Sandia National Laboratories

Towards dynamic quantile function models for anomaly detection

PRESENTED BY

Peter Jacobs and **Lekha Patel**

Joint work with Anirban Bhattacharya and Debdeep Pati (TAMU)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of

boratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LL Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE

Outline

1. Introduction: Problem Definition, Motivating Applications, and Prior Work

2. Our Approach: General Framework, Constraints, and Current Model

3. Conclusions

Outline

1. Introduction: Problem Definition, Motivating Applications, and Prior Work
2. Our Approach: General Framework, Constraints, and Current Model
3. Conclusions

Problem Definition

We observe univariate $X_1 \sim F_1, X_2 \sim F_2, \dots$ in a stream

$$F_1, F_2, \dots$$

vary through time. At each time $t \geq 1$, we want to estimate the

$$0 < q_1 < q_2 < \cdots < q_K < 1$$

quantiles of F_t

Motivating Applications

Motivating Applications

Detecting Malicious Activity in a Stream of Computer Network Data

- Multivariate points in the stream of network data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ are converted via feature engineering to a discriminative 1d stream $X_1, X_2, \ldots$ as in (Barata, 2021)
- Raises in the $.9$ quantile without changes in the $.85$ quantile could indicate a small group of machines with a common characteristic have become infected
- Quantile tracking also

Motivating Applications

Detecting Malicious Activity in a Stream of Computer Network Data

- Multivariate points in the stream of network data $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are converted via feature engineering to a discriminative 1d stream $X_1, X_2, \ldots$ as in (Barata, 2021)
- Raises in the $.9$ quantile without changes in the $.85$ quantile could indicate a small group of machines with a common characteristic have become infected
- Quantile tracking also

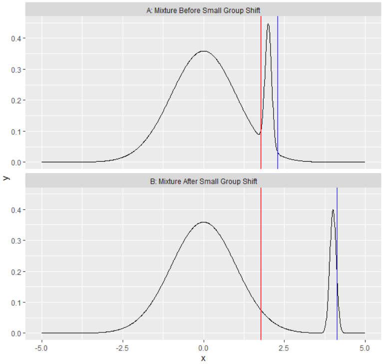

Figure: Red line is $.86$ quantile. Blue line is $.99$ quantile

Motivating Applications

Motivating Applications

Monitoring for Oil Price Shocks over Days

- Changes in the $.9$ quantile of world oil prices without changes in the $.99$ quantile indicates that oil producing nations that had expensive prices became more expensive but the countries with the most costly oil still had the most costly oil
- Useful for quantile quantile regression as in Barata (2021)

Motivating Applications

Monitoring for Oil Price Shocks over Days

- Changes in the $.9$ quantile of world oil prices without changes in the $.99$ quantile indicates that oil producing nations that had expensive prices became more expensive but the countries with the most costly oil still had the most costly oil
- Useful for quantile quantile regression as in Barata (2021)

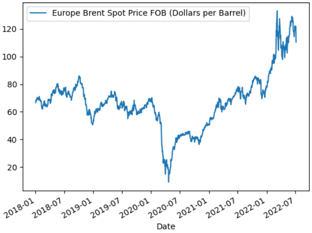

Figure: Time Series of BRENT Crude oil Dollars Per Barrel

Prior Work

- Dynamic **point estimation** of many quantiles that is **order preserving** is well studied in Barata (2021) and Barata (2021)

Prior Work

- Dynamic **point estimation** of many quantiles that is **order preserving** is well studied in Barata (2021) and Barata (2021)
  - Basic idea is update rule such as

$$\hat{Q}_q(n+1) = \begin{cases} (1+\lambda q)\hat{Q}_q(n) & X_n > \hat{Q}_q(n) \\ (1-\lambda(1-q))\hat{Q}_q(n) & X_n \leq \hat{Q}_q(n) \end{cases}$$

Prior Work

- Dynamic **point estimation** of many quantiles that is **order preserving** is well studied in Barata (2021) and Barata (2021)
  - Basic idea is update rule such as

$$\hat{Q}_q(n+1) = \begin{cases} (1+\lambda q)\hat{Q}_q(n) & X_n > \hat{Q}_q(n) \\ (1-\lambda(1-q))\hat{Q}_q(n) & X_n \leq \hat{Q}_q(n) \end{cases}$$

  - Extension to multiple quantiles via rules that constrain the order on the estimates

Prior Work

- Dynamic **point estimation** of many quantiles that is **order preserving** is well studied in Barata (2021) and Barata (2021)
  - Basic idea is update rule such as

$$\hat{Q}_q(n+1) = \begin{cases} (1+\lambda q)\hat{Q}_q(n) & X_n > \hat{Q}_q(n) \\ (1-\lambda(1-q))\hat{Q}_q(n) & X_n \leq \hat{Q}_q(n) \end{cases}$$

  - Extension to multiple quantiles via rules that constrain the order on the estimates
  - Consistency in iid setting is proved: If $F_1 = F_2 = \cdots = F$, then in probability if $Q_q = F^{-1}(q) > 0$, and $Q_q(0) > 0$

$$\lim_{n\lambda \to \infty, \lambda \to 0} \hat{Q}_q(n) = Q_q$$

Prior Work

- Dynamic **point estimation** of many quantiles that is **order preserving** is well studied in Barata (2021) and Barata (2021)
  - Basic idea is update rule such as

$$\hat{Q}_q(n+1) = \begin{cases} (1+\lambda q)\hat{Q}_q(n) & X_n > \hat{Q}_q(n) \\ (1-\lambda(1-q))\hat{Q}_q(n) & X_n \leq \hat{Q}_q(n) \end{cases}$$

  - Extension to multiple quantiles via rules that constrain the order on the estimates
  - Consistency in iid setting is proved: If $F_1 = F_2 = \cdots = F$, then in probability if $Q_q = F^{-1}(q) > 0$, and $Q_q(0) > 0$

$$\lim_{n\lambda \to \infty, \lambda \to 0} \hat{Q}_q(n) = Q_q$$

- In the non-dynamic setting $F = F_1 = F_2 = \ldots$, frequentist approaches have worked out asymptotically valid interval estimation for a wide range of distributions Barata (2021), Barata (2021), Barata (2021)

Prior Work

- Dynamic **point estimation** of many quantiles that is **order preserving** is well studied in Barata (2021) and Barata (2021)
  - Basic idea is update rule such as

$$\hat{Q}_q(n+1) = \begin{cases} (1+\lambda q)\hat{Q}_q(n) & X_n > \hat{Q}_q(n) \\ (1-\lambda(1-q))\hat{Q}_q(n) & X_n \le \hat{Q}_q(n) \end{cases}$$

  - Extension to multiple quantiles via rules that constrain the order on the estimates
  - Consistency in iid setting is proved: If $F_1 = F_2 = \cdots = F$, then in probability if $Q_q = F^{-1}(q) > 0$, and $Q_q(0) > 0$

$$\lim_{n\lambda \to \infty, \lambda \to 0} \hat{Q}_q(n) = Q_q$$

- In the non-dynamic setting $F = F_1 = F_2 = \ldots$, frequentist approaches have worked out asymptotically valid interval estimation for a wide range of distributions Barata (2021), Barata (2021), Barata (2021)
  - These methods are inherently unable to adapt to change because they weight each member of the sample equally

Prior Work

- Dynamic credible interval construction for a single quantile is also well studied Barata (2021)
- This work suffers from a stochastic ordering violation
  - The model can be fit to 2 quantiles separately but if estimating quantiles q_1 and q_2 where $q_1 < q_2$ it is possible there will be a $c \in \mathbb{R}$ s.t $P(Q_{1t} > c|X_1, \dots, X_t) > P(Q_{2t} > c|X_1, \dots, X_t)$
  - This type of contradiction is likely to occur when estimating extreme low quantiles

Outline

1 Introduction: Problem Definition, Motivating Applications, and Prior Work

2 Our Approach: General Framework, Constraints, and Current Model

3 Conclusions

General Framework: HMM

We will approach the dynamic many quantile estimation problem using a hidden markov model structure as well. Recall we seek to estimate the $0 < q_1 < \cdots < q_K < 1$ quantiles of F_t for each t. We propose the following model

Initial Distribution

$$\boldsymbol{Q}_0 = (Q_{10}, \ldots, Q_{K0}) \sim G_0$$

where G_0 is some to be determined parametric family of distributions on $\{(x_1, \ldots, x_K) | -\infty < x_1 < x_2 < \cdots < x_K < \infty\}$.
And for $t \geq 1$

Transition Distribution

$$\boldsymbol{Q}_t | \boldsymbol{Q}_{t-1} \sim G(\boldsymbol{q}_{t-1}, \boldsymbol{V}_t)$$

where G is some to be determined parametric family of distributions on $\{(x_1, \ldots, x_K) | -\infty < x_1 < x_2 < \cdots < X_K < \infty\}$ parameterized by the underlying quantile vector at time $t-1$ and a vector of parameters dictating the variance of transition at time

General Framework: HMM

Emission Distribution

$$X_t|\mathbf{Q}_t \sim Hist_{q_1,\ldots,q_K}(\mathbf{Q}_t, s_t, e_t)$$

where $Hist_{q_1,\ldots,q_K}(\mathbf{Q}_t, s_t, e_t)$ is the distribution coming from the family

$$\begin{aligned} \mathcal{F}_{(q_1,\ldots,q_K)} &= \{Hist_{q_1,\ldots,q_K}(\mathbf{Q}, s, e) = p(x|\mathbf{Q}, s, e) \\ &= q_1(s * exp(-(q_1 - x)s)\mathbb{I}(x < Q_1) + \\ &\sum_{i=2}^{K-1}(q_i - q_{i-1})\mathbb{I}(Q_i \leq x < Q_{i+1})\frac{1}{Q_i - Q_{i-1}} \\ +(1 - q_k)\mathbb{I}(x \geq Q_K)(e * exp(-(x - q_{M-1})e))|\infty &< Q_1 < \cdots < Q_K, s > 0, e > 0\} \end{aligned} \quad (1)$$

Constraints

The following 3 constraints allow this work to build upon what has already been done

- Order preservation
- Consistent estimation
- Adaptable

Constraints: Order Preservation

Constraint

Order Preserving Estimation: *For each time $t \geq 0$ the marginal filter distributions should be stochastically ordered. In other words* $p(Q_{1t}|X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) \preceq \cdots \preceq p(Q_{Kt}|X_1 = x_1, \ldots, X_{t-1} = x_{t-1})$

Constraints: Consistent Estimation

Constraint

Consistent in iid Circumstances *For every F, if* $F_1 = F_2 = \cdots = F$*, i.e if the stream is truly an iid stream possessing true quantiles* $Q_1 = F^{-1}(q_1)$*, ...*$Q_K = F^{-1}(q_k)$*, then for* $1 \leq k \leq K$ *and every* $\epsilon > 0$

$$\lim_{t\|\mathbf{V}_t\|\to\infty, \|\mathbf{V}_t\|\to 0} p(|Q_{kt} - Q_k| > \epsilon | X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = 0$$

Constraints: Adaptable

Constraint

Fast Adaptation *In practice we require* $\mathbf{V}_t$ *to not tend to zero so that adaptation to distributional change is possible. At least, we must empirically verify fast adaption to change for non-varying* $\mathbf{V}$*. Ideally we can specify a deterministic algorithm for dynamically varying* $\mathbf{V}_t$ *that maintains consistency but allows adaptation as in Barata (2021).*

Specific Model Under Analysis

This model maintains the $(1/K, 2/K, \ldots, (K-1)/K)$ quantiles in a transformed space that allows for multivariate normal transitions Let $\boldsymbol{a}_t = (a_{1t}, a_{2t}, \ldots, a_{(K-1)t}, \alpha_t, \beta_t)$ for $t \geq 1$. Also define for $t \geq 1$.

$$Q_{1t}(\boldsymbol{a}_t) = a_{1t}$$

and for $2 \leq j \leq K-1$

$$Q_{jt}(\boldsymbol{a}_t) = a_{1t} + \sum_{t=2}^{j} exp(a_{jt})$$

and

$$s(\boldsymbol{a}_t) = exp(\alpha_t)$$

and

$$e(\boldsymbol{a}_t) = exp(\beta_t)$$

In the “a” space, the first dimension is the $1/K$ quantile, the second dimension is the log of the difference between the $2/K$ and $1/K$ quantile, the third dimension is the log of the difference between $3/K$ and $2/K$ quantile, and so on.

Transition Distributions: Specification of G

$\boldsymbol{a}_t|\boldsymbol{a}_{t-1} \sim MVN(\boldsymbol{a}_{t-1}, \Sigma_t)$ where Σ_t is $K+1 \times K+1$ diagonal and

$$Var(\boldsymbol{a}_{1t}|\boldsymbol{a}_{1(t-1)}) = \sigma_L^2(t)$$

and for $2 \leq j \leq K-1$

$$Var(\boldsymbol{a}_{jt}|\boldsymbol{a}_{j(t-1)}) = \zeta(t, K)$$

and

$$Var(\alpha_t|\alpha_{t-1}) = Var(\beta_t|\beta_{t-1}) = \sigma_B^2(t)$$

Note in this model that $\boldsymbol{V}_t := (\sigma_L^2(t), \zeta(t, K), \sigma_B^2(t))$

Initial Distribution: Specification of G_0

$\boldsymbol{a}_1 \sim MVN(\left[0, c, c, \ldots c, d, d\right]^T, \Sigma_0)$. Σ_0 is diagonal; the first entry is $\sigma_L^2(0)$. The last two entries are $\sigma_B^2(0)$. The other diagonal entries are $\zeta(0, K)$

Simulations

Simulations

Figure: Quantiles over a standard Normal.

Simulations

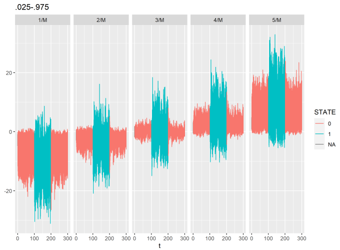

Figure: Quantiles over a standard Normal.

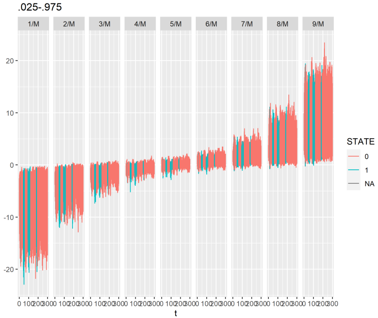

Figure: Quantiles over a t-distribtution.

Outline

Conclusions

- We have built a flexible framework for modeling multiple quantiles for streaming data.
- Simulated streaming data using an example model class shows promise in using this method for detecting anomalies.
- Regulatory conditions are needed to ensure the estimated quantile function is proper along the stream.
- Establishing posterior consistency for the dynamic setting requires focus on the iid setting which is currently in progress.
- Once esatblished, will focus on credible intervals of the quantiles can be used to do anomaly detection with risk quantification.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

Hammer, H. L., A. Yazidi, M. A. Riegler, and H. Rue (2022). Efficient quantile tracking using an oracle. *Applied Intelligence*, 1–12.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

Hammer, H. L., A. Yazidi, M. A. Riegler, and H. Rue (2022). Efficient quantile tracking using an oracle. *Applied Intelligence*, 1–12.

Hammer, H. L., A. Yazidi, and H. Rue (2020). Tracking of multiple quantiles in dynamically varying data streams. *Pattern Analysis and Applications 23*(1), 225–237.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

Hammer, H. L., A. Yazidi, M. A. Riegler, and H. Rue (2022). Efficient quantile tracking using an oracle. *Applied Intelligence*, 1–12.

Hammer, H. L., A. Yazidi, and H. Rue (2020). Tracking of multiple quantiles in dynamically varying data streams. *Pattern Analysis and Applications 23*(1), 225–237.

Hammer, H. L., A. Yazidi, and H. Rue (2021). Joint tracking of multiple quantiles through conditional quantiles. *Information Sciences 563*, 40–58.

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

Hammer, H. L., A. Yazidi, M. A. Riegler, and H. Rue (2022). Efficient quantile tracking using an oracle. *Applied Intelligence*, 1–12.

Hammer, H. L., A. Yazidi, and H. Rue (2020). Tracking of multiple quantiles in dynamically varying data streams. *Pattern Analysis and Applications 23*(1), 225–237.

Hammer, H. L., A. Yazidi, and H. Rue (2021). Joint tracking of multiple quantiles through conditional quantiles. *Information Sciences 563*, 40–58.

Hutson, A. D. (1999). Calculating nonparametric confidence

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

Hammer, H. L., A. Yazidi, M. A. Riegler, and H. Rue (2022). Efficient quantile tracking using an oracle. *Applied Intelligence*, 1–12.

Hammer, H. L., A. Yazidi, and H. Rue (2020). Tracking of multiple quantiles in dynamically varying data streams. *Pattern Analysis and Applications 23*(1), 225–237.

Hammer, H. L., A. Yazidi, and H. Rue (2021). Joint tracking of multiple quantiles through conditional quantiles. *Information Sciences 563*, 40–58.

Hutson, A. D. (1999). Calculating nonparametric confidence

References

Barata, R. A. (2021). *Flexible Dynamic Quantile Linear Models*. Ph. D. thesis, University of California, Santa Cruz.

Chen, S. X. and P. Hall (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 1166–1181.

Chu, F. and M. K. Nakayama (2012). Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 22*(2), 1–25.

Hammer, H. L., A. Yazidi, M. A. Riegler, and H. Rue (2022). Efficient quantile tracking using an oracle. *Applied Intelligence*, 1–12.

Hammer, H. L., A. Yazidi, and H. Rue (2020). Tracking of multiple quantiles in dynamically varying data streams. *Pattern Analysis and Applications 23*(1), 225–237.

Hammer, H. L., A. Yazidi, and H. Rue (2021). Joint tracking of multiple quantiles through conditional quantiles. *Information Sciences 563*, 40–58.

Hutson, A. D. (1999). Calculating nonparametric confidence