second, the paper formulates an optimal Bayesian problem for MALG estimation from data. We propose a version of Expectation Minimization (EM) method [1] for iterative computation of the MALG parameters. E step of the EM uses scalable closed-form expression computations. M step is decomposed into three convex optimization problems for estimating the body, the tails, and the regression parameters.

Third, the paper shows the links between the EM method and known estimators. The regression and body estimation in the MALG model is related to Huber's robust regression. The estimation of the tails is related to Hill's tail estimator.

Finally, the paper applies the results to data modeling and stochastic optimization in day-ahead electricity market. It is shown that using the MALG model instead of a simpler normal model can save a few million dollars per year.

II. MODEL FORMULATION

Consider dataset

$$\mathcal{D} \;=\; \{X_i, y_i\}_{i=1}^{N}\,, \tag{1}$$

where scalars y_i are response variables and $X_i \in \Re^n$ are explanatory variables (regressors). In the motivating example, index i describes time sample; N is the number of samples.

We assume that (1) is an i.i.d. realization of a mixture of asymmetric Laplace and Gaussian (MALG) distributions

$$\begin{aligned} y &= \beta^T X + (1-z)\, v_N + z v_{AL}, && (2)\\ z &\sim B(1,q), && (3)\\ v_N &\sim N(0,\sigma^2), && (4)\\ v_{AL} &\sim AL(0,\lambda_L,\lambda_R), && (5) \end{aligned}$$

where z is Bernoulli random variable with probability q for $z=1$ and $\beta \in \Re^n$ is linear regression parameter vector. Normal (Gaussian) distribution $N(\mu,\sigma^2)$ has mean μ, covariance σ^2, and probability density $p_N(x|\mu,\sigma^2)$. The Asymmetric Laplace distribution $AL(\mu,\lambda_L,\lambda_R)$ has probability density $p_{AL}(x|\mu,\lambda_L,\lambda_R)$ given in Appendix.

The model (2)-(5) is a special case of Generalized Linear Model (GLM) mixture model class, see [5]. The probability density functions (PDF) of the two mixture distribution are

$$\begin{aligned} p(y|X,\theta,z=0) &= p_N(y|\beta^T X,\sigma^2), && (6)\\ p(y|X,\theta,z=1) &= p_{AL}(y|\beta^T X,\lambda_L,\lambda_R). && (7) \end{aligned}$$

This GLM mixture is special in having the same linear model $\beta^T X$ of the distribution means for (6) and (7). In prior work, such as [5], [9], the component distributions in GLM mixture represent clusters of the data centered at different locations. In this work, GLM mixture components represent the body and the tail of the same distribution

$$\begin{aligned} y|X &\sim MALG(\theta), && (8)\\ \theta &= \begin{bmatrix}\beta^T & q & \sigma & \lambda_L & \lambda_R\end{bmatrix}^T, && (9) \end{aligned}$$

where MALG combines the normal (N) distribution PDF p_N, (4) and the Asymmetric Laplace (AL) PDF p_{AL}, (5) as

$$p_{MALG}(y|X,\theta) = (1-q)p_N(y) + q\, p_{AL}(y). \qquad (10)$$

We will use an example MALG distribution to illustrate the approach in this paper. In the example, $X=1$, $\beta^T X = \beta_1 \cdot 1$, and the distribution parameters are

$$\theta = [\beta_1 \;\; q \;\; \sigma \;\; \lambda_L \;\; \lambda_R]^T = [5 \;\; 0.05 \;\; 20 \;\; 15 \;\; 0.05]^T. \qquad (11)$$

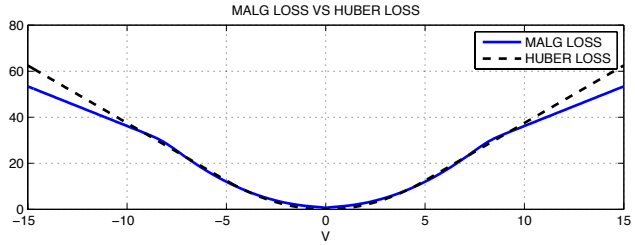

Fig. 1. Comparison of the MALG log likelihood and the Huber loss index.

Figure 1 plots the negative log likelihood index $L(v) = -\log p_{MALG}(v|1,\theta)$ for parameters (11). In what follows, we relate the optimal Bayesian estimation based on MALG distribution to robust regression based on Huber's loss index optimization. The second curve in Figure 1 is an example Huber loss index. Both functions in Figure 1 grow linearly for large absolute values of the argument. This provides robustness of the estimates to large outliers. Huber loss index is convex and quadratic for small arguments. MALG loss index is non-convex and non-quadratic for small arguments.

III. EM Algorithm

The objective is to find a Maximum Likelihood Estimate (MLE) of the parameter vector θ (9) in (8) from i.i.d. data (1). MLE is the solution of the optimal Bayesian problem

$$\theta = \arg\max \sum_{i=1}^{N} \log p_{MALG}(y_i|X_i,\theta). \qquad (12)$$

To solve MLE problem (12) we use Expectation Maximization (EM) algorithm, see [1]. EM for a GLM mixture is discussed in [5], [15]. The EM algorithm described below accounts for the linear models $\beta^T X$ in the mixture being the same. The EM formulation allows to solve (12) as series of convex optimization steps and finite expression computations. The EM algorithm iterates between two major steps, the expectation step (E step) and the minimization step (M step), that are detailed in the next two subsections.

A. E Step

The expectation step assumes that the parameter vector θ (9) is known. It computes posterior probabilities w_{iz} for data point i in the data set (1) belonging to either distributions (6) with $z=0$ or (7) with $z=1$. Standard EM algorithm, see [5], computes these posterior probabilities as

$$\begin{aligned} w_{ij} &= p(z_i = j|\theta, y_i), && (13)\\ &= \frac{p(y_i|X_i,\theta,z_i=j)p(z_i=j)}{\sum_{k=0}^{1} p(y_i|X_i,\theta,z_i=k)p(z_i=k)}, && (14) \end{aligned}$$

where $p(z_i=1)=q$, $p(z_i=0)=1-q$, and $p(y_i|X_i,\theta,z)$ are given by (6), (7). Weights w_{ij} are complementary,

$$w_{i0} + w_{i1} = 1, \quad (i=1,...,N). \qquad (15)$$

Large weight w_{i0} means data point i is likely generated by the normal distribution and belongs to MALG body. For large w_{i1} and small w_{i0}, the data point is likely generated by the AL distribution and belongs to the distribution tail.

B. M Step

M step assumes that posterior probabilities w_{ij} are known and estimates parameters θ of the MALG distribution. This is done by maximizing a convex lower bound $J(\theta)$ of the log likelihood function in (12), $L(\theta) = \sum_i \log p_{MALG}(y_i|X_i,\theta)$

$$J(\theta) = \sum_{i=1}^{N}\sum_{k=0}^{1} w_{ik} \log p(y_i|X_i,\theta,z=k), \qquad (16)$$

$$\begin{aligned} \log p(y_i|X_i,\theta,z=0) =& -(1/2)\sigma^{-2}(y_i-\beta^T X_i)^2 \\ &+ \log(1-q) - \log\sigma^2, && (17)\\ \log p(y_i|X_i,\theta,z=1) =& \lambda_L (v_i)_- - \lambda_R (v_i)_+ \\ &+ \log q - \log(\lambda_L^{-1} + \lambda_R^{-1}), && (18) \end{aligned}$$

where $(x)_+ = \max\{x, 0\} \geq 0$ and $(x)_- = \min\{0, x\} \leq 0$.

Maximizing (16) instead of (12) is a standard M step, e.g., see [1]. Expressions (17), (18) are specific to the MALG distribution. Our version of the EM approach performs M step as a sequence of three sub-steps: Robust Regression, Body Estimation, and Tail Estimation. These compute partial optimums over subsets of parameters in vector θ (9) and are detailed below. The three sub-steps make a block coordinate descent algorithm maximizing the objective $J(\theta)$.

1) Robust Regression: The robust regression step estimates regression parameter vector β assuming that other MALG parameters in θ (9) are known. Maximizing (16) with respect to β, assuming that σ, λ_L, and λ_R in (17), (18) are known, leads to the following convex optimization problem

$$\hat{\beta} = \arg\min \sum_{i=1}^{N} \Big[\frac{w_{i0}}{2\sigma^2}(y_i - \beta^T X_i)^2 - w_{i1}\lambda_L(y_i - \beta^T X_i)_- + w_{i1}\lambda_R(y_i - \beta^T X_i)_+ \Big], \quad (19)$$

where the decision vector β has a moderate size, the same as the number of regressors. The quadratic programming (QP) problem (19) can be numerically solved using one of the standard QP solvers.

2) Body Estimation: MALG distribution body is described by the second summand in (10) that is defined by parameters q and σ. To find these, we maximize (16) with respect to q and σ^2 assuming that β, λ_L, and λ_R in (17), (18) are known. This results in two independent problems for q and σ^2; each has a closed form analytical solution. The optimal estimate of q is

$$\hat{q} \quad = \quad \frac{1}{N}\sum_{i=1}^{N} w_{i1}. \quad (20)$$

The estimate of the normal distribution covariance σ^2 is

$$\hat{\sigma}^2 \quad = \quad \left(\sum_{k=1}^{N} w_{k0}\right)^{-1} \sum_{i=1}^{N} w_{i0}(y_i - \beta^T X_i)^2. \quad (21)$$

3) Tail Estimation: MALG distribution tails are described by the third summand in (10). The estimated parameters are λ_L and λ_R. We assume that q is already known. Maximizing (16) with respect to λ_L, and λ_R for given β, σ, and q in (17), (18), yields the following problem

$$\{\hat{\lambda}_L, \hat{\lambda}_R\} = \arg \max_{\lambda_L, \lambda_R} \sum_{i=1}^{N} w_{i1} \big[\lambda_L(y_i - \beta^T X_i)_- - \lambda_R(y_i - \beta^T X_i)_+ - \log(\lambda_L^{-1} + \lambda_R^{-1})\big]. \quad (22)$$

Differentiating yields a non-linear system that allows the following closed form solution for the optimum (note that a_L and a_R are always non-negative)

$$\hat{\lambda}_L = (a_L + \sqrt{a_L a_R})^{-1}, \quad \hat{\lambda}_R = (a_R + \sqrt{a_L a_R})^{-1}, \quad (23)$$

$$a_L = \Big(\sum_{i=1}^{N} w_{i1}\Big)^{-1} \sum_{i=1}^{N} -w_{i1}(y_i - \beta^T X_i)_-, \quad a_R = \Big(\sum_{i=1}^{N} w_{i1}\Big)^{-1} \sum_{i=1}^{N} w_{i1}(y_i - \beta^T X_i)_+. \quad (24)$$

C. Algorithm

The proposed expected conditional maximization (ECM) version of the EM algorithm starts with an initial guess θ_0 of the parameter vector. One approach to computing θ_0 is discussed below in Section IV-B. The algorithm logic is presented in Algorithm 1 panel. In examples of Sections IV and V, it takes about 15 EM algorithms iterations to achieve convergence.

Data: Data set $\mathcal{D}$ (1), parameter guess θ_0, accuracy ϵ
Result: $\hat{\theta}$
Initialize: $\theta = \theta_0$
while $\|\Delta\theta\| > \epsilon$ **do**
  E step: Compute weights w_{ij} from (14)
  M step:
    Robust Regression: get $\hat{\beta}$ from QP problem (19),
    Body Estimation: compute $\hat{q}$ (20) and $\hat{\sigma}$ (21),
    Tail Estimation: compute $\hat{\lambda}_L$, $\hat{\lambda}_R$ (23),
    $\theta = [\hat{\beta}\ \hat{\sigma}\ \hat{\lambda}_L\ \hat{\lambda}_R\ \hat{q}]^T, \quad \Delta\theta = \hat{\theta} - \theta, \quad \hat{\theta} \leftarrow \theta$
end

Algorithm 1: EM algorithm for identification of MALG

IV. EM Algorithm Discussion

The EM algorithm with the block coordinate descent in M-step described in Subsection III-B is known to converge. The variant on the EM with the M step broken into partial minimization steps is called the Expectation Conditional Minimization (ECM) algorithm. It was earlier considered by Meng and Rubin [23], who discuss the ECM convergence.

A. Idealized Models

Weights w_{ij} (14) provide 'soft' switching between the distribution body and tail, the normal and AL distributions in the mixture model. Additional insight into the formulated algorithm can be obtained by considering a 'hard threshold' modification of the EM algorithm. Consider an idealized model with weights w_{ij} are replaced by

$$w_{ij}^* = \text{nint}(w_{ij}), \quad (25)$$

where $\text{nint}(\cdot)$ rounds its argument to the nearest integer.

The idealized model helps to establish the initial quess of the parameter vector θ for the EM update. Consider the tails of the MALG distribution first, where $w_{i0} \leq w_{i1}$. Expressions (14) for w_{i0} and w_{i1} have common denominator. In accordance with (6), (7), for the tails

$$qp_{AL}(y_i - \beta^T X_i, \lambda_L, \lambda_R) \geq (1-q)p_N(y_i - \beta^T X_i, \sigma^2). \quad (26)$$

This paper assumes the tail intensity q is small. For $q < 1/2$, the left tail consists of data points, where $y_i - \beta^T X_i < t_L$; for the right tail $y_i - \beta^T X_i > t_R$. From (26), we get

$$\begin{aligned} t_L &= -\sigma^2\lambda_L - \sqrt{(\sigma^2\lambda_L)^2 + 2\sigma^2 \log\rho}, \\ t_R &= \sigma^2\lambda_R + \sqrt{(\sigma^2\lambda_R)^2 + 2\sigma^2 \log\rho}, \\ \rho &= (1-q)\left(\lambda_L^{-1} + \lambda_R^{-1}\right) / \left(\sqrt{2\pi}q\sigma\right). \end{aligned} \quad (27)$$

For the tail points, $w_{i0} \leq w_{i1}$ and (15) imply that $w_{i0} \leq 1/2$ and $w_{i1} \geq 1/2$. The idealized model is then

$$w^*_{i0} = 0, w^*_{i1} = 1, \quad \text{for} \quad i \in \mathcal{S}_L \cup \mathcal{S}_R, \tag{28}$$
$$w^*_{i0} = 1, w^*_{i1} = 0, \quad \text{for} \quad i \in \mathcal{S}_B, \tag{29}$$

where (28) is the idealized model for the tail indexes and (29) for the distribution body indexes. The left tail index set $\mathcal{S}_L$, the right tail set $\mathcal{S}_R$, and the body indexes $\mathcal{S}_B$ are

$$\mathcal{S}_L \equiv \{i : y_i - \beta^T X_i \leq t_L\}, \tag{30}$$
$$\mathcal{S}_R \equiv \{i : y_i - \beta^T X_i \geq t_R\}, \tag{31}$$
$$\mathcal{S}_B \equiv \{i : t_L < y_i - \beta^T X_i < t_R\}. \tag{32}$$

Figure 2 compares weights w_{ij} (14) and the idealized model (28), (29) for the example (11). The plot argument is $v = y - \beta^T X$. In Figure 2, $t_L = -0.187$ and $t_R = 0.165$.

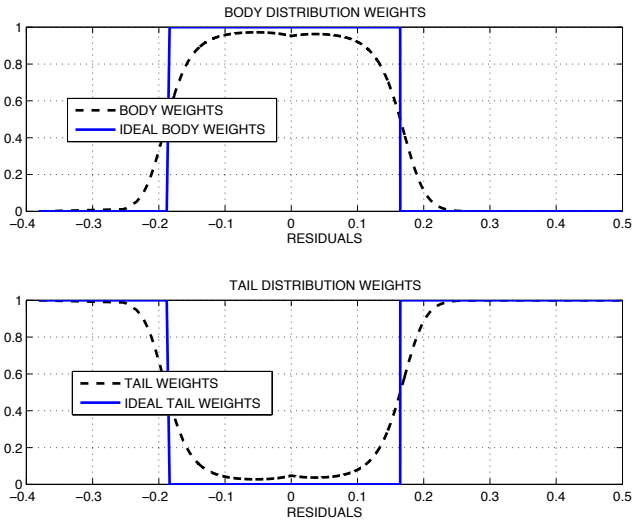

Fig. 2. Weights w_{i0} for the distribution body (upper plot) and w_{i1} for the tail (lower plot)

B. Initialization and Approximate Algorithms

1) Initialization: Our goal is to establish the initial parameter vector guess θ_0 in EM Algorithm 1. This can be done by executing the M step of Algorithm 1, with idealized model (28), (29) used in place of the E step weights (14). The idealized model is defined through (30), (31), (32) that depend on the two tail threshold parameters t_L and t_R in (27) and the regression parameter vector β.

The initialization starts from assuming a Gaussian model. This is the same as having $q = 0$ in MALG. The solution is then the standard least square regression

$$\begin{aligned} \beta_0 &= \arg\min \textstyle\sum_{i=1}^N (y_i - \beta^T X_i)^2, \\ \sigma_0^2 &= N^{-1} \textstyle\sum_{i=1}^N (y_i - \beta_0^T X_i)^2. \end{aligned} \tag{33}$$

These initial estimates can be further improved by removing the outlier data points, where $y_i - \beta_0^T X_i$ is outside the $[t_L, t_R]$ interval. Several iterative updates usually provide convergence, see [4] for discussion. The tail thresholds define the initial idealized model and are initialized as

$$t_{L,0} = -3.5\sigma_0, \quad t_{R,0} = 3.5\sigma_0. \tag{34}$$

2) Approximate Tail Estimation: By substituting idealized model (28)–(32) in place of weights w_{ij} in (24) we get

$$\begin{aligned} a_{L,0} &= -N_T^{-1} \textstyle\sum_{i \in \mathcal{S}_L} (y_i - \beta_0^T X_i), \\ a_{R,0} &= N_T^{-1} \textstyle\sum_{i \in \mathcal{S}_R} (y_i - \beta_0^T X_i), \end{aligned} \tag{35}$$

where $N_T = \text{card } \mathcal{S}_L + \text{card } \mathcal{S}_R$ is the total number of the tail points. Initial estimates of MALG tail parameters $\lambda_{L,0}$ and $\lambda_{R,0}$ are then computed from (23) using approximate initial values (35) in place of a_L and a_R.

Estimates (35) are related to MLE estimation of exponential tail using peaks over threshold method. In the motivating example, response variable y is logarithm of the physical variable. In that case, estimates (35) are closely related to Hill's estimator for Pareto distribution, see [6].

3) Approximate Body Estimation: The initial estimate q_0 of q is deduced by approximating MALG cumulative distribution function (CDF) in the tails with AL CDF (50) as follows

$$P\left(y - \beta^T X \leq t_L | \theta\right) \approx q\lambda_R(\lambda_L + \lambda_R)^{-1} e^{\lambda_L t_L}, \tag{36}$$
$$P\left(y - \beta^T X \geq t_R | \theta\right) \approx q\lambda_L(\lambda_L + \lambda_R)^{-1} e^{-\lambda_R t_R}. \tag{37}$$

We approximate probabilities (36) and (37) with $\text{card}(\mathcal{S}_L)/N$ and $\text{card}(\mathcal{S}_R)/N$, to estimate q_0

$$q_0 = (q_L + q_R)/2, \tag{38}$$
$$q_L = N^{-1}\text{card}(\mathcal{S}_L)(1 + \lambda_{L,0}\lambda_{R,0}^{-1}) e^{-\lambda_{L,0} t_{L,0}}, \tag{39}$$
$$q_R = N^{-1}\text{card}(\mathcal{S}_R)(\lambda_{R,0}\lambda_{L,0}^{-1} + 1) e^{\lambda_{R,0} t_{R,0}}. \tag{40}$$

The initial guess q_0 can be iteratively refined by recomputing weights in (14), updating thresholds (27), and then computing a new value of q_0 using (20).

The idealized model yields covariance estimate (21) as

$$\sigma_0^2 = \frac{1}{N - N_T} \sum_{i \in \mathcal{S}_B} (y_i - \beta_0^T X_i)^2, \tag{41}$$

where $N - N_T = \text{card } \mathcal{S}_B$. This estimate is consistent with computing σ_0 after removing the outliers (tail points).

4) Approximate Robust Regression: For the idealized model weights (28)–(32), (34), QP problem (19) becomes the well-known robust regression problem formulated as Huber loss function optimization, see [16]. For $\lambda_L \neq \lambda_R$, the loss function is asymmetric. The described procedure for iterative outlier removal when estimating β from (33) provides robust regression approach that is roughly equivalent to the Huber robust regression. Additional discussion of the two robust regression approaches can be found in [4].

5) Initial Guess: Initial parameter vector θ_0 in Algorithm 1 is set to $\theta_0 = [\beta_0 \ \ \sigma_0 \ \ \lambda_{L,0} \ \ \lambda_{R,0} \ \ q_0]^T$.

C. Verification

The performance of the described EM algorithm for MLE of MALG distribution parameters was verified for simulated data with known ground truth. For that purpose, 40,000 points were generated following MALG distribution with parameters (11). In the verification example, $\beta^T X_i = \beta_1 \cdot 1$, with $X_i = 1$. The results obtained after five iterations of Algorithm 1 are shown in Table I. The estimates obtained by the algorithm are close to the ground truth parameters.

TABLE I

TRUE AND ESTIMATED PARAMETERS FOR SIMULATED DATA

	μ	σ	λ_L	λ_R	q
Estimated Value	4.9997	0.0502	21.1615	14.4549	0.0446
True Value	5	0.05	20	15	0.05

V. POWER LOAD DEMAND MODEL

The motivating example for the MALG model is forecasting of electrical power demand for a utility. Anonymous US utility data described in [8] include hourly loads for 20 zones served by the utility and ambient temperature. The methodology described above was applied to the aggregate load across all these zones. The range of the aggregate load is 0.8 to 3.2GW, with the average value being 1.6GW. The data covers a time range of approximately 4 years with sampling interval of one hour, $N = 38,070$ samples at all.

Let $L(t)$ be the load demand. The data is sampled every hour and t is the number of hours elapsed since the start of the data collection. We use logarithmic load, normalized by $L_0 = 1$ GW, as response variable y_t

$$y_t = \log\left(L(t)/L_0\right). \tag{42}$$

We use linear regression model of the hourly load demand described in [22]. This model has the form $y_t = \beta^T X_t$ with 58 non-linear regressors X_t that depend on temperature, load values, and time. The regressors are calculated from the available data at each time sample to provide dataset (1).

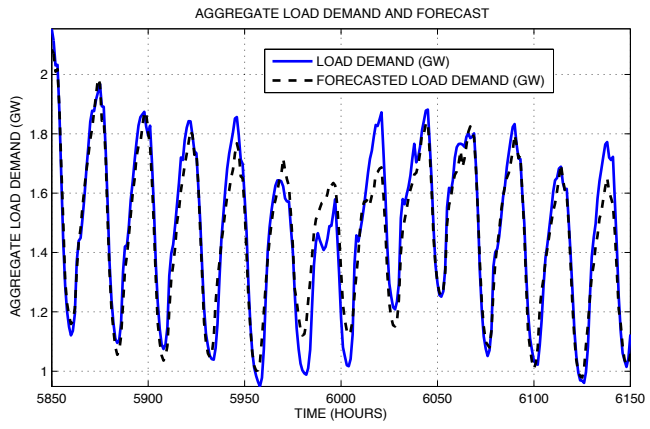

Fig. 3. Comparison of aggregated load demand and forecast.

Algorithm 1 was used to fit the MALG mixture of Generalized Linear Models to the data $\{X_t, y_t\}_{t=1}^N$. Model forecast was calculated as $\hat{y}_t = \hat{\beta}^T X_t$, where $\hat{\beta}$ is the regression parameter vector estimated by the EM algorithm. Figure 3 compares the forecast $\hat{y}_t$ to the actual logarithmic load demand y_t for 301 hour period.

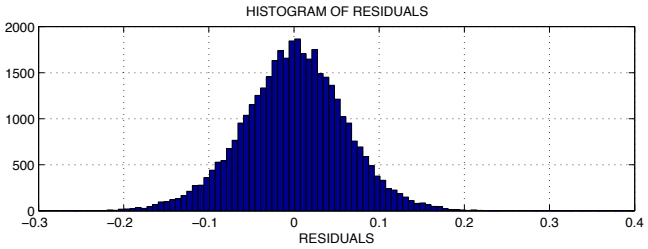

Fig. 4. Residual histogram

A histogram of the prediction residuals $y_t - \hat{y}_t$ is shown in Figure 4. The long tails of the residuals are not modeled by a normal distribution. A zero-mean MALG distribution model provides much better fit. The estimated MALG model parameters are summarized in Table 3. Vector $\hat{\beta} \in \Re^{58}$ is not included here because of the space limitations. The shown parameters are roughly the same as in the example (11).

TABLE II

ESTIMATED MALG PARAMETER FOR THE LOG LOAD MODEL

Distribution	Normal $\hat{\sigma}$	AL $\hat{\lambda}_L$ and $\hat{\lambda}_L$	Bernoulli $\hat{q}$
	0.0593	(18.2594,19.3068)	0.0659

For comparison, a regression model based on the normal distribution was fitted to the data. This model is a special case of the described MALG model with mixture parameter $q = 0$. Zooming in on the right tail, Figure 5 shows a comparison between the PDFs of the fitted normal and MALG distribution. One can see that the MALG distribution provides much better for the tail than the normal distribution.

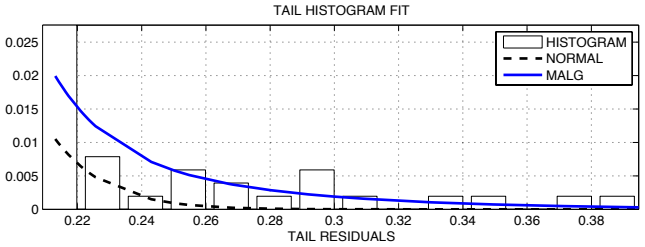

Fig. 5. Histogram of tail data

VI. STOCHASTIC OPTIMIZATION

This section considers a practically important problem that can be solved using the model described in Section V. Utilities order electrical power from power generating companies 24 hours in advance. The demand that exceeds the advance order has to be fulfilled in the spot market, at much higher price. There is a trade-off between overpaying upfront at day-ahead power cost and the risk of paying for excessive

demand at much higher spot price. The MALG model allows stochastic optimization of this trade-off taking into account the long tail risk of unusually large demand.

Models of day-ahead electricity market order were considered in earlier work. Related approach in [7] considers much simpler forecasting model for power load and no optimization. Stochastic optimization in [11] uses a simple regression model and normal model without the long tails.

A. Expected Total Cost Formulation

This section assumes that accurate day ahead temperature forecast is available. In that case the regressors vector X_t in load demand model of Section V is available day ahead. The problem below assumes that regressor vector X is deterministic and known. The next day demand y is a random variable with distribution $y \sim \text{MALG}(X, \theta)$; parameter θ can be identified from historical data as described in Section V.

The goal is to model and optimize the total expected cost that will be paid for the electricity in the day's time. The forecasted mean log-load (42) is $\beta^T X$. The forecasted physical load (in GW) is $L_0 \exp(\beta^T X)$. The day ahead order is then $M + L_0 \exp(\beta^T X)$, where M is the margin above the mean forecast. The total cost can be broken up into deterministic day-ahead upfront cost $A(M)$ and stochastic spot cost $C(M)$. The expected total cost is

$$U(M) = A(M) + \mathbf{E}[C(M)], \tag{43}$$

where $\mathbf{E}[\cdot]$ is the random variable expectation. The goal is to optimize margin M, such that $U(M)$ (43) is minimal.

The day-ahead cost $A(M)$ and the spot cost $C(M)$ are

$$A(M) = \pi_{adv} \cdot \left(M + L_0 e^{\beta^T X} \right), \tag{44}$$

$$C(M) = \pi_{spot} \cdot (D)_+, \tag{45}$$

$$D = L_0 e^{y} - M - L_0 e^{\beta^T X}, \tag{46}$$

where π_{adv} is electricity price on the day ahead market; it is considered as a deterministic constant. The load forecast error D is a stochastic variable depending on y, where $y \sim MALG(X, \theta)$. The spot price π_{spot} is also a stochastic variable. Spot cost $C(M)$ (45) includes the term $(D)_+$. If demand error D (46) is positive, the electricity must be procured at the spot price π_{spot}. If the error D is negative, the advance cost $A(M)$ is not recovered. The trade off is then between wasting money on buying too much power in advance or potentially paying a high spot cost later on if demand is not met. The goal is to find M such that the expected cost $U(M)$ (43) is minimized.

It is well known that the spot price is driven by the spot demand. We assume that spot price π_{spot} is affine in the demand (day-ahead forecast error D). Related models are discussed in [7], [11], [12]. Our model is

$$\pi_{spot} = aD + b + \epsilon, \tag{47}$$

where a, b are constants, ϵ is the modeling error for the prices, and D is given by (46). We consider ϵ as a stochastic variable independent of y and with zero mean $\mathbf{E}[\epsilon] = 0$. Related affine model for spot prices is considered in [13].

Substituting (47) into (45) yields $C(M) = aD(D)_+ + b(D)_+$. By taking expectation we get

$$\mathbf{E}[C(M)] = a\mathbf{E}[D^2 | D \geq 0] + b\mathbf{E}[D | D \geq 0]. \tag{48}$$

The exact expression for $\mathbf{E}[C(M)]$ is given in the Appendix.

B. Results

The stochastic optimization was applied to the power load data and the model described in Section V. The price models in Subsection VI-A were established based on the results of [19]. The advance price in (44) was set to π_{adv} =\$1920/MW-day. The affine spot cost model (47) was set to a = \$19,200-day/(MW-day)^2 and b = \$1,920,000/MW-day.

To illustrate the results, we pick sample t such that the forecasted mean demand $L_0 \exp(\beta^T X_t) = 2.845$ GW. Figure 6 plots MALG expected total cost $U(M)$ (43) and day ahead advance cost $A(M)$ (44) against the margin M argument. Expected cost $\mathbf{E}[C(M)]$ is the difference between the two curves. For small values of margin M the expected spot cost $\mathbf{E}[C(M)]$ dominates the total cost $U(M)$. This means not enough power was ordered in advance and there is high risk of covering substantial amount of power deficit at the high spot price. For large values of margin M, the expected spot cost becomes negligible and the advanced cost $A(M)$ dominates the total cost. This means too much power is bought in advance. The optimum is achieved somewhere in the middle, for $M = 467.6$MW.

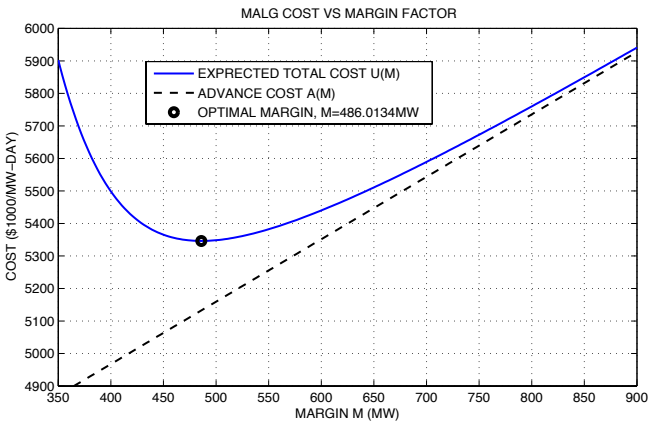

Fig. 6. Expected total cost.

We compared the stochastic optimization results obtained using the MALG model and a simpler model based on normal distribution. The latter can be considered a special case of the MALG model with $q = 0$. Table III is the summary of total expected costs based on the MALG and normal distribution model. The total expected cost shown for the normal distribution is computed assuming the optimal margin M is computed using the fitted normal model; the total cost computation then uses the same MALG model. In accordance with Table III, using MALG model instead of the simpler normal model in advance power order decision would save \$23,567 per day, that is about \$8.6 M per year.

TABLE III
OPTIMAL MARGIN AND COST

	Normal Model	MALG Model
Margin	432.59 MW	486.01 MW
Total Expected Cost	\$5,393,384/day	\$5,346,341/day

VII. CONCLUSION

This paper proposed a mixture model, MALG, that allows combined modeling of distribution body and distribution tails in regression problems. The paper demonstrated EM algorithm for optimal estimation of MALG parameters from data as sequence of robust regression, distribution body estimation, and tail estimation steps.

In the example application, the MALG model was used as a basis for stochastic optimization of the total expected cost for day-ahead electricity market. Modeling long-tail risks of the peak load demand events using MALG is shown to save a few million dollars per year compared to use of a normal distribution model and least squares regression.

REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1-38, 1977.

[2] C. Harris, *Electricity Markets: Pricing, Structures and Economics*. England: Wiley, 1st ed., 2006.

[3] C. Sigauke, A. Verster, and D. Chikobvu, "Extreme daily increase in peak electricity demand: Tail-quantile estimation," *Energy Policy*, vol. 53, pp. 90–96, 2013.

[4] D. J. Cummins and C. W. Andrews, "Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation," *Journal of Chemometrics*, vol. 9, pp. 489–507, 1995.

[5] G. McLachlan and D. Peel, *Finite Mixture Models*. Canada: Wiley-Interscience, October 2000.

[6] J. Beirlant, P. Vynckier, and J. L. Teugels, "Tail index estimation, Pareto quantile plots, and regression diagnostics," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1659–1667, December 1996.

[7] J. H. Kim, "Quantile optimization in the presence of heavy-tailed stochastic processes, and an application to electricity markets," PhD Dissertation, Princeton University, November 2011.

[8] Kaggle.com, "Global energy forecasting competition 2012 - Load forecasting," Available: https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting.

[9] L. A. Hannah, D. M. Blei, and W. B. Powell, "Dirichlet process mixtures of generalized linear models," *Journal of Machine Learning Research*, vol. 12, pp. 1923–1953, 2011.

[10] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction*. New York: Springer, 2006.

[11] L. Jia and L. Tong, "Optimal pricing for residential demand response: A stochastic optimization approach," *Communication, Control, and Computing*, pp. 1879–1884, October 2012.

[12] L. Jinying and L. Jinchao, "Next-day electricity price forecasting based on support vector machines and data mining technology," *27th Chinese Control Conference*, pp. 630–633, July 2008.

[13] M. A. Crew, C. S. Fernando, and P. R. Kleindorfer, "The theory of peak-load pricing: A survey," *Journal of Regulatory Economics*, vol. 8, pp. 215–248, 1995.

[14] M. Haas, S. Mitnik, and M. S. Paolella, "Modeling and predicting market risk with Laplace-Gaussian mixure distributions," *Applied Financial Economics*, vol. 16, no. 15, pp. 1145-1162, 2006.

[15] M. Wedel and W. S. DeSarbo, "A mixture likelihood approach for generalized linear models," *Journal of Classification*, vol. 12, pp. 21–55, 1995.

[16] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. New Jersey: Wiley, 2nd ed., 2009.

[17] P. N. Jones and G. J. McLachlan, "Laplace-normal mixtures fitted to wind shear data," *Journal of Applied Statistics*, vol. 17, no. 2, pp. 271–276, 1990.

[18] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. England: Wiley, 1st ed., 2006.

[19] S. Borenstein, M. Jaske, and A. Rosenfeld, *Dynamic Pricing, Advanced Metering and Demand Response in Electricity Markets*, Hewlett Foundation, San Francisco, CA, October 2002.

[20] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 896–904, September 2005.

[21] S. Mirasgedis et al., "Models for mid-term electricity demand forecasting incorporating weather influences," *Energy*, vol. 31, pp. 208–227, 2006.

[22] S. Shenoy and D. Gorinevsky, "Risk adjusted forecasting of electric power load," *American Control Conference (ACC)*, pp. 914–919, Portland, OR, June 2014.

[23] X. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

APPENDIX

The asymmetric Laplace distribution $AL(\mu, \lambda_L, \lambda_R)$ has the following PDF

$$p_{AL}(x) = \begin{cases} \kappa\, e^{\lambda_L(x-\mu)}, & x < \mu \\ \kappa\, e^{-\lambda_R(x-\mu)}, & x \geq \mu \end{cases}. \tag{49}$$

The CDF of AL is given by ($\kappa = 1/(\lambda_L^{-1} + \lambda_R^{-1})$)

$$F_{AL}(x) = \begin{cases} \lambda_L^{-1}\kappa\, e^{\lambda_L(x-\mu)}, & x < \mu \\ 1 - \lambda_R^{-1}\kappa\, e^{-\lambda_R(x-\mu)}, & x \geq \mu \end{cases}. \tag{50}$$

Section VI-A requires computation of expected spot cost (48), $\mathbf{E}[C(M)] = a\mathbf{E}[D^2|D \geq 0] + b\mathbf{E}[D|D \geq 0]$. In the computation, we hold parameters $\sigma, \lambda_L, \lambda_R, q, \beta^T X, L_0$ constant do not show dependence on these parameters explicitly. We define $m = \inf\{y - \beta^T X | D(y) \geq 0\}$. By solving $\exp(m + \beta^T X) = \exp(\beta^T X) + ML_0^{-1}$ we get

$$m = \log\left(1 + ML_0^{-1}\exp(-\beta^T X)\right). \tag{51}$$

Next we compute the expected values $\mathbf{E}[D^2|D \geq 0]$ and $\mathbf{E}[D|D \geq 0]$ based on (10)

$$\mathbf{E}[D^2|D \geq 0] = L_0^2 \int_m^\infty \left(e^y - \gamma(M)\right)^2 p_{MALG}(y|X,\theta)dy \tag{52}$$

$$= L_0^2\left[\Psi_2(m) - 2\gamma(M)\Psi_1(m) + \gamma(M)^2\Psi_0(m)\right], \tag{53}$$

$$\mathbf{E}[D|D \geq 0] = L_0 \int_m^\infty \left(e^y - \gamma(M)\right) p_{MALG}(y|X,\theta)dy \tag{54}$$

$$= L_0\left[\Psi_1(m) - \gamma(M)\Psi_0(m)\right], \tag{55}$$

where $\gamma(M) = e^{\beta^T X} + ML_0^{-1}$ and $\Psi_z(m)$ is defined as

$$\Psi_z(m) = \int_m^\infty e^{zy} p_{MALG}(y|X,\theta)dy \tag{56}$$

$$= e^{z\beta^T X}\left[(1-q)\,\eta_z(m) + q\,\kappa\,\phi_z(m)\right] \tag{57}$$

$$\eta_z(m) = \frac{1}{2}e^{z^2\sigma^2/2}\mathrm{erfc}\left((m\sigma^{-1} - z\sigma)/\sqrt{2}\right) \tag{58}$$

$$\phi_z(m) = \begin{cases} \kappa_R\, e^{-\kappa_R^{-1}m}, & m \geq 0 \\ \kappa_R + \kappa_L\left(1 - e^{\kappa_L^{-1}m}\right), & m < 0 \end{cases} \tag{59}$$

where $\kappa_L = (\lambda_L + w)^{-1}$ and $\kappa_R = (\lambda_R - w)^{-1}$. These integrals exist provided that $\lambda_R > 2$.