Quantile Tail Model

Cássio Jandir Pagnoncelli

November 30, 2025

Abstract

A robust econometric model for estimating conditional surface with explicit tail treatment. The model combines a grid of quantile estimates that individually contribute to a stacked, monotonic surface fit of Generalised Pareto, adressing quantile crossing, high variance, and extrapolation power.

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.

Finally, QBM is a computationally efficient method for estimating conditional quantiles serving as the backbone for the more complex quantile tail model ahead.

3 Quantile Tail Model

Quantile tail (qtail) is a composite model that integrates a grid of non-linear quantile estimates from QBM with an extreme-value-theory (EVT) tail extension, addressing the dual challenges of

quantile crossing, high variance, and extrapolation power beyond observed data.

The construction couples 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 $\tau = \{\tau_1, \tau_2, \ldots, \tau_K\}$, each corresponding to a QBM fit, supplying non-linear quantile surfaces across the feature space.

These raw curves often cross in finite samples. To enforce logical ordering, qtail projects the grid onto the monotone cone projection, which smooths idiosyncratic variation and reduces variance where data is sparse.

Combining multiple quantile estimates provides a more balanced estimate into a single robust forecast at each level. After monotonicity, qtail can stack the grid, namely a level 2 regressor, with shrinking weights to adjust for explosive extreme tail behaviour.

The tail extension begins at a chosen threshold quantile inside the grid, τ^*. Exceedances beyond this threshold are modeled with a Generalized Pareto Distribution (GPD) tail, delivering shape and scale parameters that describe the marginal tail beyond the fitted boundary. Formally, exceedances over the threshold quantile $\hat{q}_{\tau^*}(x)$ are defined as $e_i = y_i - \hat{q}_{\tau^*}(x_i)$, for for $y_i > \hat{q}_{\tau^*}(x_i)$. These exceedances are then used to fit a GPD, estimating the tail shape $\hat{\xi}$ and scale $\hat{\beta}$ parameters via maximum likelihood.

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

$$\hat{q}_p = \hat{q}_{\tau^*} + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{1-p}{1-\tau^*} \right)^{-\hat{\xi}} - 1 \right), \qquad \hat{\xi} \neq 0$$

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.

Adequate exceedance count, stable tail index estimates, and acceptable calibration at the threshold are top priorities. If these fail, qtail defaults to the stacked monotone grid without extrapolation yielding baseline predictions.

Empirical checks are recommend for stability across nearby thresholds, particularly checking for tail index consistency, calibration, and sufficient exceedance count. Thresholds set too low bias the tail shape; thresholds set too high starve the EVT fit.

Qtail combines a logically ordered grid with an EVT-based extension into an unified tail surface. The gain over a single quantile fit is stability and extrapolation power at and beyond quantile threshold.

4 Implementation

An R package implementing QBM and Quantile Tail is available at

`https://www.github.com/cassiopagnoncelli/qboost`.

Discussion

Quantile Tail displays a robust, creative framework for estimating conditional quantiles with tweakable tail model.

Tail extremes are heavily influenced by heteroskedasticity, entropy, and assymetry in the data. Introspectively, these can weight differently across quantiles, which QBM captures due to its non-parametric nature in picking splits based on informational gain.

Tweaking pinball loss assymetry is encouraged for different, extreme quantile targets when feature engineering provides little to no marginal gain.

Tree-based boosting machine family models currently perform satisfactorily in the non-linear regression landscape, with QBM leveraging quantile estimation. While ensemble methods provide some marginal edge over individual estimators, the stacked GPD model using these estimates consistently displays superior performance on tail extrapolation even in challenging scenarios.

QBM does the heavy lifting when edge-predicting variables are available, with the tail module providing a balanced surface extension.

References

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

[2] R. Koenker; J. Machado *Goodness of Fit and Related Inference Processes for Quantile Regression.* Journal of the American Statistical Association, 2012.

[3] E. Alpaydin. *Introduction to Machine Learning.* MIT, 2010.