Quantile Tail Model

Cássio Jandir Pagnoncelli

November 30, 2025

Abstract

This paper sets out a framework for non-linear quantile regression with an ex This paper sets out a framework for non-linear quantile regression with an explicit treatment of the tails. We review the econometric underpinnings of conditional quantiles, describe a tree-based boosting approach for estimating them, and present a tail module that marries a quantile grid to an EVT bridge. The emphasis is on calibration, stability, and validation standards appropriate for applied econometric work.

1 Foundational Landscape

Classical models are designed to minimise central moment residuals and, as a result, these tend to be dispersed near-symmetrically around the estimated conditional mean.

Irrespective of the presence of heteroskedasticity and assymetry in the data, models tend to underestimate events in peripheric, underpopulated regions. Because extreme events lie in the tails of the distribution and are rare by definition, they are overshadowed by the bulk of the data that are conditioned on the central moment behaviour, generally yielding underestimated risk measures.

Quantile regression (QR) is a nonparametric method that provides a natural framework for estimating conditional quantiles and is pivotal in risk assessment, particularly in Conditional Value at Risk (CVaR) models.

By replacing the traditional least squares loss gradient function with the asymmetric pinball loss, QR directly targets specific quantiles of the response variable's conditional distribution in a logistic penalisation fashion. Vanilla pinball loss heavily penalises under-predictions whilst reinforcing exceedances much less severely, however, this can be adapted to account for different risk metrics.

Extreme Value Theory (EVT) provides asymptotic structure for exceedances producing a richer glimpse into the tail behaviour. Peaks-over-threshold arguments motivate *Generalised Pareto* models at high thresholds, yielding shape (ξ) and scale (β) descriptors that can extend a quantile curve beyond observed order statistics. Extrapolation based on tail geometry remains defensible only when exceedances are sufficient and diagnostics stable.

These models provide quantile surfaces rather than point-forecast estimates, hence the analysis and validation metrics must be adapted from traditional regression to reflect model quality, namely, coverage, calibration, baseline pseudo-R^2, Kendall rank concordance, and Probability Integral Transform (PIT) histograms.

2 Quantile Boosting Machine (QBM)

Quantile Boosting Machine is a quantile regression-inspired tree-based learner designed to estimate conditional quantiles using pinball loss gradient function and tree-based boosting machines, in this case, LightGBM model with vanilla pinball loss.

Model fit demands careful hyperparameter governance. Because quantile targets are more sensitive to overfitting than means, models tend to better generalise with novel data when the architecture display shallow trees with limited leaves, conservative learning rates, and early stopping based on validation pinball loss to halt training before overfitting ensues. Traditional regularisation techniques such as bootstrapping and subsampling, L1/L2 penalties on leaf weights, and minimum child samples are effective in QBM, curbing variance without distorting the quantile objective.

QBM regressor optimises for separating the τ-quantile from the rest of the distribution, akin to the logistic regressor. As the data gets farther away from the τ-quantile, the gradient signal weakens, thus reducing the point-forecast accuracy incentive. Because it does not learn the full distribution nor the conditional distribution, traditional evaluation metrics should be analysed from a quantile-aware angle.

Empirical coverage is the realised frequency with which model predicted quantile bounds actually contain the observed outcomes. This is the primary metric for quantile model validity and should align with nominal levels, as well as its error, namely quantile calibration error (QCE), which quantifies deviations. Note deviations from nominal levels indicate miscalibration, with positive QCE suggesting underestimation and negative QCE indicating overestimation of the quantile.

Kendall correlation, similar to Spearman's rank correlation, measures the ordinal association between predicted and observed quantiles. Here, the method is adapted to quantify rank-based concordance between predicted and observed quantiles.

Goodness-of-fit metric R^2 surveys the full distribution of the response variable. In central-moment regressors the scale of explained variance is expected to be substantially higher than in quantile regressors; reason is in quantile models only a fraction of the data near the quantile contribute significantly to the loss gradient, therefore a comparably much lower R^2 is not necessarily a bad fit. Koenker-Machado provide a pseudo-R^2 definition tailored for quantile regression that contrasts the pinball loss of the fitted model against a naive intercept-only model. For a proper interpretation, depending on the mass beyond the quantile of interest, is that pseudo-R^2 indicates the proportional reduction in pinball loss achieved by the model relative to a naive benchmark, hence lower values may merely indicate the model successfully capturing the quantile in data-sparse regions, while a higher level may explain the shape and structure of the tail.

Deviations from ideal uniformity in PIT histograms are common in quantile models and signal miscalibration. Left/right-skewed indicate systematic under/over-prediction, while U-shaped/hump-shaped histograms suggest under- or overestimated variability.

An attribute of interest is monotonicity across quantile levels, that is, estimated surface should be non-decreasing (or non-increasing) in τ for all x. Isotonic regression during training can enforce this property while fitting the surface to the dependent variable, that is, $\hat{q}_{\tau_1}(x) \leq \hat{q}_{\tau_2}(x)$ for $\tau_1 < \tau_2$, while minimising $\sum_{i=1}^{n} \big(y_i - \hat{q}_\tau(x_i)\big)^2$ with an unique solution, this is efficiently recovered by the Pool Adjacent Violators Algorithm (PAVA), which iteratively merges neighbouring points that violate the contrainst, at the same time making no assumptions regarding the distribution beyond the ordering constraint. Effectively this is a projection onto the monotone cone, having the model obeying order, not shape.

3 Quantile Tail Model

Quantile tails demand structure that a single fit cannot provide. The qtail construction answers this by coupling a bank of non-linear quantile estimates from QBM with an extreme-value overlay, so that the bulk and the tail remain coherent.

The starting point is a grid of quantile levels. Higher resolution near the tail of interest captures curvature and local shape. Each level is fit with QBM, which supplies non-linear conditional quantiles without hand-crafted features.

These raw curves often cross in finite samples. To enforce logical ordering, qtail projects the grid onto the monotone cone via a row-wise least-squares projection,

$$\min_{Q_{i1} \leq \cdots \leq Q_{iK}} \sum_{k=1}^{K} \big(Q_{ik} - \hat{q}_{\tau_k}(x_i) \big)^2,$$

which smooths idiosyncratic variation and reduces variance where data are sparse.

After monotonicity, qtail can stack the grid with shrinkage weights. This pooling borrows strength across adjacent levels and tempers the instability that arises when extreme quantiles are fit in isolation.

The tail extension begins at a chosen threshold quantile inside the grid. Exceedances beyond that threshold are modeled with a Generalized Pareto tail, delivering shape and scale parameters that describe the marginal tail beyond the fitted boundary.

Extrapolation to a more extreme level p is then obtained by adding a modeled exceedance to the threshold quantile,

$$\hat{q}_p = \hat{q}_{\tau^*} + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{1-p}{1-\tau^*} \right)^{-\hat{\xi}} - 1 \right],$$

with the exponential limit as $\hat{\xi} \to 0$. This links the tree-based bulk fit to an EVT-consistent tail in a single surface.

Alignment between the grid and the EVT layer is deliberate: the threshold coincides with a fitted quantile from QBM, avoiding discontinuities at the splice point. This keeps CVaR and other tail functionals consistent with the underlying quantile surface.

Diagnostics gate the EVT step. Adequate exceedance count, stable tail index estimates, and acceptable calibration at the threshold are prerequisites. If these fail, qtail defaults to the stacked monotone grid without extrapolation.

Uncertainty should be propagated from both components. Reporting tail risk without accounting for variance in the grid and in the GPD fit overstates confidence; interval estimates should reflect both sources.

Threshold choice balances bias and variance. Thresholds set too low bias the tail shape; thresholds set too high starve the EVT fit. Empirical checks on stability across nearby thresholds are recommended.

Monotonicity is more than housekeeping. By shrinking adjacent levels, it dampens noise amplification in the far tail and makes PIT and calibration curves more stable across levels.

Stacking is likewise functional. Shrinkage toward equal weights curbs over-reliance on any single quantile level and mitigates the high variance inherent in tail targets.

The EVT bridge provides structure where QBM alone is extrapolating. By importing asymptotic shape information, it offers a disciplined extension beyond observed support—when the data justify it.

Coherence across risk measures is a central advantage. With a calibrated tail, CVaR and stressed shortfall align with the quantile surface rather than being bolted on separately.

From a computational standpoint, the projection and stacking steps are lightweight relative to fitting the grid, so the marginal cost of enforcing monotonicity and pooling is small.

In dependent data, declustering exceedances before fitting the GPD avoids inflated tail estimates due to serial correlation. This keeps the EVT layer compatible with econometric time-series practice.

Model monitoring remains necessary after deployment. Drift in calibration or tail index stability signals the need to adjust the grid, retrain qrb fits, or revise thresholds.

The qtail design is modular: change the grid density, swap the base learner, or toggle EVT based on diagnostics. This flexibility lets practitioners match model complexity to data depth and risk appetite.

In summary, qtail ties a calibrated bulk (via qrb), a logically ordered grid, and an EVT-consistent extension into a single tail surface. The gain over a single quantile fit is stability, logical coherence, and a disciplined path to extrapolation when the evidence supports it.

Conclusion

Quantile methods are the natural language of risk because they target conditional bounds rather than averages. Tree-based quantile regression delivers these bounds with non-linear flexibility and operational speed, while the quantile grid with monotone pooling addresses crossing and variance in the tails. The EVT bridge adds principled extrapolation when exceedances warrant it; otherwise, the stacked monotone fit remains the conservative choice. A disciplined workflow—coverage/QCE and calibration checks, careful capacity control, and cautious threshold selection—turns these pieces into a practical model for tail-aware forecasting and event monitoring in applied econometrics. The architecture is modular: the quantile engine can be swapped, the grid can be densified or thinned, and the EVT layer can be toggled based on diagnostics. This modularity allows practitioners to balance flexibility with interpretability and to tailor the model to data availability and tail risk appetite.

Open directions include adaptive threshold selection for EVT, automated monitoring of calibration drift in streaming settings, and closer coupling between quantile surfaces and downstream risk measures such as CVaR and stressed expected shortfall. Empirical studies across assets and macro series can further quantify how extrapolation risk trades off against tail coverage in live environments. The overarching aim is a calibrated, transparent pipeline for tail risk: non-linear quantiles that respect coverage, a tail layer that is only invoked when justified, and diagnostics that surface instability before it harms decisions. Continued work on uncertainty quantification for the combined model—and on stress-testing under regime shifts—will make the framework more robust for high-stakes econometric applications.

References

[1] R. S. Tsay. *Analysis of Financial Time Series*. Wiley, 2010.

[2] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[3] R. Koenker and J. Machado. Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999.

[4] G. Ke, Q. Meng, T. Finley, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *NeurIPS*, 2017.