Likelihood over Estimation

Robust Quadratic Discriminant Analysis for Heavy-Tailed Distributions — with Theory and Evidence

Niranjana Ambadi LawMate Private Limited
Eugene Pinsky Boston University, Metropolitan College

Seoul, South Korea · Paper #1380

ICML
International Conference On Machine Learning

MOTIVATION

Heavy tails are the rule, not the exception

QDA is a workhorse generative classifier: **closed-form training, $O(Kp^2)$ prediction, calibrated probabilities, inspectable quadratic boundaries.**

But it assumes each class is Gaussian — and much of the real world is not:

Finance

returns & transaction amounts follow stable-like laws, $\alpha \approx 1.4$–1.9

Networks

traffic and attack dynamics exhibit power-law behavior

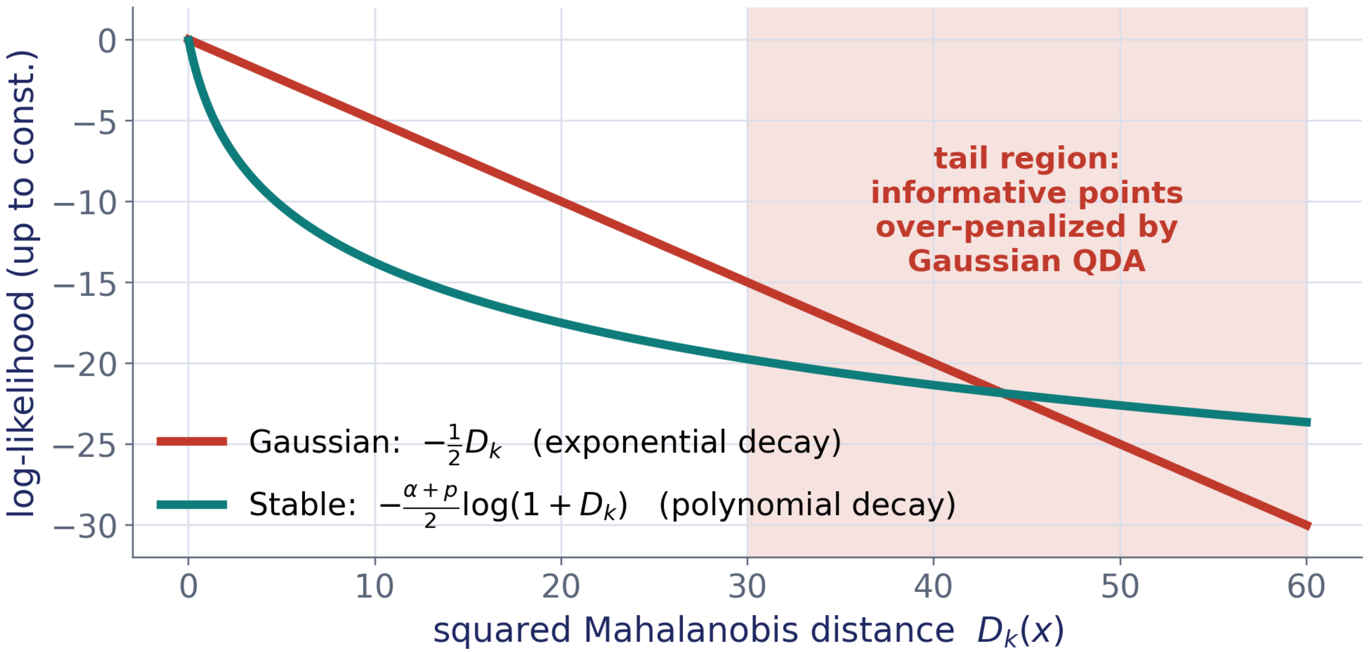
Astronomy

faint pulsar signals hide inside heavy-tailed noise

In these domains the tails are the point: extreme transactions signal fraud, anomalous packets signal intrusions, faint outliers signal discoveries.

THE PROBLEM

Gaussian QDA systematically misclassifies the tails



The Gaussian likelihood decays exponentially in D_k; α-stable decays polynomially.

Two distinct failure modes:

Wrong model — exponential decay assigns vanishing likelihood to tail points, so the classifier is wrong even with perfect parameters.

Wrong estimation — under infinite variance ($\alpha < 2$) the sample covariance blows up and never converges.

Result: distorted boundaries and systematic misclassification exactly where correct decisions matter most.

KEY INSIGHT

Fix the likelihood, not the estimator.

Decision boundaries depend on likelihood ratios. Estimation errors that hit every class alike cancel in those ratios — likelihood misspecification does not.

Robust estimators (Tyler, spatial median)

fix parameter instability — but keep the Gaussian likelihood, leaving the model wrong.

Stable-QDA

fixes the likelihood (polynomial tails) and keeps simple, well-conditioned estimators — this drives most of the improvement.

Why α-stable? The Generalized CLT

The α-stable family generalizes the Gaussian through a stability index $\alpha \in (0, 2]$: **smaller α = heavier tails; α = 2 recovers the Gaussian; α = 1 is Cauchy.**

Ordinary data	**Central Limit Theorem**	averages of finite-variance variables → Gaussian
Heavy-tailed data	**Generalized CLT**	sums of heavy-tailed variables → α-stable — the only possible non-degenerate limits

So α-stable is the "right" model for heavy-tailed phenomena by the same foundational logic that makes Gaussian the right model for ordinary data.

Stable-QDA: an order-equivalent closed-form discriminant

The multivariate stable density is intractable — but classification needs only the likelihood ordering. Its polynomial tail yields the surrogate:

$$\delta_k(x) = \log \hat{\pi}_k \; - \; \frac{\alpha + p}{2} \log\left(1 + \hat{D}_k(x)\right) \; - \; \frac{1}{2} \log \left|\hat{\Sigma}_k\right|$$

$$\hat{D}_k(x) = (x - \hat{\mu}_k)^\top \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k)\,, \qquad \hat{y}(x) = \arg\max_k \delta_k(x)$$

Gaussian: $-\frac{1}{2} D_k$ exponential — crushes tail points

Stable: $-(\alpha+p)/2 \log(1+D_k)$ polynomial — keeps tail signal

$O(Kp^2)$ prediction — same as QDA

$\alpha = 1.5$ fixed default, no tuning

Calibrated softmax log-posteriors

Which estimators? Scale is signal

Tyler's M-estimator is trace-normalized: it estimates shape but discards scale. **QDA's log-determinant term uses scale differences between classes** — so “robust” can mean throwing away signal.

STANDARD (default)

sample mean + Ledoit–Wolf

moderate tails ($\alpha \geq 1.5$) or heteroscedastic classes; shrinkage supplies implicit robustness

ROBUST

spatial median + Tyler

very heavy tails ($\alpha < 1.5$) with similar class scales

Deterministic rule — estimate α (McCulloch quantiles) and the determinant ratio r, then read off:

det ratio r	< 10	10–100	100–1000	> 1000
robust if α <	2.0	1.8	1.7	1.6

Open-source diagnostic: `$ python diagnose_dataset.py --data my.csv --target label`

Theoretical guarantees

Bayes consistency (Thm 6.1)

$$R(\hat{y}_n) \xrightarrow{P} R^* \quad \text{as} \quad n \to \infty$$

With enough data, Stable-QDA converges to the best possible classifier — for α ∈ (1, 2], no finite variance required.

Gaussian QDA is inconsistent (Thm 6.2)

$$\liminf_{n\to\infty} R(\hat{g}_{\text{Gauss}}) \geq R^* + \varepsilon$$

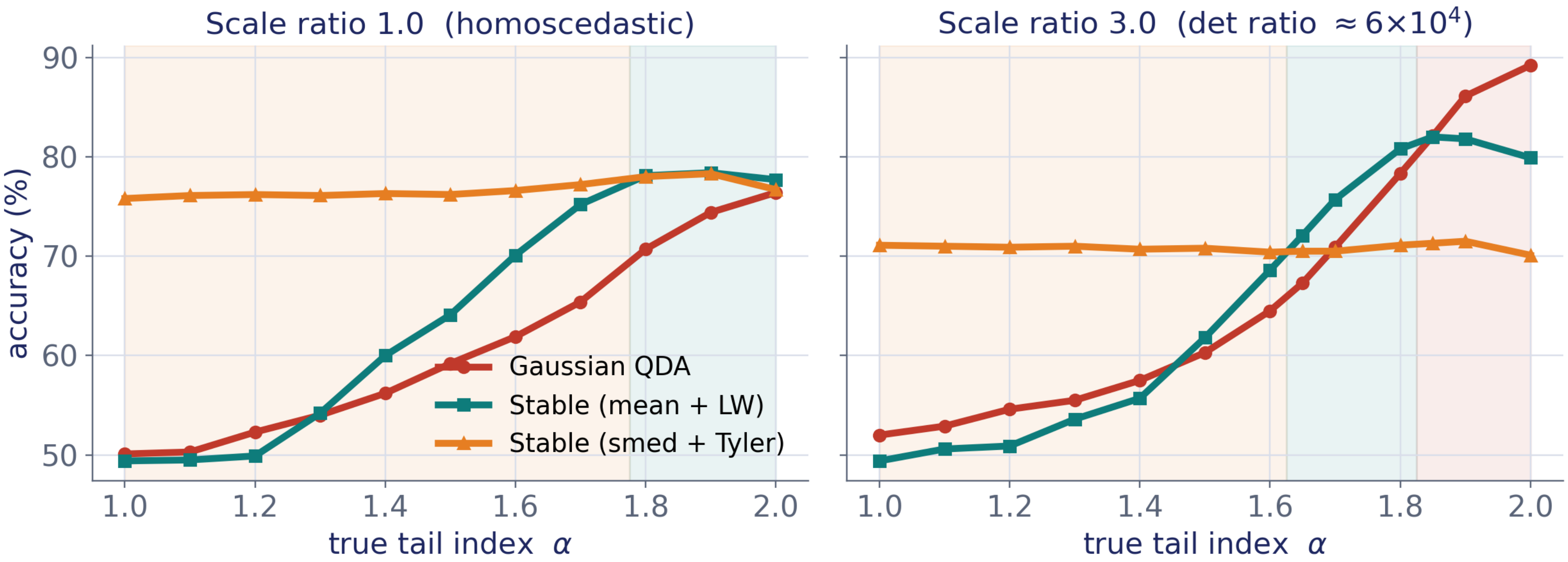
Under heavy tails it converges to the wrong answer permanently, no matter how much data it sees.

Insensitive to α (Prop 6.4)

$$|R(\alpha) - R(\tilde{\alpha})| \leq C\,|\alpha - \tilde{\alpha}|$$

Risk is Lipschitz in α — the log(1+D) form matters, not the exact index. Fixed α = 1.5 ≈ oracle.

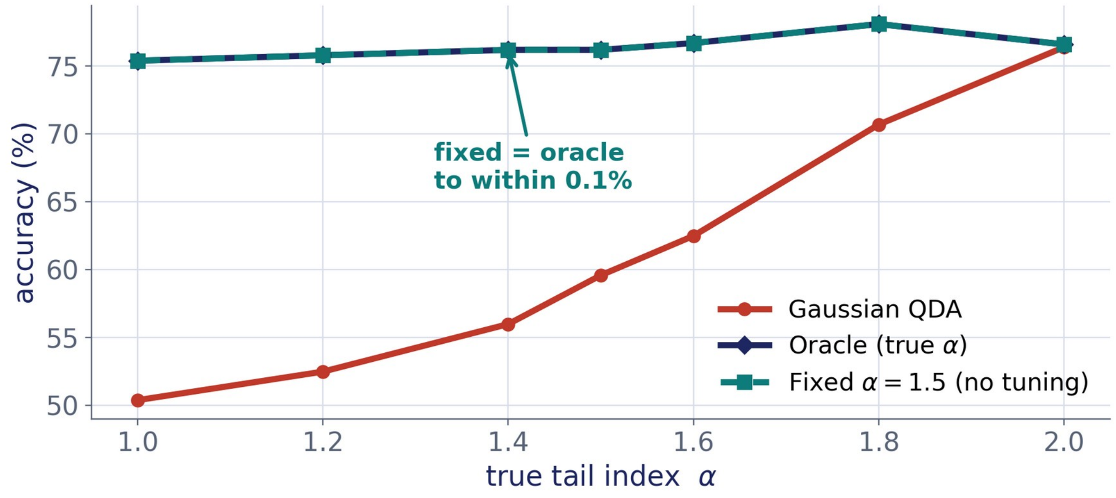
Synthetic study: when does what win?



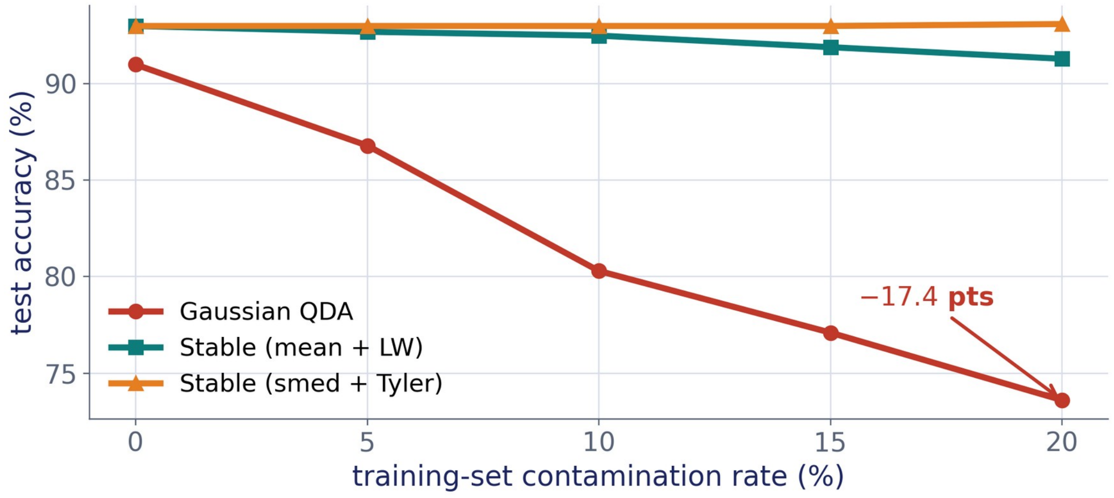
Shaded = best regime, d = 10, n = 500/class, 15 seeds.

Heavy tails (α < 1.5): robust estimators win regardless of scale. Moderate tails + heteroscedasticity: standard estimators preserve the scale signal Tyler destroys. Light tails (α > 1.8): Gaussian suffices.

No tuning needed — and free contamination robustness



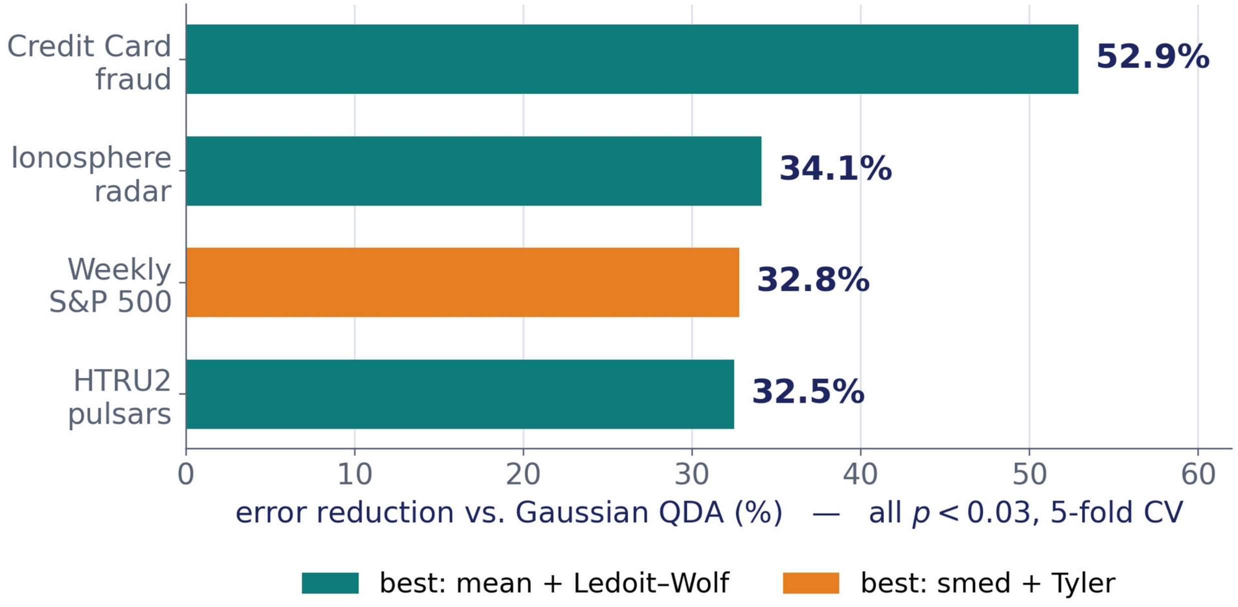
Fixed α = 1.5 matches the oracle within 0.1% at every true α; +15.3% over Gaussian on average.



20% contaminated training data (α = 1.8): Gaussian drops 17.4 pts; Stable-QDA barely moves.

Recommendation: default to α = 1.5 and do not tune it by cross-validation — per-class α estimation never helps.

Real-world benchmarks: 15–53% error reduction



5-fold CV; all improvements significant at $p < 0.03$.

+51.4% PR-AUC on credit-card fraud (det ratio > 10^7) — Tyler's normalization destroys this signal

99.2% tail recall on NetML malware vs. 89.3% for Gaussian QDA (n = 114K)

4 / 4 datasets where the diagnostic picked the winning estimator configuration

TAKEAWAYS

Treat heavy tails as structure, not nuisance

Specify the likelihood first. Correcting the model beats aggressive robust estimation — “outliers” are frauds, intrusions, and pulsars, and a polynomial-tail likelihood lets the classifier use them.

α-stable is the principled choice. The Generalized CLT makes it the natural model for heavy-tailed sums, just as the CLT makes Gaussian natural for ordinary data.

Plug-and-play. Fixed $\alpha = 1.5$ is near-oracle everywhere; same $O(Kp^2)$ cost, closed-form training, calibrated probabilities.

Works with deep learning. A generative head for linear probing on frozen ViT/CLIP/DINOv2 features — empirically heavy-tailed — where generative classifiers converge in $O(\log n)$ vs. $O(n)$ samples.

Future work. Asymmetric stable laws for skewed data; sparse covariance in $p \gg n$.