Non-Parametric Regression Modeling for Stochastic Optimization of Power Grid Load Forecast

Saahil Shenoy[1], Dimitry Gorinevsky[2], and Stephen Boyd[2]

***Abstract*— This paper develops a method for building non-parametric stochastic models of multivariate distributions from large data sets. The motivation is stochastic optimization based on time series forecasting models. The proposed non-parametric stochastic modeling approach is based on multiple quantile regressions with inter-quantile smoothing. The models are built using ADMM optimization approach scalable to large datasets. As an application example, the paper considers forecasting of the loads in the electrical power grid. The forecasted load is used for the electricity procurement in the day-ahead power market. The stochastic optimization trades the costs of advance and spot procurements of the electricity. This problem is currently important because the random variability in the grid power load increases with integration of renewable generation.**

I. INTRODUCTION

This paper studies non-parametric multivariate stochastic models that can be built from large data sets and used for stochastic optimization of decisions. This work is motivated by electrical power market applications where time series forecasts of the load are used for day-ahead procurement.

A flexible non-parametric multivariate approach to modeling the empirical distribution function is offered by quantile regression, see [1]. In this approach, the domain of the inverse empirical distribution function at given value is described by a linear combination of the regressors. Quantile regression is used in many applications.

Quantile regression can be computed as a solution to a LP (Linear Programming) problem, see [1]. For a given quantile, it can be considered as a generalization of Generalized Linear Model with 'link function' defined by asymmetric Laplacian distribution. The solution is robust to the data outliers.

The quantile regression function is included with major statistical software packages. However, so far, it has found limited use for building non-parametric multivariate models that can be used in stochastic calculus.

Most of the related quantile regression work is for kernel-type models of multivariate non-linear distributions, e.g., see [2], [3], [4], [5], [6]. The kernel models use as many regressors as there are data points. One known issue with quantile regression is *quantile crossing*. The problem is that for some model arguments different quantile regression solutions might be improperly ordered. For kernel models, this problem can be addressed by introducing the ordering constraint at each data point and solving the optimization problems for all quantiles simultaneously, e.g., see [3], [4]. An alternative is adding a regularization penalty term that encourages the smooth transition of the model between different quantiles, e.g., see [5].

The literature on multivariable non-parametric quantile regression models, such as the forecasting problem in this paper, is limited. Estimation of multi-quantile model with smoothing is discussed in [7]. Some of the theoretical issues related to smoothed quantile regression and reordering for the quantile crossing are discussed in [8], [9], [10]. These papers do not address computational scalability of the model estimation and modeling of the distribution tails. High and low quantiles do not have enough data points to fit multivariate models. A viable approach is to fix the regression slope and use a parametric model of the tail, see [11].

Non-parametric multivariate quantile regression modeling of the entire distribution for stochastic optimization appears to be new. Earlier, quantile regression was applied to off-line analysis of the data, such as risk estimation in finance [12]. The quantile model in [13] is suitable for forecasting of wind power generation; yet this model separates each regressor variable and is not truly multivariate.

To address the quantile regression crossing, we introduce a notion of a solution ball where the quantiles are guaranteed to be ordered. This paper develops a characterization of the ball radius, which can be increased by enhancing the smoothing.

The multi-quantile regression formulation in this paper is a quadratic programming (QP) problem. We present a computational procedure for solving this QP problem using alternating direction method of multipliers (ADMM) method known as block splitting, see [14], [15]. The procedure is scalable to extremely large training data sets.

The contributions of this paper are as follows. (i) A multi-quantile regression formulation for a non-parametric model suitable for use in stochastic optimization. The smoothed formulation enables numerical differentiation of the quantiles to obtain the probability density. (ii) A scalable computational method for optimization solution of the proposed formulation. (iii) A constructive approach to addressing the issues of quantile regression crossing and tail modeling. (iv) An application to stochastic optimization of power load forecasting in the day-ahead power market. We train the non-parametric model using historical data and demonstrate its use for on-line stochastic optimization.

II. SINGLE QUANTILE REGRESSION

Consider a dataset

$$D = \{Z_i, y_i\}_{i=1}^N, \tag{1}$$

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

[2]Dimitry Gorinevsky and Stephen Boyd are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA {gorin, boyd}@stanford.edu

where scalars y_i are response variables, vectors $Z_i \in \Re^n$ are explanatory variables (regressors), i is the sample number, and N is the number of the samples, which can be large. In what follows, we assume that data (1) are i.i.d., and follow unknown conditional multivariate distribution $p(y_i|Z_i)$. In forecasting applications, i is the time sample and the i.i.d. assumption means time-invariance of the underlying process.

A. Quantile Regression Problem

We assume that the generating distribution $p(y_i|Z_i)$ for (1) is described by the model

$$\mathbf{P}\left(y_i \leq y|Z_i\right) = q, \quad y(q) = Z_i\beta(q) + \alpha(q), \tag{2}$$

where $q \in (0,1)$ is the quantile level; $\beta \in \Re^n$ and $\alpha \in \Re$ are the quantile regression hyperplane parameters. For a given q, model (2) is a solution to the LP problem, see [1],

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \left[q\left(y_i - Z_i\beta - \alpha\right)_+ + (q-1)\left(y_i - Z_i\beta - \alpha\right)_-\right],$$

where $(x)_+ = \max\{x, 0\}$ and $(x)_- = \min\{x, 0\}$. This problem can be compactly written as

$$\begin{aligned} &\underset{\alpha,\beta}{\text{minimize}} \;\; h(y - Z\beta - \alpha\mathbf{1}_N; q), \\ &h(x;q) = \frac{1}{2}||x||_1 + \left(q - \frac{1}{2}\right)\mathbf{1}_N^T x, \end{aligned} \tag{3}$$

where matrix $Z = [Z_1 \dots Z_N] \in \Re^{N\times n}$ and $\mathbf{1}_N \in \Re^N$ is a column vector of ones. For $q = 1/2$, the quantile regression is the median regression. The regression hyperplanes for different quantiles might be not parallel to each other.

Practical use of the quantile regression model has to deal with two issues. The first issue is the large variance of the solution (3) when there are few data points on one side of the hyperplane. The left ($q \ll 1$) are right ($1 - q \ll 1$) tail quantiles have few to none points, which is a problem.

The second issue is that for some data the quantiles might be unordered such that

$$Z_i\beta(q_1) + \alpha(q_1) > Z_i\beta(q_2) + \alpha(q_2)$$

for $q_1 < q_2$, where $\{\alpha(q), \beta(q)\}$ solves (3) for quantile level q. This is the quantile crossing problem mentioned in the introduction. In fact, if the hyperplanes are not parallel, $\beta(q_1) \neq \beta(q_2)$, they cross. This means one can always find test data such that the quantiles will be unordered.

Both issues are addressed in Section IV.

B. ADMM Block Formulation

The ADMM formulation is a scalable way to solve LP problems. The ADMM algorithm is formulated for the problem in the following general graph form, see [15],

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \sum_{i=1}^{N} f(z_i) + g(w), \\ \text{subject to} \quad & z = Aw, \end{aligned} \tag{4}$$

where $w \in \Re^K$, $z \in \Re^J$, and $A \in \Re^{J\times K}$, and $f(z)$, $g(w)$ are two convex closed functions.

The ADMM update at iteration $k+1$ is, see [15],

$$\begin{aligned} w^{(k+1/2)} &= \mathbf{prox}_g\left(w^{(k)} - \tilde{w}^{(k)}\right), \\ z^{(k+1/2)} &= \mathbf{prox}_f\left(z^{(k)} - \tilde{z}^{(k)}\right), \\ \left(w^{(k+1)}, z^{(k+1)}\right) &= \Pi_A\left(w^{(k+\frac{1}{2})} + \tilde{w}^{(k)}, z^{(k+\frac{1}{2})} + \tilde{z}^{(k)}\right), \\ \tilde{w}^{(k+1)} &= \tilde{w}^{(k)} + w^{(k+1/2)} - w^{(k+1)}, \\ \tilde{z}^{(k+1)} &= \tilde{z}^{(k)} + z^{(k+1/2)} - z^{(k+1)}, \end{aligned} \tag{5}$$

where $\mathbf{prox}_{f,g}$ are the proximal operators and Π_A denotes projection onto $\{(w,z) \in \Re^{J+K} | z = Aw\}$, see [16].

For quantile regression (3), we have $J = N$, $K = n+1$,

$$f(z_i) = \frac{1}{2}\left|y_i - z_i\right| + \left(\frac{1}{2} - q\right)z_i, \; (i = 1, \dots, N), \tag{6}$$

$$g(w) = 0, \tag{7}$$

$$A = \left[\mathbf{1}_N^T \;\; Z^T\right]^T, \tag{8}$$

$$w = \left[\alpha \;\; \beta^T\right]^T. \tag{9}$$

For functions f (6) and g (7), the operators in (5) are

$$\begin{aligned} \left(\mathbf{prox}_f(z)\right)_i =& y_i + \left(z_i - y_i - \frac{1-q}{\rho}\right)_+ \\ &- \left(y_i - z_i - \frac{q}{\rho}\right)_+, \end{aligned} \tag{10}$$

$$\begin{aligned} \left(\mathbf{prox}_g(w)\right)_i =& w_i, \\ \Pi_A(w,z) =& \left(\left(I + A^TA\right)^{-1}\left(w + A^Tz\right), \; Aw\right), \end{aligned} \tag{11}$$

where ρ is a scalar penalty parameter, see [14].

The ADMM algorithm in general is extremely scalable and parallelizable. This is covered in depth in [15]. We are specifically interested in scalability for large data set size N in (3). Updates (10) scale by separating components $i = 1, \dots, N$. In (11), matrix $(A^TA)^{-1}$ has small size $n \times n$. It can be right-multiplied by A one column at a time. This is very scalable, as is the multiplication Aw. Finally, the scatter matrix A^TA can be computed as the running sum

$$A^TA = \sum_{i=1}^{N} [1 \;\; Z_i]^T \cdot [1 \;\; Z_i].$$

III. Multi-quantile Regression

We need a model (2), where $\alpha = \alpha(q)$ and $\beta = \beta(q)$ are smooth functions that can be differentiated to compute the probability density. Solving single quantile regression problems (3) on a grid of q might not give the desired result. To get a better solution, we solve multiple quantile regression problems jointly, with a smoothing penalty. The optimization

problem on grid of n_q quantiles q_j is formulated as

$$\begin{aligned}\underset{\{\alpha_i,\beta_i\}_{i=1}^{n_q}}{\text{minimize}} \quad & \sum_{j=1}^{n_q} h(y - Z\beta_j - \alpha_j \mathbf{1}_N; q_j) \\ & + \lambda \sum_{j=2}^{n_q} ||\beta(q_j) - \beta(q_{j-1})||_2^2 \\ & + \mu \sum_{j=2}^{n_q-1} \left(\alpha_{j+1} + \alpha_{j-1} - 2\alpha_j\right)^2, \\ \text{subject to} \quad & \beta_L = \beta_i, \quad (i = 1, \ldots, L), \\ & \beta_R = \beta_i, \quad (i = R, \ldots, n_q),\end{aligned} \tag{12}$$

where $h(x, q)$ has the same form as in (3); λ is a penalty on the first difference on β_j; μ, on the second difference of α_j. The constraints on β_j are introduced because for the low and the high quantiles there is not enough data on one side of the hyperplane to get accurate estimates of both regression slope β_j and its intercept α_j. For these quantiles, we keep β_j constant and just estimate α_j.

A. ADMM Formulation

We cast (12) into ADMM form (4). With the overload of the notation, for problem (12) we have $K = n_q(n + 1)$, $J = n_q N + (2n_q - 3)(n + 1)$,

$$\begin{aligned} f(z_i) &= \frac{a_i}{2}\left|Y_i - z_i\right| + \left(\frac{1}{2} - Q_i\right) a_i z_i + (1 - a_i) z_i^2, \\ g(w) &= \mathbf{I}(Cw = 0), \\ A &= \left[\left(I_{n_q} \otimes [\mathbf{1}_N \; Z]^T\right) \; F^T\right]^T. \end{aligned} \tag{13}$$

In (13), $\mathbf{I}(\cdot)$ is the indicator function, $\otimes$ is the Kronecker product, and

$$w = \left[\begin{array}{ccccc} \alpha_1 & \beta_1^T & \cdots & \alpha_{n_q} & \beta_{n_q}^T \end{array}\right]^T, \tag{14}$$

$$Y = \left[\mathbf{1}_{n_q}^T \otimes y^T \quad \mathbf{0}_{1,(2n_q-3)(n+1)}\right]^T, \tag{15}$$

$$\mathbf{q} = \left[q_1 \;\; q_2 \;\; \ldots \;\; q_{n_q}\right]^T, \tag{16}$$

$$Q = \left[\mathbf{q}^T \otimes \mathbf{1}_N^T \quad \mathbf{0}_{1,(2n_q-3)(n+1)}\right]^T, \tag{17}$$

$$F = \left[\begin{array}{cc} \sqrt{\lambda} D_{n_q,1}^T \otimes \mathcal{I}_R & \sqrt{\mu} D_{n_q,2}^T \otimes e_1 \end{array}\right]^T, \tag{18}$$

$$a_i = \left\{\begin{array}{ll} 1, & i = 1, \ldots, n_q N, \\ 0, & \text{otherwise} \end{array}\right., \tag{19}$$

$$C = \left[\begin{array}{cc} D_{L,1}^T \otimes \mathcal{I}_R & \mathbf{0}_{K_L, K-K_L} \\ \mathbf{0}_{K_R, K-K_R} & D_{R,1}^T \otimes \mathcal{I}_R \end{array}\right], \tag{20}$$

where $\mathcal{I}_R = I_{n+1} - e_1 e_1^T$ with $e_1^T = [1 \;\; 0 \;\; \ldots \;\; 0] \in \Re^{n+1}$ and I_m an identity matrix of size m, $K_L = L(n+1)$, $K_R = R(n + 1)$, $D_{m,1}^T \in \Re^{(m-1)\times m}$ is the first difference matrix, $D_{m,2} \in \Re^{(m-2)\times m}$ is the second difference matrix, $\mathbf{1}_m \in \Re^m$ is a vector of ones, and $\mathbf{0}_{m,p} \in \Re^{m\times p}$ is a matrix of zeros; L and R are the numbers of constrains in (12).

The ADMM update steps for (13) have the form (5), where the proximal operators are given by

$$\begin{aligned} \left(\mathbf{prox}_f(z)\right)_i &= Y_i - \left(Y_i - \frac{\rho z_i + a_i/2 - (1/2 - Q_i)a_i}{2(1 - a_i) + \rho}\right)_+ \\ &\quad + \left(\frac{\rho z_i - a_i/2 - (1/2 - Q_i)a_i}{2(1 - a_i) + \rho} - Y_i\right)_+, \\ \mathbf{prox}_g(w) &= \Pi_C(w), \end{aligned} \tag{21}$$

where Π_C is the Euclidean projection onto a convex set given by constraint matrix C, see [17]. Since we have expressed the problem in graph ADMM form, this problem is extremely scalable to large datasets with a large number of independent variables. For large number N of the data points, the same type of reasoning as in Section II-B applies.

IV. Model Predictive Power

Two issues with the predictive power of the smoothed multi-quantile model of Section III are brought up in Subsection II-A. The first issue is quantile crossing for large regressors, $\|Z_j\| \gg 1$. The second issue is with the distribution tail modeling for large absolute values of response variables y. This section examines these two issues in more depth.

A. Large Regressors

The multi-quantile model is obtained by solving (14). The solution is the set of the slopes β_j and intercepts α_j (12), defined on the quantile grid q_j (16), where $j = 1, \ldots, n_q$. In what follows, we assume this set describes the functions $\alpha(q)$ and $\beta(q)$ in (2). The probability density $p(y) = \frac{dq}{dy}$ can be obtained by differentiating $y = Z_i\beta + \alpha$. In practice, a secant method of differentiation will use β_j and α_j.

The quantile crossing is avoided if $\frac{dq}{dy} > 0$. This is equivalent to $\frac{dy}{dq} > 0$, which can be expressed as

$$Z_i \frac{d\beta(q)}{dq} + \frac{d\alpha(q)}{dq} > 0. \tag{22}$$

We will use an equivalent form of (22)

$$-Z_i M^{-1} \cdot \frac{d\left(M\beta(q)\right)}{dq} \cdot \left[\frac{d\alpha(q)}{dq}\right]^{-1} \leq 1, \tag{23}$$

where M is a preconditioner matrix. Consider the regressor scatter matrix $Z^T Z$, where $Z = [Z_1 \ldots Z_N]$, for data set (1). Preconditioner M can be selected such that the matrix $M^{-1}(Z^T Z)M^{-1}$ has condition number of 1.

Consider the following use case. Model (2) is estimated (trained) for the historical data set (1). It is then used as a basis of on-line stochastic optimization for new data points. The quantile hyperplanes intersect somewhere in the regressor space, unless $\beta(q)$ is constant. This means one can always find a new regressor Z_* such that model has quantile crossing. Below is a simple condition that for a given Z_* there is no quantile crossing.

Using Cauchy-Schwarz inequality, a sufficient condition for (23) to hold is

$$\|Z_* M^{-1}\|_2 \le \Omega, \tag{24}$$

$$\Omega = \max_q \left\| \frac{d(M\beta(q))}{dq} \cdot \left[\frac{d\alpha(q)}{dq}\right]^{-1} \right\|_2^{-1}, \tag{25}$$

where Ω is the radius of the scaled regressors ball where the model is guaranteed to have no crossing.

B. Large Response Variable

The described smoothed model interpolates the training data. The models for low or high quantiles can extrapolate beyond that range if the parametric form of the distribution tails is known. Extreme Value Theory (EVT) predicts that in many cases the distribution tails, which describe the extreme events, follow a Pareto (power law) distribution. The tails can be estimated using peaks over threshold (POT) method, where the tail model is fitted to the data exceeding a threshold, see [18]. The exceedance data are usually sparse and parametric fit procedures used, such as the Hill's estimator for Pareto tail model [19].

The application example in Section V and VI uses log coordinates. Thus, the Pareto distribution becomes an exponential distribution. Consider the first q_1 and the last q_{n_q} quantiles on the modeling grid (16) as the tail thresholds. The POT exceedances are

$$e_{L,j} = y_j - Z_j\beta_1 - \alpha_1, \quad j \in J_L, \tag{26}$$

$$e_{R,k} = y_k - Z_k\beta_{n_q} - \alpha_{n_q}, \quad k \in J_R, \tag{27}$$

where $J_L \equiv \{j : y_j < Z_j\beta_1 + \alpha_1\}$ and $J_R \equiv \{k : y_k > Z_k\beta_{n_q} + \alpha_{n_q}\}$.

We model the probability distributions of $e_{L,j}$ and $e_{R,k}$ as

$$-e_{L,j} \sim q_1 \cdot \text{Exp}(\theta_L), \quad e_{R,k} \sim q_{n_q} \cdot \text{Exp}(\theta_R). \tag{28}$$

The maximum likelihood estimates (MLE) of θ_L and θ_R in (28) are

$$\hat{\theta}_L^{-1} = -\text{mean}\{e_{L,j} | j \in J_L\}, \ \hat{\theta}_R^{-1} = \text{mean}\{e_{L,k} | k \in J_R\}. \tag{29}$$

V. Power Load Model

The motivating example for development of the proposed non-parametric approach is forecasting of electrical power demand for a utility. The hourly load and price data from an anonymous US utility are described in [20]. The smoothed quantile regression modeling methodology was applied to the total system load. The range of the loads is 11.54 to 33.22 GW, with the average value being 18.02 GW. The data time range is from January 2011 to June 2013. The sampling interval is one hour. There are $N = 21,696$ samples in all.

Let P_t be the load demand. The data is sampled every hour and index t is the number of hours elapsed since the start of the data collection. Logarithmic load, normalized by $P_0 = 1$ GW, was used as response variable y_t

$$y_t = \log\left(P_t/P_0\right). \tag{30}$$

The 45 non-linear regressors Z_t included the day-behind log load and the time related regressors from [21], such as the hour, the week day, and the calendar month.

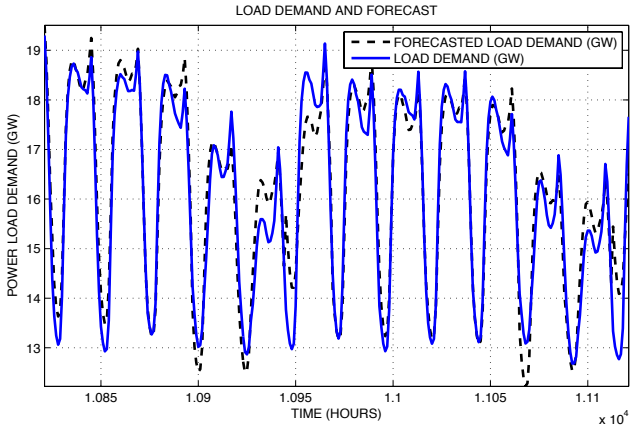

Fig. 1. Plot of median regression forecast.

Median regression, described by (3) with $q = 1/2$, is used to illustrate the quantile regression use for forecasting of the power load data in Figure 1. The forecast plotted is $y_t = Z_t\beta(1/2) + \alpha(1/2)$. One can see that the forecast matches the data reasonably well.

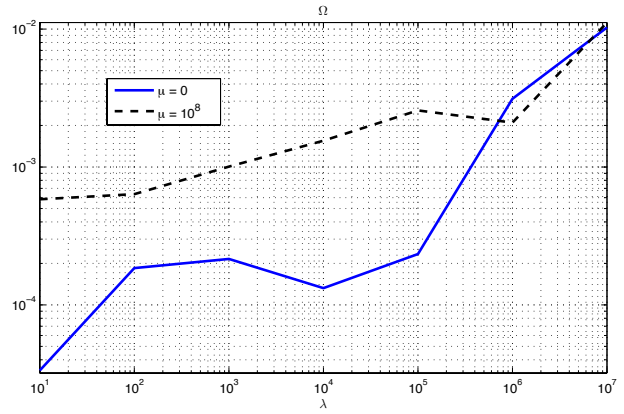

Fig. 2. Log-log plot of cluster radius Ω as a function of λ and μ.

To set up the smoothed multi-quantile regression model estimation problem (12), we experimented with the smoothing parameters μ and λ. For each combination of these parameters, we solved (12) with the available dataset using the ADMM method of Subsection III-A. We then numerically differentiated the obtained multi-quantile model data $\alpha(q)$ and $\beta(q)$ to compute the empirical estimate of the radius Ω in (25) plotted in Figure 2.

The problem parameters used in the example are summarized in Table I. These parameters yield $\Omega = 0.0021$ in the condition (25). Sufficient condition (24) for the absence of level crossing is then satisfied for 19,451 points in the data set, leaving approximately 2200 points out.

The PP (probablity-probablity) plot in Figure 3 illustrates the accuracy of the data fit for the developed model. The

abscissa is quantile level q in the fitted model (2), $\mathbf{P}(y_t < Z_t\beta(q) + \alpha(q)) = q$. The ordinate is the empirical quantile level estimated as the fraction of the data points in the set where the inequality $y_t < Z_t\beta(q) + \alpha(q)$ holds. For ideal model, the data points would be on the diagonal. As shown, the data points are close to the diagonal.

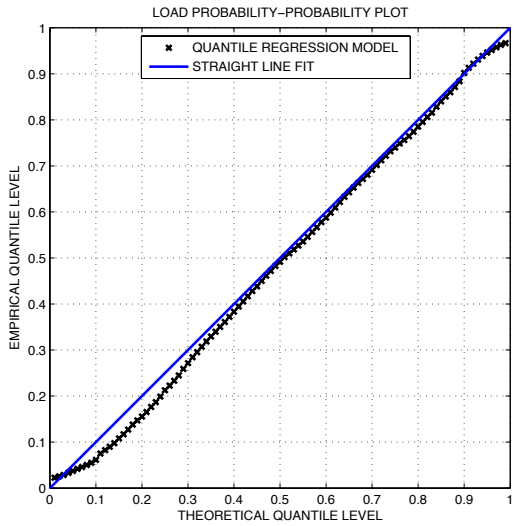

Fig. 3. PP plot for non-parametric distribution model fit for the load data.

VI. Stochastic Optimization of Cost

This section considers an example of using the non-parametric model described in Section V for stochastic optimization. The utilities order power in the electricity market a day in advance. If the actual power load is higher, the utility has to buy additional electricity at much higher spot price. If it is lower, then overpayment results. The goal of the stochastic optimization in this section is to minimize the total expected cost.

A. Advance and Expected Spot Cost

Consider a given time t, when regressor Z_t is known. The future (day-ahead) log-load y_t is defined by the conditional quantile model of the form (2) discussed in Section V. This section assumes that the advance and the spot prices of the electricity are known.

The dataset [20] used in Section V includes electricity spot prices. The spot prices range from \$12.52 to \$363.80, with average value being \$48.51.

The stochastic optimization requires the estimation of the advance cost and the expected spot cost components of the total expected cost. These costs depend on the advance order

$$P_a(t) = P_0 e^{y_t}, \tag{31}$$

TABLE I
Parameters of Multi-quantile Model for Load

n_q	q_1	q_{n_q}	q_L	q_R	λ	μ
99	0.01	0.99	0.1	0.9	10^6	10^8

where y_t is the logarithmic load. One can always find a quantile level s for Section V model such that $y(s) = y_t$ in (31).

$$y(s) = \begin{cases} y(q_1) + \log(s/q_1)/\theta_L, \ s < q_1 \\ Z_t\beta(s) + \alpha(s), \ q_1 \le s \le q_{n_q} \\ y(q_{n_q}) - \log((1-s)/q_{n_q})/\theta_R, \ s > q_{n_q} \end{cases}. \tag{32}$$

The advance cost is the deterministic value $\pi_{adv,t}P_a(t)$, where $\pi_{adv,t}$ is the advance price at time t.

$$A_t(s) = \pi_{adv,t} P_0 e^{y(s)}. \tag{33}$$

The spot cost is the random variable defined by the day-ahead future load. The expected spot cost can be computed using the model $p(y|Z_t)$ of the log-load y distribution conditional on the regressors Z_t. The utility has to pay the spot price only when the advance order $P_a(t) = P_0 e^{y(s)}$ is exceeded. The expectation of the spot cost $C(s)$ is

$$\begin{aligned} \mathbf{E}_y[C(s)] = \int_{-\infty}^{\infty} P_0\pi_t \left(e^{y(q)} - e^{y(s)}\right)_+ \\ \cdot p(y(q)|Z_t)dy(q), \end{aligned} \tag{34}$$

where q and s are quantile levels and the integrand is the spot cost times the excess demand. The spot price π_t is assumed deterministic and known ahead of time. Assuming the spot prices are known yields the same result as assuming they are forecasted and the forecast error is independent of the load.

The pdf in (34) can be expressed in terms of the quantile levels q in accordance with (2) by using

$$p(y|Z_t) = dq/dy(q). \tag{35}$$

Changing the integration variable in (34) to q and using (35) yields

$$\mathbf{E}_y[C(s)] = P_0\pi_t \int_s^1 \left(e^{y(q)} - e^{y(s)}\right) dq. \tag{36}$$

Integral (36) is computed by substituting the model (32) for $y(q)$. Expressions (32), (36) break into three parts: the numerical model for the middle part of the distribution and the two analytical models for the tails.

$$\mathbf{E}_y[C(s)] = P_0\pi_t \cdot (B_L(s) + B_M(s) + B_R(s)), \tag{37}$$

where the subscripts L, M, and R indicate the left tail, the middle part, and the right tail respectively. The integrals $B_L(s)$, $B_M(s)$, and $B_R(s)$ in (37) are computed as follows

$$\begin{aligned} B_L(s) &= \int_{\min(s,q_1)}^{q_1} \left(e^{y(q)} - e^{y(s)}\right) dq \\ &= q_1^{1-\gamma_L} \left(q_1^{\gamma_L} - \min(s,q_1)^{\gamma_L}\right)/\gamma_L \cdot e^{y(q_1)} \\ &\quad - e^{y(s)} \left(q_1 - \min(s,q_1)\right), \end{aligned} \tag{38}$$

$$B_M(s) = \int_{\min(\max(s,q_1)),q_{n_q})}^{q_{n_q}} \left(e^{y(q)} - e^{y(s)}\right) dq, \tag{39}$$

$$\begin{aligned} B_R(s) &= \int_{\max(s,q_{n_q})}^{1} \left(e^{y(q)} - e^{y(s)}\right) dq \\ &= q_{n_q}^{1-\gamma_R} \left(1 - \max(s,q_{n_q})\right)^{\gamma_R}/\gamma_R \cdot e^{y(q_{n_q})} \\ &\quad - e^{y(s)} \left(1 - \max(s,q_{n_q})\right), \end{aligned} \tag{40}$$

where $\gamma_L = 1 + 1/\theta_L$ and $\gamma_R = 1 - 1/\theta_R$.

The integral in (39) is evaluated numerically for given s using the smoothed non-parametric models described above. The integrals involving the tails have been evaluated analytically assuming they converge, which requires $\theta_R > 1$. The tail parameters estimated for the example load demand dataset were $\theta_L = 39.2248$ and $\theta_R = 31.7821$. This means the tail integrals converge.

B. Cost Optimization Results

The total cost can be computed from (33) and (36) as $T(s) = A(s) + \mathbf{E}[C(s)]$. Based on (33), advance cost $A(s)$ is a non-decreasing function of s. Based on (36), the expected spot cost $\mathbf{E}[C(s)]$ is a non-increasing positive function of s. The numerical results show that there is an optimal trade-off between the advance cost and the spot cost that is defined by s in log-load expression (32) and minimizes the total cost $T(s)$. To find the optimum, $T(s)$ is computed numerically on a grid of the values of s by evaluating formulas (38)–(40) for the trained non-parametric model of the load.

The non-parametric model was trained on the data set as described in Section V. As an example, the time sample of January 11, 2011 at 10 PM was selected, when $P_t = 20.371$ GW and $\pi_t =$ \$69.19/MW-hour. The assumed advance price was $\pi_{adv,t} =$\$10/MW-hour. Figure 4 shows the total cost $T(s)$ computed for this time sample t using the trained model.

This stochastic optimization result was compared to the baseline case of the median regression model, where log-load (32) for $s = 0.5$ quantile is used. The baseline total cost is higher by approximately \$5,096.40/hour. This hourly difference corresponds to \$44,644,464/year. A summary of the total cost and the savings is shown in Table II.

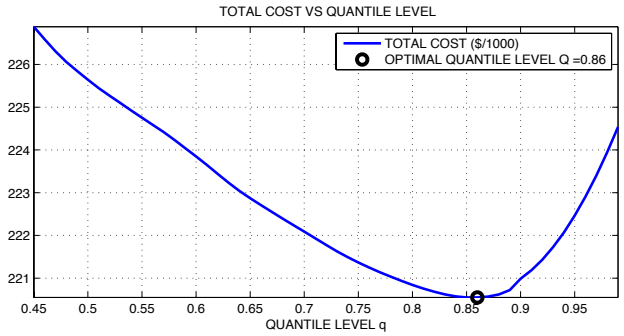

Fig. 4. The total cost and the optimal quantile level Q gives the minimum of the total cost.

TABLE II
COST RESULTS

Strategy/Model	Smoothed QR	Median
Total Cost	\$220,550	\$225,640
Percentage Savings	2.26%	0%

ACKNOWLEDGEMENT

We wish to thank Christopher Fougner for his proximal operator graph solver and adaptive ρ updates that helped us produce the results of this paper. His open source code can be found at [22].

REFERENCES

[1] R. Koenker, *Quantile regression*. No. 38, Cambridge university press, 2005.

[2] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability & Its Applications*, vol. 14, no. 1, pp. 153–158, 1969.

[3] R. Koenker, P. Ng, and S. Portnoy, "Quantile smoothing splines," *Biometrika*, vol. 81, no. 4, pp. 673–680, 1994.

[4] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola, "Nonparametric quantile estimation," *The Journal of Machine Learning Research*, vol. 7, pp. 1231–1264, 2006.

[5] X. He, "Quantile curves without crossing," *The American Statistician*, vol. 51, no. 2, pp. 186–192, 1997.

[6] J. Zhuang, I. W. Tsang, and S. C. Hoi, "A family of simple non-parametric kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 1313–1347, 2011.

[7] L. Jiang, H. D. Bondell, and H. J. Wang, "Interquantile shrinkage and variable selection in quantile regression," *Computational statistics & data analysis*, vol. 69, pp. 208–219, 2014.

[8] V. Chernozhukov, I. Fernández-Val, and A. Galichon, "Quantile and probability curves without crossing," *Econometrica*, vol. 78, no. 3, pp. 1093–1125, 2010.

[9] L. Kong and I. Mizera, "Quantile tomography: Using quantiles with multivariate data," *Statsitica Sinica*, vol. 22, no. 4, pp. 1589–1610, 2008.

[10] M. Hallin, D. Paindaveine, and M. Šiman, "Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth," *The Annals of Statistics*, vol. 38, pp. 635–669, 04 2010.

[11] H. J. Wang, D. Li, and X. He, "Estimation of high conditional quantiles for heavy-tailed distributions," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1453–1464, 2012.

[12] L. Qian, L. Yongli, and W. Chong, "The risk linkage effects of stock indexes based on quantile regression and granger causality test," in *Control and Decision Conference (CCDC), 2013 25th Chinese*, pp. 4252–4257, May 2013.

[13] H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts," *Wind Energy*, vol. 9, no. 1-2, pp. 95–108, 2006.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[15] N. Parikh and S. Boyd, "Block splitting for distributed optimization," *Mathematical Programming Computation*, vol. 6, no. 1, pp. 77–102, 2014.

[16] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

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

[19] 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, pp. 1659–1667, December 1996.

[20] Crowdanalytix.com, "Global energy forecasting competition 2014 probabilistic electricity price forecasting." Available: `https://crowdanalytix.com/contests/global-energy-forecasting-competition-2014-probabilistic-electricity-price-forecasting#`.

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

[22] C. Fougner, "Proximal operator graph solver." Available: `https://github.com/foges/pogs`.