Closed-Form Beta Distribution Recovery from Sparse Statistics with Random-Forest Implicit Regularization

Jonathan R. Landers

jonathan.robert.landers@gmail.com
github.com/jonland82/seatgeek-beta-modeling
https://orcid.org/0000-0003-1872-6179

Abstract

This work advances distribution recovery from sparse data and ensemble classification through three main contributions. First, we introduce a closed-form estimator that reconstructs scaled Beta distributions from limited statistics (minimum, maximum, mean, and median) via composite quantile and moment matching. The recovered parameters (α, β), when used as features in Random Forest classifiers, improve pairwise classification on time-series snapshots, validating the fidelity of the recovered distributions. Second, we establish a link between classification accuracy and distributional closeness by deriving error bounds that constrain total variation distance and Jensen–Shannon divergence, the latter exhibiting quadratic convergence. Third, we show that zero-variance features act as an implicit regularizer, increasing selection probability for mid-ranked predictors and producing deeper, more varied trees. A SeatGeek pricing dataset serves as the primary application, illustrating distributional recovery and event-level classification while situating these methods within the structure and dynamics of the secondary ticket marketplace. The UCI handwritten digits dataset confirms the broader regularization effect. Overall, the study outlines a practical route from sparse distributional snapshots to closed-form recovery and improved ensemble accuracy, with reliability enhanced through implicit regularization.

Keywords: Scaled Beta Distribution, Random Forest, Implicit Regularization, Time-Series Classification, Ticket Pricing

1. Introduction

Recovering probability distributions from limited information is a central problem in data analysis. In many applied settings, only a small set of summaries (minimum, maximum, mean, and median) is available rather than full samples. For time-series classification, target examples can be compared through compact snapshots of representative, underlying distributions. In this context, ensemble methods such as Random Forests benefit from features derived from reconstructed distributions, provided estimation is tractable and theoretically grounded. In practice, we interpret out-of-sample Random Forest accuracy as a pragmatic gauge of distributional integrity.

This paper presents the following contributions:

1. **Closed-form distribution recovery from limited statistics.** A method is introduced to reconstruct scaled Beta laws from incomplete statistics using composite quantile and moment matching, producing parameters (α, β) that convey shape information beyond location summaries (Sections 4.1 and 4.2). Injecting (α, β) into Random Forests improves pairwise classification of time-series snapshots, indicating that the recovered distributions preserve class-distinctive structure. Recovery is not only theoretically sound but also efficient, with a constant-time closed-form estimator that contrasts with iterative alternatives.

2. **Accuracy–fidelity theory.** A link between predictive accuracy and distributional fidelity is established: Theorem 4.1 bounds total-variation distance by classification error, and Theorem 4.2 strengthens the connection to a quadratic convergence rate in Jensen–Shannon divergence, showing that modest accuracy gains imply disproportionately larger reductions in information-theoretic divergence. Accordingly, classifier performance can be used as an operational proxy for how closely the recovered distributions match the (unobserved) ground truth.
3. **Implicit regularization via zero-variance features.** When zero-variance (constant-value) features are added to the Random Forest ensemble, Theorem 5.1 (*Zero-Variance Dilution Effect*) formalizes how split-selection probabilities are rebalanced away from over-dominant predictors. Corollary 5.1 (*Increased Expected Tree Depth*) and Corollary 5.2 (*Reduced Ensemble Correlation*) show that this dilution yields deeper, more varied trees and lowers inter-tree correlation. Theorem 5.2 (*Continuous Approximation via Zero-Variance Dilution*) together with Corollary 5.3 (*Continuous Accuracy Expansion via Selection Probability*) then shows that the mechanism enables fine-grained control over selection probabilities and tree variety. Although accuracy gains are intentionally modest, the quadratic bound ensures that they correspond to meaningful improvements in distributional correspondence, reinforcing trust in recovery fidelity.

At a high level, the narrative arc is: sparse distributional snapshot from time series → closed-form scaled Beta recovery (α, β) → fidelity evidenced by Random Forest classification gains and amplified by implicit regularization from constant-value features.

The empirical study centers on two datasets. The primary application is a curated SeatGeek time-series dataset collected via the SeatGeek API from May 2023 to May 2024 SEAT GEEK (SeatGeek). It contains approximately 130,000 events, 15,400 artists or acts, and 6,700 venues across the United States, spanning globally recognized performers (e.g., Metallica, Taylor Swift) as well as local acts. The dataset reflects an expanding secondary market that is projected to grow by USD 132.1 billion between 2024 and 2028 at a 34.25% CAGR (PR Newswire, 2024). SeatGeek's position in this market has been reinforced through strategic partnerships, including serving as the official secondary marketplace for Paciolan, a leading provider of ticketing solutions for college athletics and live entertainment venues (Learfield, 2023). The ability to explicitly reverse-engineer event pricing distributions not only provides practical value for market applications, but also holds methodological interest for a wide range of domains where inference from limited statistics is required. Additional experiments on the UCI ML handwritten digits dataset (Alpaydin and Kaynak, 1998) show that the implicit-regularization mechanism is not specific to ticketing data and carries over to a standard benchmark.

The remainder of the paper proceeds as follows. Section 2 reviews prior work on distribution estimation from limited statistics, quantile- and moment-based inference, time-series classification, and implicit regularization in ensembles. Section 3 describes the SeatGeek dataset, the raw time-series representation, and the transformation to distributional features. Section 4 details the scaled-Beta estimation based on composite quantile and moment matching and states the formal accuracy–fidelity bounds. Section 5 develops the implicit-regularization results (Zero-Variance Dilution Effect; Continuous Approximation via Zero-Variance Dilution) and provides supporting evidence. Section 6 concludes with implications and outlook.

2. Related work

The literature relevant to this research spans areas that directly correspond to our contributions: (1) distribution-based parameter estimation using limited statistics, (2) statistical learning theory connecting estimation accuracy with classification performance, and (3) implicit regularization in ensemble methods. We briefly review each in turn.

2.1. Distribution-based parameter estimation and classification

Estimating parameters without full distributions and classifying time series are long-standing problems across machine learning, statistics, and econometrics. We estimate scaled-Beta parameters via composite quantile and moment matching from limited summaries, connecting to quantile-based estimation, moment matching, statistical learning theory, and feature-based TSC. Classic TSC baselines such as DTW/1-NN (Berndt and Clifford, 1994; Rakthanmanon et al., 2012) and a recent survey (Middlehurst et al., 2024) frame our comparisons, including ensemble methods like TSF and CIF. Shapelets (Ye and Keogh, 2009) introduce discriminative subsequences; in our setting, subsequences aid distribution recovery that ultimately supports act classification. For Beta distributions, (Krishnamoorthy, 2016) covers traditional fitting under full samples; we extend to estimators from limited summaries, enabling event-specific modeling with minimal data.

Feature toolkits such as TSFresh (Christ et al., 2018) and Catch22 (Lubba et al., 2019) extract broad or minimal sets of interpretable statistics; our approach instead learns a bounded-support distribution with few parameters. Quantile-centric and bounded-support works further motivate this stance: Quantile Flows for GFlowNets (Zhang et al., 2024) show how quantiles can replace point estimates; Beta Diffusion (Zhou et al., 2023) highlights Beta's flexibility for range-bounded inference; QUANT (Dempster et al., 2024) attains SOTA TSC using only quantiles from dyadic intervals; LQM (Wei et al., 2024) demonstrates that limited quantiles preserve properties needed downstream; and a black-box simulation framework (Lenzi and Rue, 2025) targets parameter recovery under limited information.

Beyond quantiles, moment-matching methods inform our estimators. Moment matching accelerates diffusion sampling by aligning conditional expectations (Salimans et al., 2024) and improves denoising Gibbs sampling in energy-based models (Zhang et al., 2023). Reliability estimation for the exponentiated Pareto distribution from only lower record values Saini (2025) similarly infers parameters from severely limited statistics. In this spirit, we combine quantile- and moment-based constraints to recover scaled-Beta parameters for ticket pricing when observations are sparse.

2.2. Learning theory and estimation accuracy

(Lin, 1991) introduced the Jensen–Shannon divergence as a symmetrical, bounded measure of distributional distance, demonstrating how classification accuracy can deteriorate significantly when estimated and true distributions diverge. (Devroye et al., 1996) then provided probabilistic bounds on classification risk, directly linking distribution-estimation error to predictive accuracy. (Tsybakov, 2004) introduced margin conditions under which classification error rates converge optimally, establishing a deeper connection between parameter-estimation precision and classification performance.

Our findings follow these foundational insights: improved estimation of Beta parameters leads to more accurate classification of event types, while misestimation propagates into downstream classification error. By mapping the proposed parameter-estimation method to these theoretical frameworks, we demonstrate how precisely characterizing the underlying distribution supports robust predictive performance. Moreover, the observed classification accuracy itself provides indirect validation that the estimated distributions faithfully capture key aspects of the true underlying pricing dynamics. In this way, our scaled Beta approach echoes the broader principle in statistical learning that well-characterized data distributions are essential for achieving strong generalization, and conversely, strong generalization serves as empirical evidence of distributional fidelity.

2.3. Implicit regularization and entropy in random forests

Work on ensembles, especially Random Forests, has long emphasized implicit regularization for robust generalization. Early foundations include Jensen–Shannon divergence as a tool for measuring distributional shifts (Lin, 1991), bagging for variance reduction (Breiman, 1996), the Random Subspace Method (RSM) to limit over-reliance on any feature subset (Ho, 1998), and Extremely Randomized Trees, which further inject randomness into splits (Geurts et al., 2006).

(Breiman, 2001) formalized generalization in terms of tree "strength" and inter-tree "correlation," showing that lower correlation improves accuracy. Stability and entropy-based perspectives (Bousquet and Elisseeff, 2002) complement this view. Despite being capable of interpolation, Random Forests can generalize via ensemble self-averaging (Wyner et al., 2017). In causal forests, adaptive neighborhood selection provides implicit regularization that reduces estimation variance (Wager and Athey, 2018). Random feature selection (`max_features` or m) likewise lowers variance and acts as an implicit regularizer (Mentch and Zhou, 2020), a theme connected to budget-aware hyperparameter tuning (Cironis et al., 2022). Relatedly, sparse Bayesian learning with automatic-weighting Laplace priors shows how structural constraints induce implicit regularization (Bai and Sun, 2023).

Our contribution highlights an additional, often overlooked mechanism: introducing zero-variance features reshapes the feature-selection distribution, acting as an "entropy-based stabilizer." By reducing the probability that dominant predictors monopolize splits, this increases ensemble variety, can promote deeper or more varied trees, and reduces inter-tree correlation. In our scaled-Beta setting with limited summaries, such redistribution ensures newly introduced distributional parameters (α, β) are not overshadowed by obvious predictors and can inform split decisions.

Regularization by explicit penalties is classical, e.g., ridge regression and Tikhonov regularization (Hoerl and Kennard, 1970; Tikhonov, 1943). From a KL/entropy viewpoint, Random Forests, while not Bayesian, can still exhibit entropy-driven smoothing akin to the stability perspective of (Bousquet and Elisseeff, 2002). The probability-redistribution effect parallels function-smoothing in FDA; for instance, roughness penalties in free-knot spline estimation (De Magistris et al., 2024) avoid over-concentration and preserve balanced structure. Finally, the link we establish among implicit regularization, feature-selection probabilities, and classification accuracy resonates with recent work formulating hyperparameter optimization for randomized algorithms as a stochastic inverse problem solved via Ensemble Kalman Inversion (Dunbar et al., 2025).

3. Preliminaries and data

3.1. Event time-series data

The main use-case dataset consists of daily snapshots of secondary concert ticket prices collected through the SeatGeek API, covering the period from May 2023 to May 2024. In total it includes approximately 130,000 events, 15,400 artists or acts, and 6,700 venues across the United States. For each event, price information was recorded from the initial sale date (or first available date) through the event date, yielding a comprehensive view of the pricing lifecycle. We denote the raw time-series data as

$$\mathcal{D}_{\text{raw}} = \{\mathbf{x}_t\}_{t=1}^{T},$$

where $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(d)}]^\top$ is a vector of d observed variables at time t, and T is the total number of recorded time steps. Variables include artist, event date/time, venue, price collection date/time, mean price, median price, low price, high price, and listing count. This can equivalently be represented as a matrix:

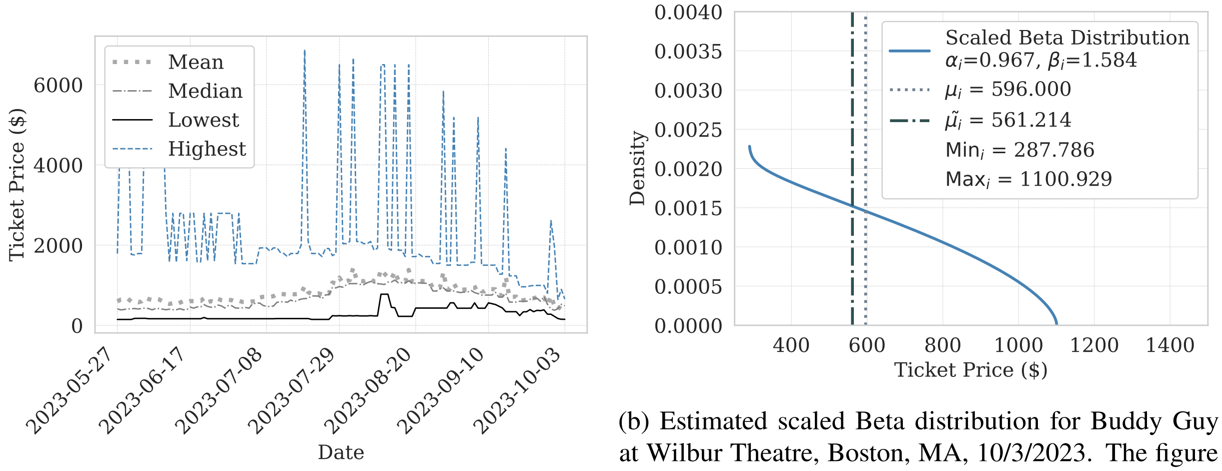
$$\mathbf{X} \in \mathbb{R}^{T \times d},$$

with rows corresponding to time steps and columns to variables. Fig. 1a illustrates this representation using ticket price data for blues guitarist Buddy Guy at the Wilbur Theatre in Boston on 10/3/2023.

3.2. Representations for classification

To prepare the data for artist classification, we define a transformation f from the raw time-series data into a structured feature space:

$$f : \mathbb{R}^{T \times d} \to \mathbb{R}^{E \times n}.$$



(a) Ticket prices over time for Buddy Guy at Wilbur Theatre, Boston, MA, 10/3/2023, showing the Mean, Median, Lowest, and Highest prices.

SEAT GEEK

(b) Estimated scaled Beta distribution for Buddy Guy at Wilbur Theatre, Boston, MA, 10/3/2023. The figure shows the estimated α_i and β_i parameters, the mean price (μ_i), the median price ($\tilde{\mu}_i$), the lowest price (Min$_i$), and the highest price (Max$_i$). These quantities define the scaled Beta distribution at snapshot T', leading up to event i on 10/3/2023. They represent the economic signature and corresponding feature value set for this event in the Random Forest model.

Figure 1: Event Overview, Buddy Guy at Wilbur Theatre, Boston, MA, 10/3/2023

The resulting training dataset is given by:

$$\mathcal{D}_{\text{train}} = f(\mathcal{D}_{\text{raw}}) = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E},$$

where $\mathbf{z}_i \in \mathbb{R}^n$ is the derived feature vector for the i-th event, and y_i is the corresponding artist label. This structured format is amenable to standard machine learning methods.

We frame artist classification as pairwise binary tasks (e.g., The Pixies vs. Billy Joel), using each model both for identification and as a test of whether artist-specific pricing distributions are separable. High accuracy signals a faithful representation of underlying distributional dynamics, while lower accuracy warrants further study. Random Forests suit this problem because they capture complex feature interactions and are robust to noise, and feature-based time-series classification shows that summary statistics can be strongly discriminative (Middlehurst et al., 2024). Toolkits like TSFresh (Christ et al., 2018) and Catch22 (Lubba et al., 2019) achieve high accuracy with handcrafted statistics, and ensemble methods such as Canonical Interval Forest (CIF) further improve performance via interval-based features (Middlehurst et al., 2024). Building on these insights, we summarize pricing time series into compact distributional representations that capture artist-specific patterns in pricing dynamics.

Basic statistical features: Initially, we derive basic summary statistics for the subsequence T' leading up to each event:

$$x = \frac{1}{|T'|} \sum_{t \in T'} x_t, \quad x_t \in \{\mu_t, \tilde{\mu}_t, \text{Max}_t, \text{Min}_t\}, \tag{1}$$

where $\mu_t, \tilde{\mu}_t, \text{Max}_t, \text{Min}_t$ denote mean, median, maximum, and minimum prices respectively. This yields the feature vector and dataset:

$$\mathbf{z}_i = [\mu_i,\ \tilde{\mu}_i,\ \text{Max}_i,\ \text{Min}_i]^\top \qquad \mathcal{D}_{\text{basic}} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E}. \tag{2}$$

Distribution-augmented features: Basic statistics alone omit nuanced distributional shapes. To address this, we estimate scaled Beta distribution parameters α_i, β_i for each event-artist pair over $[\text{Min}_i, \text{Max}_i]$ (Section

4). These parameters enrich the feature vector and yield the dataset:

$$\mathbf{z}_i = [\mu_i,\ \tilde{\mu}_i,\ \mathrm{Max}_i,\ \mathrm{Min}_i,\ \alpha_i,\ \beta_i]^\top \qquad \mathcal{D}_{\alpha\beta} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E}. \tag{3}$$

3.3. Implicit regularization via zero-variance features

To implicitly regularize our Random Forest models, we augment dataset $\mathcal{D}_{\alpha\beta}$ with n_{ZV} zero-variance (constant-value) features $\mathbf{c} \in \mathbb{R}^{n_{\mathrm{ZV}}}$: yielding the updated feature vector and dataset:

$$\mathbf{z}_i = [\mathbf{z}_i,\ \mathbf{c}]^\top \qquad \mathcal{D}_{\alpha\beta}^{(\mathrm{reg})} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E}. \tag{4}$$

Although counterintuitive, constant-value features subtly shift Random Forest feature-selection probabilities, implicitly promoting deeper, more robust trees and improved generalization, as explored in detail in Section 5.

Additional validation (handwritten digits): To verify that implicit regularization effects generalize beyond ticket pricing data, we replicate our approach using the standard UCI handwritten digits dataset (Alpaydin and Kaynak, 1998). Specifically, we form two analogous datasets: $\mathcal{D}_\delta$, containing the original digit features, and $\mathcal{D}_\delta^{(\mathrm{reg})}$, which includes additional zero-variance features to mirror the ticket pricing methodology. This parallel validation confirms (Section 5) the consistency and generalizability of the observed regularization effects across distinct data domains.

4. Distribution recovery from limited statistics

Concert ticket price distributions are modeled for each event-artist pair using a scaled Beta distribution. We estimate α_i and β_i via composite quantile and moment matching from the SeatGeek API minimum, maximum, mean, and median. Scaled Beta flexibly captures bounded shapes and provides a nuanced snapshot of pricing dynamics. Unlike (Wei et al., 2024), which matches multiple quantiles in a latent space, and (Zhang et al., 2024), which parameterizes flows via quantiles, we fit α and β using only the median and mean, an effective summary-based strategy on minimal data. (Dempster et al., 2024) also show the value of quantile features, though they require richer raw coverage than SeatGeek. Classical formulas in (Krishnamoorthy, 2016) provide the Beta moments and parameter relations we use.

Two complementary validations are presented. Sections 4.4 and 4.5 test fidelity indirectly via out-of-sample classification accuracy consistent with the accuracy fidelity theory. Section 4.6 directly verifies the theory in a controlled synthetic pipeline: generate a scaled-Beta law P_θ with $\theta = (\alpha_i, \beta_i)$, compute $\mathrm{Min}_i, \mathrm{Max}_i, \mu_i, \tilde{\mu}_i$, reconstruct $\hat{\theta} = (\alpha_i', \beta_i')$ from $\mu_i, \tilde{\mu}_i$ after standardizing by $(\mathrm{Min}_i, \mathrm{Max}_i)$, then measure relationships among logistic loss, $\mathrm{TV}(P_{\hat{\theta}}, P_\theta)$, and $\mathrm{JS}(P_{\hat{\theta}}, P_\theta)$. Loss to divergence curves fall within $\frac{1}{2}\,\mathrm{TV}^2 \leq \mathrm{JS} \leq (\ln 2)\,\mathrm{TV}$ and align with the Lipschitz margin arguments, supporting accuracy as a practical surrogate for fidelity when only $\{\mathrm{Min}_i, \mathrm{Max}_i, \mu_i, \tilde{\mu}_i\}$ are available.

4.1. Scaled beta distribution

With variables Min_i, Max_i, μ_i, and $\tilde{\mu}_i$ defined in Section 3, the probability density function (PDF) of the scaled Beta distribution is given by:

$$f(x; \alpha_i, \beta_i, \mathrm{Min}_i, \mathrm{Max}_i) = \frac{(x - \mathrm{Min}_i)^{\alpha_i - 1}\,(\mathrm{Max}_i - x)^{\beta_i - 1}}{(\mathrm{Max}_i - \mathrm{Min}_i)^{\alpha_i + \beta_i - 1}\,B(\alpha_i, \beta_i)}, \tag{5}$$

where x is the ticket price, and $B(\alpha_i, \beta_i)$ is the Beta function. This formulation transforms the standard Beta distribution from $[0, 1]$ to $[\mathrm{Min}_i, \mathrm{Max}_i]$. Such a scaled Beta framework is also seen in other contexts, like (Zhou et al., 2023), who exploit Beta distributions for bounded data in generative modeling, underscoring the flexibility of Beta-based parameterizations. Classical discussions in (Krishnamoorthy, 2016) elaborate on these Beta formulations and offer general moment-based inference approaches that set the stage for our scaled version.

4.2. Closed-form estimation: composite quantiles and moments

To estimate the parameters α_i and β_i, we reparameterize the scaled Beta distribution using the μ_i and $\tilde{\mu}_i$ provided by the SeatGeek API. The μ_i and $\tilde{\mu}_i$ for the Beta distribution on $[\text{Min}_i, \text{Max}_i]$ are given by:

$$\mu_i \;=\; \text{Min}_i \;+\; \frac{\alpha_i}{\alpha_i + \beta_i}\,(\text{Max}_i - \text{Min}_i), \tag{6}$$

$$\tilde{\mu}_i \;\approx\; \text{Min}_i \;+\; (\text{Max}_i - \text{Min}_i)\Big(\frac{\alpha_i - \frac{1}{3}}{\alpha_i + \beta_i - \frac{2}{3}}\Big). \tag{7}$$

We first scale the mean and median to $[0, 1]$:

$$s \;=\; \frac{\mu_i - \text{Min}_i}{\text{Max}_i - \text{Min}_i}, \quad q \;=\; \frac{\tilde{\mu}_i - \text{Min}_i}{\text{Max}_i - \text{Min}_i}. \tag{8}$$

From the mean equation, we express β_i in terms of α_i and s, and substitute into the median equation to obtain:

$$\beta_i \;=\; \alpha_i\Big(\frac{1-s}{s}\Big) \quad \text{and} \quad q \;=\; \frac{\alpha_i - \frac{1}{3}}{\frac{\alpha_i}{s} - \frac{2}{3}}. \tag{9}$$

This simplifies to:

$$\alpha_i \;=\; \frac{s\,(2q-1)}{3\,(q-s)}, \quad \beta_i \;=\; \frac{(\,1-s)\,(2q-1)}{3\,(q-s)}. \tag{10}$$

This method estimates the underlying price distribution from minimal statistics, capturing central tendency and shape, which improves predictive performance in downstream models. (Salimans et al., 2024) show that matching selected moments can preserve generative behavior, supporting our use of a mean plus a single quantile (the median) to influence the inferred distribution. While (Dempster et al., 2024) use many raw-data quantiles, our composite quantile-and-moment matching uses only SeatGeek API summaries. Canonical Beta identities from (Krishnamoorthy, 2016) validate fitting α and β from so few statistics.

(Wei et al., 2024) show that matching multiple quantiles in latent space can further align distributions. Our approach is simpler, fitting scaled-Beta parameters directly in the observable ticket-price space using only the mean and median, yet it demonstrates how a small, well-chosen set of statistics yields useful distributional insight. (Lubba et al., 2019) likewise find that compact feature sets can preserve classification strength, supporting our reliance on $\{\mu_i,\ \tilde{\mu}_i,\ \alpha_i,\ \beta_i\}$ with Min_i and Max_i. An example snapshot appears in Fig. 1b.

In terms of efficiency, the estimator runs in constant time $O(1)$ per event with a few arithmetic operations. Root-finding for (α, β) from mean and median typically needs $O(I)$ special-function evaluations with $I \approx 5$ to 20 iterations. Constrained optimizers incur $O(I\,C_{\text{grad}})$ from gradient computations and line search. Grid search costs $O(G)$ for a grid of size G unless supported by precomputation and interpolation. Simulation-based or Bayesian methods are more general but scale as $O(N_{\text{iter}})$ with larger constants. Our closed-form solution is the most efficient among these options and is well suited to large SeatGeek-scale applications.

4.3. Kernel density estimation for distributional features

Given the derivations for α_i and β_i alongside the original statistical features, we compare these components across events for specific acts to identify where α_i and β_i add predictive power. Larger distances between the feature distributions of two acts indicate greater separability. Consider two acts $\{1, 2\}$ in a pairwise setting.

Formally, for a given act, define

$$\mathbf{z}_i = \big[\mu_i,\ \tilde{\mu}_i,\ \text{Max}_i,\ \text{Min}_i,\ \alpha_i,\ \beta_i\big]^\top$$

for each event i. Let $x \in \{\mu, \tilde{\mu}, \text{Max}, \text{Min}, \alpha, \beta\}$. The kernel density estimate (KDE) for each feature is

$$\hat{f}_x(x) = \frac{1}{E\,h} \sum_{i=1}^{E} K\Big(\frac{x - x_i}{h}\Big), \tag{11}$$

where K is the kernel, E is the number of events, and h is the bandwidth. Using the KDE for each feature and act, $\{\hat{f}^{act}_{\mu}, \hat{f}^{act}_{\tilde{\mu}}, \hat{f}^{act}_{\text{Max}}, \hat{f}^{act}_{\text{Min}}, \hat{f}^{act}_{\alpha}, \hat{f}^{act}_{\beta}\}$, we assess distributional similarity with Hellinger distance $H(\hat{f}^1_x, \hat{f}^2_x)$ and Jensen–Shannon divergence $JS(\hat{f}^1_x \parallel \hat{f}^2_x)$:

1. Hellinger Distance:

$$H(\hat{f}^1_x, \hat{f}^2_x) = \frac{1}{\sqrt{2}} \sqrt{\int \big(\sqrt{\hat{f}^1_x(t)} - \sqrt{\hat{f}^2_x(t)}\big)^2 dt.} \tag{12}$$

2. Jensen–Shannon Divergence:

$$JS(\hat{f}^1_x \parallel \hat{f}^2_x) = \tfrac{1}{2}\, KL(\hat{f}^1_x \parallel M) + \tfrac{1}{2}\, KL(\hat{f}^2_x \parallel M), \tag{13}$$

   where $M = \frac{1}{2}(\hat{f}^1_x + \hat{f}^2_x)$ and

$$KL(\hat{f}^1_x \parallel \hat{f}^2_x) = \int \hat{f}^1_x(t)\, \log\Big(\frac{\hat{f}^1_x(t)}{\hat{f}^2_x(t)}\Big)\, dt. \tag{14}$$

These distances score each feature's ability to separate acts. The estimated parameters α_i and β_i often improve Random Forest accuracy by sharpening separability. For example, in Fig. 2 comparing Drake and Olivia Rodrigo, the KDEs show α_i is more distinctive than the original features, which is reflected in both Hellinger and JS.

While (Dempster et al., 2024) extract many quantiles from raw data, our setting uses summary statistics to compute α_i and β_i. Krishnamoorthy's discussion (Krishnamoorthy, 2016) emphasizes how Beta shape parameters capture subtle differences. Here those shape and skew measures both aid classification and are validated by it.

Our use of the Jensen–Shannon distance here differs from Section 4.4, where it supports formal bounds on convergence. In this subsection it is an empirical tool for comparing feature-density profiles across artists.

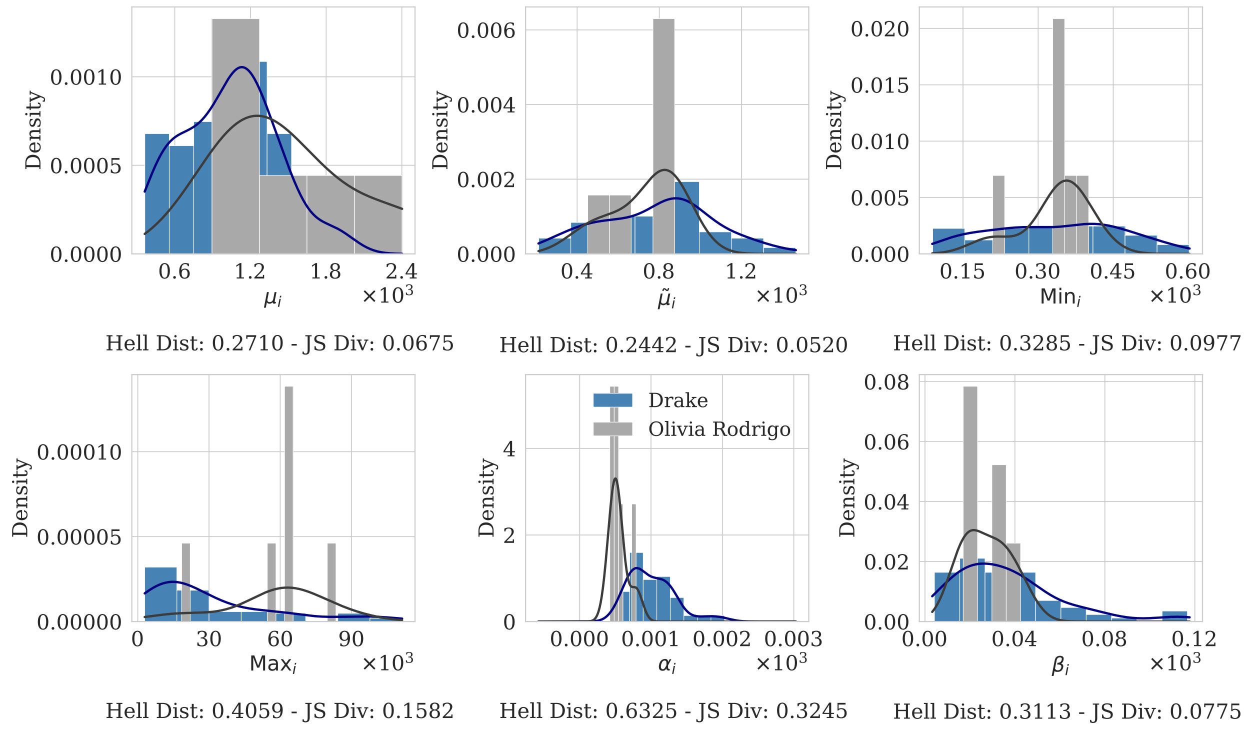
4.4. Validation via classification accuracy

We justify classification as a validation tool by linking classification accuracy to parameter estimation accuracy. If the estimated parameters $\hat{\theta}$ are close to the true parameters θ then classification performance should rise. Conversely, high classification performance provides empirical evidence that the estimates capture the underlying distribution, so accuracy can serve as a measurable proxy for distributional correctness when ground truth is unavailable. This builds on Tsybakov's margin assumption (Tsybakov, 2004) and the probabilistic bounds of Devroye et al. (Devroye et al., 1996). We extend these results by characterizing the connection between accuracy and distributional similarity through total variation distance and Jensen-Shannon divergence (Lin, 1991), showing that improvements in accuracy yield stronger guarantees of distributional reliability with quadratic convergence in the information-theoretic setting.

Proposition 4.1 (Parameter Estimation Consistency via Classification Accuracy)**.** *Let $\Theta \subset \mathbb{R}^d$ be the space of parameters, where each probability distribution P is parameterized by $\theta \in \Theta$. Define a feature map*

$$\phi(\theta) = \Big(f_1(P), f_2(P), \ldots, f_{k-d}(P), \theta\Big), \tag{15}$$

where $f_i(P)$ represents summary statistics of P, such as Min_i, Max_i, $\tilde{\mu}_i$, and μ_i. A classifier $f : \mathbb{R}^k \to \{0, 1\}$ is trained to distinguish between two classes based on $\phi(\hat{\theta})$, where $\hat{\theta}$ is an estimated parameter obtained from



Hell Dist: 0.2710 - JS Div: 0.0675

Hell Dist: 0.2442 - JS Div: 0.0520

Hell Dist: 0.3285 - JS Div: 0.0977

Hell Dist: 0.4059 - JS Div: 0.1582

Hell Dist: 0.6325 - JS Div: 0.3245

Hell Dist: 0.3113 - JS Div: 0.0775

Figure 2: The plots show the distributions of each feature across all events for artists Drake and Olivia Rodrigo. The Hellinger Distance and Jensen-Shannon divergence are calculated between each distribution. In this particular comparison of artists, the α_i parameter offers the most distinctive density profile across all events, as indicated by the distribution distance metrics.

observed data. If f achieves a classification error rate ε, then there exists a function $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$ such that the estimation error satisfies

$$\|\hat{\theta} - \theta\| \;\leq\; \delta(\varepsilon). \tag{16}$$

Proof. **Propagation of estimation error to feature space:** Define the "true" feature vector by $X^* = \phi(\theta)$ and the observed feature vector by $X = \phi(\hat{\theta})$. By the Lipschitz condition,

$$\| X - X^*\| = \|\phi(\hat{\theta}) - \phi(\theta)\| \;\leq\; L\,\|\hat{\theta} - \theta\|. \tag{17}$$

Relating feature perturbation to classification error: Under the margin separation assumption, the ideal feature vectors for two classes are separated by at least Δ. Suppose that a perturbation of size γ in feature space is tolerable without altering class assignment. Then,

$$R(f) \;\geq\; \mathbb{P}\big(\| X - X^*\| \;\geq\; \tfrac{\Delta}{2}\big). \tag{18}$$

Bounding the probability of large feature perturbation: Using Markov's inequality,

$$\mathbb{P}\big(\| X - X^*\| \;\geq\; \tfrac{\Delta}{2}\big) \;\leq\; \frac{2}{\Delta}\,\mathbb{E}\big[\| X - X^*\|\big]. \tag{19}$$

Combining with the Lipschitz bound,

$$\mathbb{P}(E) \;\leq\; \tfrac{2}{\Delta}\,L\,\|\hat{\theta} - \theta\|. \tag{20}$$

Then using the risk bound $R(f) \leq \varepsilon$, we obtain:

$$\varepsilon \;\geq\; \tfrac{2L}{\Delta}\,\|\hat{\theta} - \theta\|. \tag{21}$$

Rearranging,

$$\|\hat{\theta} - \theta\| \ \leq \ \tfrac{\Delta}{2L}\,\varepsilon. \tag{22}$$

Setting $\delta(\varepsilon) = \frac{\Delta}{2L}\,\varepsilon$, we see $\delta(\varepsilon) \longrightarrow 0$ as $\varepsilon \longrightarrow 0$, proving the proposition. □

Building on this foundation, we can more precisely characterize the relationship between classification accuracy and distributional similarity. The following theorems extend our theoretical analysis to establish rigorous bounds between classifier error and common measures of distributional difference.

Theorem 4.1 (Classification Accuracy and Total Variation Distance)**.** *Let $P_{\hat{\theta}}$ and P_{θ} denote distributions on a common sample space $\mathcal{X}$ with densities $p_{\hat{\theta}}(x)$ and $p_{\theta}(x)$, parameterized by the estimated parameters $\hat{\theta}$ and true parameters θ, respectively. Let ε be the misclassification error probability of a classifier built upon $P_{\hat{\theta}}$. Then the total variation distance between the distributions is bounded by:*

$$\begin{gathered} TV(P_{\hat{\theta}}, P_{\theta}) \ = \ \tfrac{1}{2}\int_{\mathcal{X}} \left| p_{\hat{\theta}}(x) - p_{\theta}(x) \right| dx \ \leq \ \eta(\varepsilon), \\ \textit{where} \quad \eta(\varepsilon) \to 0 \quad \textit{as} \quad \varepsilon \to 0. \end{gathered} \tag{23}$$

Proof. **Misclassification and distributional differences:** Consider a binary classifier with decision regions $C_{\hat{\theta}}$ and C_{θ} corresponding to estimated and true parameters, respectively, and assume equal class priors. The misclassification probability ε is given by:

$$\varepsilon \ = \ \frac{1}{2}\int_{\mathcal{X}} \left[\, p_{\hat{\theta}}(x)\, I(x \in C_{\theta}) \ + \ p_{\theta}(x)\, I(x \in C_{\hat{\theta}}) \,\right] dx, \tag{24}$$

where $I(\cdot)$ is the indicator function. We compare ε to the Bayes-optimal classification error ε^*, given explicitly by the total variation distance (Devroye et al., 1996):

$$\varepsilon^* \ = \ \frac{1}{2}\left[1 - TV(P_{\hat{\theta}}, P_{\theta})\right] \ = \ \frac{1}{2}\left[1 - \frac{1}{2}\int_{\mathcal{X}} \left| p_{\hat{\theta}}(x) - p_{\theta}(x) \right| dx\right]. \tag{25}$$

Since the achieved error ε must exceed the Bayes-optimal error ε^*, we have:

$$\varepsilon \ \geq \ \varepsilon^* \ = \ \frac{1}{2}\left[1 - TV(P_{\hat{\theta}}, P_{\theta})\right]. \tag{26}$$

Rearranging terms explicitly isolates the total variation distance:

$$TV(P_{\hat{\theta}}, P_{\theta}) \ \geq \ 1 - 2\varepsilon. \tag{27}$$

In particular, $TV(P_{\hat{\theta}}, P_{\theta}) \geq \max\{0,\ 1 - 2\varepsilon\}$, so the bound is vacuous only when $1 - 2\varepsilon < 0$. This provides a fundamental lower bound linking classification error and distributional differences. However, we also seek a meaningful upper bound.

Upper bound via parameter continuity: From the previous proposition, we have a direct parameter-based bound:

$$\|\hat{\theta} - \theta\| \leq \delta(\varepsilon), \quad \text{with} \quad \delta(\varepsilon) \longrightarrow 0 \quad \text{as} \quad \varepsilon \longrightarrow 0.$$

Assume the parametric family $\{P_{\theta}\}$ is Lipschitz-continuous in parameters in total variation, meaning there exists a constant $L > 0$ such that:

$$TV(P_{\hat{\theta}}, P_{\theta}) \ \leq \ \frac{L}{2}\,\|\hat{\theta} - \theta\|. \tag{28}$$

This condition typically holds for parametric distributions like the scaled Beta considered in this work, where densities vary smoothly with respect to parameters. Substituting the result from the Proposition, we get:

$$TV(P_{\hat{\theta}}, P_{\theta}) \ \leq \ \frac{L}{2}\,\delta(\varepsilon). \tag{29}$$

Define $\eta(\varepsilon) = \frac{L}{2}\,\delta(\varepsilon)$, which clearly approaches zero as $\varepsilon \to 0$. Thus, we have established a rigorous upper bound directly relating classifier error to total variation distance:

$$TV(P_{\hat{\theta}}, P_\theta) \;\le\; \eta(\varepsilon), \quad \eta(\varepsilon) \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

□

While the Total Variation distance provides a natural measure of distributional difference, information-theoretic measures can offer additional insights with stronger convergence properties. The following theorem establishes an even more precise relationship using the Jensen-Shannon divergence.

Theorem 4.2 (Classification Accuracy and Jensen–Shannon Divergence)**.** *Under the same conditions as the previous theorem, the Jensen–Shannon divergence between the distributions can be bounded by:*

$$JS(P_{\hat{\theta}} || P_\theta) \le \xi(\varepsilon), \quad \textit{where} \quad \xi(\varepsilon) \to 0 \textit{ as } \varepsilon \to 0. \tag{30}$$

Furthermore, in the small-error regime (i.e., for sufficiently small ε) and under mild regularity, this bound exhibits a quadratic convergence rate.

Proof. **Relationship between JS divergence and total variation distance:** The Jensen–Shannon divergence between distributions $P_{\hat{\theta}}$ and P_θ is defined as:

$$JS(P_{\hat{\theta}} || P_\theta) = \frac{1}{2} KL(P_{\hat{\theta}} || M) + \frac{1}{2} KL(P_\theta || M), \tag{31}$$

where $M = \frac{1}{2}(P_{\hat{\theta}} + P_\theta)$ is the mixture distribution, and KL is the Kullback–Leibler divergence. We recall (see, e.g., (Lin, 1991)) that, globally, JS is Lipschitz in total variation:

$$JS(P_{\hat{\theta}} || P_\theta) \;\le\; (\ln 2)\, TV(P_{\hat{\theta}}, P_\theta), \tag{32}$$

where $TV(P_{\hat{\theta}}, P_\theta)$ is the total variation distance as defined in the previous theorem.

Applying the total variation bound from the previous theorem: From the previous theorem, we have established that

$$TV(P_{\hat{\theta}}, P_\theta) \le \eta(\varepsilon) = \frac{L}{2}\,\delta(\varepsilon).$$

Substituting this into (32) yields the global vanishing bound

$$JS(P_{\hat{\theta}} || P_\theta) \;\le\; (\ln 2)\, \eta(\varepsilon) \;=\; (\ln 2)\, \frac{L}{2}\,\delta(\varepsilon) \;:=\; \xi(\varepsilon), \tag{33}$$

so that $\xi(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Quadratic convergence in the small-error regime: Moreover, when the discrepancy is small, JS admits a second-order (quadratic) control in TV under mild regularity (e.g., the relevant densities are bounded away from 0 and ∞, or the likelihood ratio is bounded). Thus there exist constants $C > 0$ and $\tau > 0$ (depending only on those regularity parameters) such that

$$\text{if} \quad TV(P_{\hat{\theta}}, P_\theta) \le \tau \quad \text{then} \quad JS(P_{\hat{\theta}} || P_\theta) \;\le\; C\, TV^2(P_{\hat{\theta}}, P_\theta). \tag{34}$$

Combining (34) with $TV(P_{\hat{\theta}}, P_\theta) \le \eta(\varepsilon)$ from above gives, whenever $\eta(\varepsilon) \le \tau$,

$$JS(P_{\hat{\theta}} || P_\theta) \;\le\; C\,\eta^2(\varepsilon) \;=\; C \left(\frac{L}{2}\,\delta(\varepsilon) \right)^2 \;:=\; \tilde{\xi}(\varepsilon), \tag{35}$$

which shows a quadratic convergence rate in the small-error regime since $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$. □

The progression from total variation distance to Jensen–Shannon divergence reveals a finer relationship: globally linear in error and quadratic in the small-error regime under mild regularity. For ticket pricing this means that as classification accuracy improves, i.e., as ε decreases, the estimated Beta parameters approach the true parameters at least linearly, with an accelerating quadratic rate once in the small-error regime. The Jensen–Shannon divergence offers advantages over total variation: (1) tighter small-error convergence via the quadratic relationship while retaining a global linear guarantee, (2) a natural information-theoretic view of distinguishability, (3) bounded range $[0, \log 2]$ or $[0, 1]$ in bits, (4) symmetry unlike KL.

In our ticket pricing context, modest gains in classifier accuracy produce increasingly large improvements in agreement between estimated and true Beta parameters, especially once accuracy is high. This supports parameter estimation from limited statistics and strengthens the theoretical basis for using the estimated α_i and β_i in downstream tasks. By relating classification error to both total variation and Jensen–Shannon divergence, these results connect practical machine-learning performance with rigorous statistical inference and tie learning theory to our modeling objectives.

Application to ticket pricing and artist classification: In our setting, P is a scaled Beta distribution for ticket prices with parameters $\theta = (\alpha_i, \beta_i)$. The summary statistics $f_i(P)$ are Min_i, Max_i, μ_i, and $\tilde{\mu}_i$, and the classifier f distinguishes artists using pricing information. According to the theorem, artist classification accuracy validates the estimated parameters $\hat{\alpha}_i$ and $\hat{\beta}_i$. Strong classifier performance implies that the reconstructed Beta distribution is close to the true pricing distribution. Empirical results in the next section confirm this, showing that including α_i and β_i in the feature set improves accuracy. This aligns with (Tsybakov, 2004; Devroye et al., 1996), which link precise estimation to better prediction, and with (Krishnamoorthy, 2016), which notes that accurate Beta inference can rely on a few well chosen statistics. Tight approximation of α_i and β_i produces measurable gains in classification accuracy.

4.5. Random forest results

Random Forests build multiple decision trees and aggregate their predictions to improve generalization (Ho, 1998; Breiman, 2001). For input x,

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} h_b(x), \tag{36}$$

where $h_b(x)$ is the bth tree and B is the number of trees. Trees train on bootstrap samples and use random feature subsets at each split, which reduces variance and limits overfitting relative to single trees. We use the standard scikit-learn implementation (Pedregosa et al., 2011).

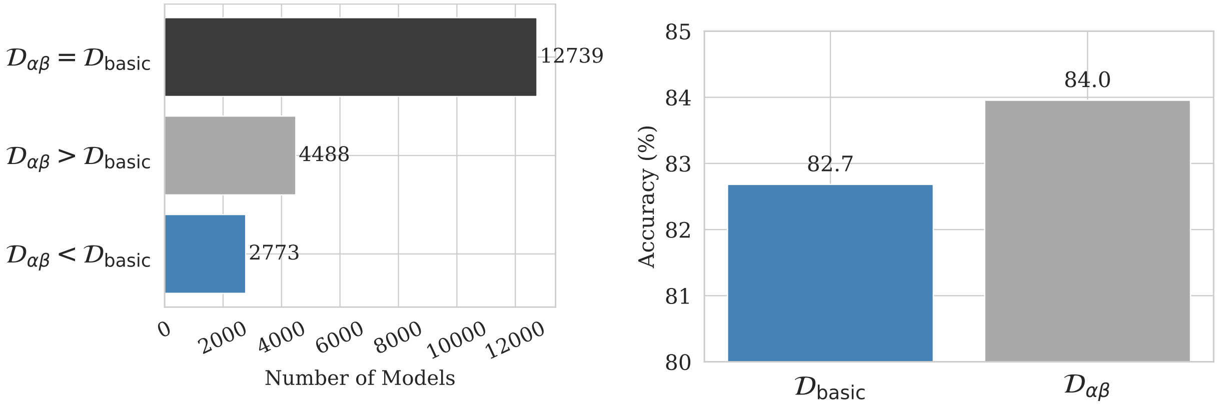
In our classification task we identify the artist from pricing information. Each Random Forest is trained on a pair of artists. This dyadic setup mitigates class imbalance compared with one versus all classification.

Empirical classification improvements with parameter estimates: To assess the effect of including α_i and β_i in the feature set, we performed a pairwise comparison of Random Forest classifiers trained on two sets. The first, $\mathcal{D}_{\text{basic}}$, contains μ_i, $\tilde{\mu}_i$, Min_i, and Max_i as in Section 3. The second, $\mathcal{D}_{\alpha\beta}$, augments these with estimated α_i and β_i from the scaled Beta distribution. The task is binary artist classification with target y_i. We compare overall accuracy and count how many models improve when using $\mathcal{D}_{\alpha\beta}$. We used $N_{\text{pair}} = 20{,}000$ paired observations of the same artist classification problem solved with both $\mathcal{D}_{\text{basic}}$ and $\mathcal{D}_{\alpha\beta}$, spanning $N_{\text{artist}} = 954$ unique artists. Each pair has $N_{\text{event}}^{\text{train}} \approx 37$ training events and $N_{\text{event}}^{\text{test}} \approx 10$ testing events under an 80/20 split, yielding $N_{\text{models}} = 20{,}000$ Random Forest models. The dataset is representative and additional subsets produced similar results. Hyperparameter specifications are available in the accompanying GitHub repository, and other configurations yielded comparable conclusions. Outcomes per pair fall into three categories: $\mathcal{D}_{\alpha\beta}$ better, equal, or worse than $\mathcal{D}_{\text{basic}}$. Of $N_{\text{pair}} = 20{,}000$ pairs we observed 12,739 ties, 4,488 better, and 2,773 worse, giving the effective sample size $N' = 7{,}261$ for statistical testing.

Under H_0 with $p = 0.5$, a normal approximation gives mean 3,630.5 and standard deviation 42.61. We use the standard binomial test via the normal approximation with a continuity correction, customary for large N' and $p = 0.5$, which yields $Z \approx 20.13$ and $p < 10^{-89}$, as summarized in Table 1. We observe a statistically

Table 1: Summary of statistical results comparing $\mathcal{D}_{\alpha\beta}$ to $\mathcal{D}_{\text{basic}}$.

Statistic	Value
Effective sample size (N')	7,261
n_{better}	4,488
n_{worse}	2,773
Mean ($\mu = N'/2$)	3,630.5
Std. dev. (σ)	42.61
Z-score	20.13
p-value	$< 10^{-89}$



(a) Random Forest model performance comparisons for $N_{\text{models}} = 20{,}000$, using the typical default feature selection size of $m = \text{round}(\sqrt{n}) = 2$. The bars show the number of cases in which models trained on $\mathcal{D}_{\alpha\beta}$ performed the same, better, or worse than models trained on $\mathcal{D}_{\text{basic}}$.

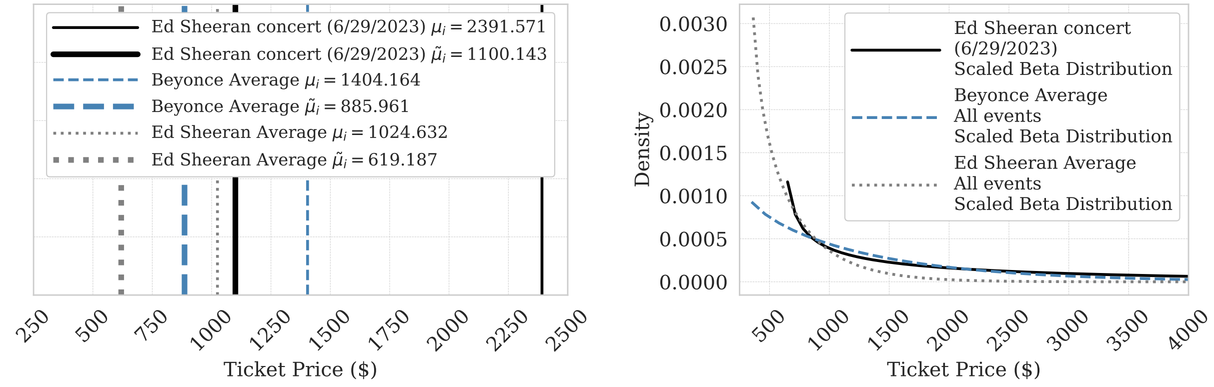
(b) Percent Accuracy by Feature Subset for $N_{\text{models}} = 20{,}000$. Although the overall accuracy difference between $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\text{basic}}$ appears relatively small, it is statistically significant given the large sample size and the substantial proportion of models showing improvement.

Figure 3: Random Forest performance comparison using $\mathcal{D}_{\text{basic}}$ vs. $\mathcal{D}_{\alpha\beta}$ features.

significant improvement using $\mathcal{D}_{\alpha\beta}$ (Fig. 3a). The mean accuracy gain across 20,000 models is modest yet consistent (Fig. 3b). The large number of improved cases together with the extremely low p-value indicates a nontrivial effect. Adding α_i and β_i improves artist discrimination and highlights the value of distributional features in dynamic pricing.

This supports the view that while (Dempster et al., 2024) exploit rich raw quantiles, limited summary statistics with principled estimation offer strong distributional characterization. Related quantile and moment findings (Zhang et al., 2024; Salimans et al., 2024) and classical Beta inference (Krishnamoorthy, 2016) align with these results, where carefully chosen statistics yield accurate parameters and improved classification.

Case study, Beyoncé vs. Ed Sheeran: To illustrate the value of incorporating estimated α_i and β_i, consider an Ed Sheeran concert on 6/29/2023 at the Boch Center Wang Theatre in Boston, MA. With only Min_i, Max_i, μ_i, and $\tilde{\mu}_i$, the summary resembles a typical Beyoncé profile and the model misclassifies the event as Beyoncé (Fig. 4a). Adding the estimated parameters yields a scaled Beta that exposes a sharper price drop characteristic of Ed Sheeran, which corrects the error (Fig. 4b). This demonstrates improved accuracy and robustness from integrating estimated distribution parameters into the Random Forest framework and it validates the inferred distribution.



(a) Ticket pricing summary for the Ed Sheeran concert on 6/29/2023 at Boch Center Wang Theatre, Boston, MA, using basic statistics (μ_i, $\tilde{\mu}_i$, Min_i, Max_i). Without distribution parameters, the mean and median prices align closely with typical Beyoncé concert values, leading to misclassification.

(b) Comparison of scaled Beta distributions after estimating α_i and β_i parameters for the Ed Sheeran concert (6/29/2023). The estimated distribution shows a more pronounced price drop relative to the typical Beyoncé concert profile, accurately reflecting Ed Sheeran's pricing pattern and correcting the previous misclassification.

Figure 4: Statistical vs. distributional pricing representations for the Ed Sheeran concert on 6/29/2023 at Boch Center Wang Theatre.

4.6. Synthetic ground-truth validation (scaled beta)

To verify the accuracy–fidelity link without relying on unknown SeatGeek densities, we construct a controlled experiment where the ground-truth law is a scaled-Beta on $[\text{Min}_i, \text{Max}_i]$ with parameters $\theta = (\alpha_i, \beta_i)$. We then mimic the estimation pipeline of Section 4 to reconstruct $\hat{\theta} = (\alpha'_i, \beta'_i)$ from the limited statistics $(\mu_i, \tilde{\mu}_i, \text{Min}_i, \text{Max}_i)$.

Loss–divergence relations: Let P_θ and $P_{\hat{\theta}}$ denote the true and reconstructed densities on $[\text{Min}_i, \text{Max}_i]$. We measure total variation and Jensen–Shannon divergence

$$\text{TV}(P_{\hat{\theta}}, P_\theta) = \tfrac{1}{2}\int_{\text{Min}_i}^{\text{Max}_i} \big|p_{\hat{\theta}}(x) - p_\theta(x)\big|\, dx, \qquad \text{JS}(P_{\hat{\theta}} \,\|\, P_\theta),$$

and, using natural logarithms (JS in nats), apply the two-sided bounds

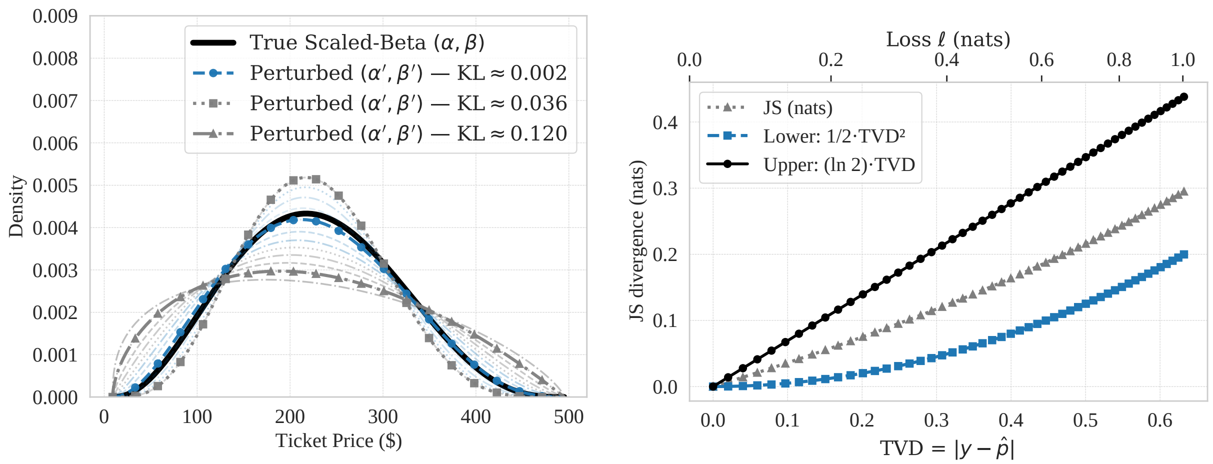
$$\tfrac{1}{2}\,\text{TV}^2(P_{\hat{\theta}}, P_\theta) \;\le\; \text{JS}(P_{\hat{\theta}} \,\|\, P_\theta) \;\le\; (\ln 2)\,\text{TV}(P_{\hat{\theta}}, P_\theta). \tag{37}$$

For a labeled example with predicted probability $\hat{p}_i$ and logistic loss $\ell_i = -\big[y_i \log \hat{p}_i + (1 - y_i)\log(1 - \hat{p}_i)\big]$, the exact identity

$$\text{TV}_i = |y_i - \hat{p}_i| \;=\; 1 - e^{-\ell_i} \tag{38}$$

implies $\ell = -\ln(1 - \text{TV})$ (small-error expansion: $\text{TV}_i = \ell_i + O(\ell_i^2)$). Composing with the bounds yields $\text{JS}_i = \frac{1}{2}\ell_i^2 + O(\ell_i^3)$ in the small-error regime, i.e., TV contracts linearly with loss, while JS contracts quadratically. These identities are model-agnostic; because our Random Forest evaluations use probabilistic cross-entropy (log loss) computed from the model's predicted class probabilities, the loss–TV/JS relationships above apply verbatim to the Random Forest results in this work.

Chain of implications: The validation proceeds along a single flow from parameters to divergence: Lipschitz continuity of the feature map and density w.r.t. (α, β) controls how parameter error propagates (Section 4.4); Theorem 4.1 (classification–TV control) links misclassification to distributional discrepancy; and JS sharpens



(a) True vs. reconstructed scaled-Beta densities (divergence-sorted). Ground-truth P_θ (black) on $[\mathrm{Min}_i, \mathrm{Max}_i]$ with (α_i, β_i) and reconstructed $P_{\hat{\theta}}$ curves (blue→grey) obtained from $(\mu_i, \tilde{\mu}_i, \mathrm{Min}_i, \mathrm{Max}_i)$. Shading reflects increasing divergence to truth and visually motivates the applied distance measures.

(b) JS vs. TV with bounds and secondary loss axis. Empirical $\mathrm{JS}(P_{\hat{\theta}} \,\|\, P_\theta)$ lies within the two-sided bounds $\frac{1}{2}\mathrm{TV}^2 \le \mathrm{JS} \le (\ln 2)\mathrm{TV}$ (nats). The top axis maps TV to logistic loss via $\ell = -\ln(1 - \mathrm{TV})$, tying divergence trends directly to classification confidence.

Figure 5: Distributional divergence analysis. (a) shows reconstructed vs. true scaled-Beta densities sorted by divergence; (b) plots Jensen–Shannon divergence against total variation with theoretical bounds and a mapped loss axis. Together these illustrate how reconstruction fidelity relates to classification-relevant divergence scales.

TV quadratically at small error in Theorem 4.2. We summarize this pipeline as

$$\underbrace{\sqrt{(\alpha' - \alpha)^2 + (\beta' - \beta)^2}}_{\text{Scaled Beta parameter error}} \xrightarrow{L} \underbrace{\varepsilon = \mathbb{P}(E)}_{\text{classification error}} \xrightarrow{\text{Theorem 4.1}} \underbrace{\mathrm{TV}(P_\theta, P_{\hat{\theta}})}_{\text{distributional distance}} \xrightarrow{\text{Theorem 4.2}} \underbrace{\mathrm{JS}(P_\theta, P_{\hat{\theta}})}_{\text{divergence fidelity}} .$$

Figure 5a anchors the construction by overlaying the true density with a divergence-sorted family of reconstructed curves, making visible how departures from (α_i, β_i) alter the shape on the observed range. Figure 5b plots empirical JS against TV with a secondary loss axis $\ell = -\ln(1 - \mathrm{TV})$, and the curve lies between the quadratic lower and linear upper bounds, confirming the globally linear and small-error quadratic regimes predicted by the theory. Together, these figures provide a compact, direct verification that out-of-sample accuracy is a reliable surrogate for distributional fidelity when only $(\mu_i, \tilde{\mu}_i, \mathrm{Min}_i, \mathrm{Max}_i)$ are available.

5. Implicit regularization via zero-variance features

Regularization improves generalization by limiting overfitting to noisy or irrelevant features. In Random Forests, bagging (Breiman, 1996) and random subspace sampling (Ho, 1998) yield implicit regularization and robustness. Bootstrapped trees reduce variance, and random feature choice decorrelates trees and controls complexity. (Breiman, 2001) decomposed generalization error into tree strength and inter-tree correlation, providing a theoretical basis.

$$\mathbb{E}[(f(x) - \mathbb{E}[y \mid x])^2] = \mathrm{Var}(f(x)) + \mathrm{Bias}^2(f(x)) + \sigma^2, \tag{39}$$

where $\mathbb{E}[y \mid x]$ is the true conditional class probability, $\mathrm{Var}(f(x))$ the model variance, $\mathrm{Bias}^2(f(x))$ the squared bias, and σ^2 irreducible noise. "Extremely Randomized Trees" increase decorrelation and lower variance (Geurts et al., 2006), while stability results (Bousquet and Elisseeff, 2002) support varied ensembles. We study another implicit mechanism by adding zero-variance (constant) features to reshape feature selection and decision boundaries. In ticket pricing, this serves as a lever to tune trust in recovered distributions, since adjustments to the Random Forest ensemble directly modulate the fidelity of the inferred parameters.

5.1. Analysis of probabilistic feature selection

Notation: For notational clarity in what follows, we let $m = \texttt{max\_features}$ denote the number of features randomly selected at each split in the Random Forest, following the scikit-learn (Pedregosa et al., 2011) implementation.

In a standard Random Forest construction with fixed-size feature selection, exactly m out of the n total features are chosen at each split node, uniformly over all $\binom{n}{m}$ subsets. Hence, the probability that a particular feature X_j is included in the candidate set at any node is

$$P(\text{Include } X_j) \;=\; \frac{m}{n}. \tag{40}$$

Across B trees, each containing an average of L split nodes, the expected total number of times X_j appears in candidate sets is then

$$\mathbb{E}[\text{Count}_{\text{in-candidate}}(X_j)] \;=\; B \cdot L \cdot \frac{m}{n}. \tag{41}$$

Feature selection via gini impurity reduction: For binary classification, the Gini impurity is

$$G \;=\; 2\,p\,(1-p), \tag{42}$$

where p is the proportion of one class. Splitting on X_j changes this impurity, reducing it by $\Delta G(X_j)$. We define the rank or score of feature X_j as

$$r(X_j) \;=\; \Delta G(X_j). \tag{43}$$

A higher $r(X_j)$ means a larger impurity reduction and thus a higher rank among the available features at a node. Moderately predictive parameters, such as α_i and β_i in the artist classification use-case, can still achieve some positive $r(X_j)$, even if not as large as top-ranked features.

Competitive advantage of highly ranked features: Although each feature X_j has a nominal $\frac{m}{n}$ chance of appearing in the size-m candidate set at a node, the final split is awarded to whichever feature yields the greatest score. If we assume a proportional "weighted by $r(X_j)$" selection among the m chosen, then for a subset

$$S \;\subseteq\; \{1,\ldots,n\}, \quad |S| \;=\; m, \tag{44}$$

we have

$$P(S) \;=\; \frac{1}{\binom{n}{m}}. \tag{45}$$

Conditioned on S, the probability that X_j wins the split is

$$P(X_j \mid S) \;=\; \frac{r(X_j)}{\sum_{k\in S} r(X_k)}. \tag{46}$$

Hence, the unconditional probability of X_j being chosen for a split is

$$P(X_j) \;=\; \sum_{S:\, j\in S} P(S) \;\cdot\; P(X_j \mid S), \tag{47}$$

which expands to

$$P(X_j) \;=\; \frac{1}{\binom{n}{m}} \sum_{S:\, j \in S} \frac{r(X_j)}{\sum_{k \in S} r(X_k)}. \tag{48}$$

Closed-form approximation: When $n \gg m$, or simply for conceptual ease, one can approximate $\sum_{k \in S} r(X_k)$ by its expectation,

$$r(X_j) \;+\; (m-1)\, \mathbb{E}\big[\, r(X_k)\big], \tag{49}$$

yielding

$$P(X_j) \;\approx \tag{50}$$

$$\frac{1}{\binom{n}{m}} \sum_{S:\, j \in S} \frac{r(X_j)}{r(X_j) + (m-1)\, \mathbb{E}\big[\, r(X_k)\big].} \tag{51}$$

Since there are $\binom{n-1}{m-1}$ subsets that include X_j, and $\frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n}$, we obtain

$$P(X_j) \;\approx\; \frac{m}{n} \cdot \frac{r(X_j)}{r(X_j) + (m-1)\, \mathbb{E}\big[\, r(X_k)\big].} \tag{52}$$

Thus, even though every feature has the same nominal $\frac{m}{n}$ chance of entering the candidate set, those with consistently higher $r(X_j)$ can dominate, overshadowing less ranked predictors.

Probabilistic effects of zero-variance variables: Earlier (Section 3), the datasets $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\delta}$ were extended to $\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})}$ and $\mathcal{D}_{\delta}^{(\mathrm{reg})}$ by adding zero-variance (constant-value) features. The count of such features is n_{ZV}. They have near-zero Gini scores because they yield no impurity reduction. Although seemingly unhelpful, these variables shift the Random Forest selection dynamics by adding low-score competition that tempers dominance of top-ranked features. Let n denote the non-constant features (e.g., $\mu_i, \tilde{\mu}_i, \mathrm{Max}_i, \mathrm{Min}_i, \alpha_i, \beta_i$) and n_{ZV} the zero-variance ones, so the feature set has $n + n_{\mathrm{ZV}}$ elements. If each zero-variance feature has rank score $r_{\mathrm{ZV}} \approx 0$, it dilutes the score sum in the denominator and boosts the selection probability of mid-ranked features compared to the case with no zero-variance features.

Theorem 5.1 (Zero-Variance Dilution Effect)**.** *Suppose n_{ZV} zero-variance features with $r_{\mathrm{ZV}} \approx 0$ are added, enlarging the feature set from n to $n_{\mathrm{eff}} = n + n_{\mathrm{ZV}}$. Let*

$$\bar{r}_{\mathrm{eff}} = \frac{\sum_{j=1}^{n} r(X_j) + n_{\mathrm{ZV}}\, r_{\mathrm{ZV}}}{n_{\mathrm{eff}}} \;\approx\; \frac{n\,\bar{r}}{n + n_{\mathrm{ZV}}} \tag{53}$$

$$\textit{so that} \quad \bar{r}_{\mathrm{eff}} < \bar{r}. \tag{54}$$

For any two informative features X_h, X_ℓ with scores $a = r(X_h) > b = r(X_\ell) > r_{\mathrm{ZV}}$, the closed-form odds ratio between their selection probabilities satisfies

$$\frac{P_h^{(\mathrm{eff})}(m)}{P_\ell^{(\mathrm{eff})}(m)} \;<\; \frac{P_h(m)}{P_\ell(m)}, \tag{55}$$

where

$$P_j(m) = \frac{m}{n}\, \frac{r(X_j)}{r(X_j) + (m-1)\,\bar{r}}, \tag{56}$$

$$P_j^{(\mathrm{eff})}(m) = \frac{m}{n_{\mathrm{eff}}}\, \frac{r(X_j)}{r(X_j) + (m-1)\,\bar{r}_{\mathrm{eff}}}. \tag{57}$$

Thus adding zero-variance features compresses the relative dominance of higher-scoring over lower-scoring variables, giving mid-ranked features more splitting opportunities.

Proof. Write $K = (m-1)\,\overline{r}$ and $K_t = (m-1)\,\overline{r}_{\text{eff}}$ with $K_t < K$. The prefactors $\frac{m}{n}$ and $\frac{m}{n_{\text{eff}}}$ cancel in the ratio, giving

$$\frac{P_h^{(\text{eff})}(m)}{P_\ell^{(\text{eff})}(m)} = \frac{a}{b}\,\frac{b+K_t}{a+K_t}, \quad \frac{P_h(m)}{P_\ell(m)} = \frac{a}{b}\,\frac{b+K}{a+K}. \tag{58}$$

Define

$$R(K) = \frac{a}{b}\frac{b+K}{a+K}. \tag{59}$$

A direct derivative gives

$$\frac{dR}{dK} = \frac{a(a-b)}{b\,(a+K)^2} > 0 \tag{60}$$

because $a > b > 0$; hence $R(K)$ is strictly increasing in K. Since $K_t < K$, we have

$$R(K_t) < R(K), \tag{61}$$

establishing the claimed inequality. □

(Geurts et al., 2006) demonstrated that increasing the randomization of split selection in Extremely Randomized Trees leads to deeper decision trees by weakening the dependence of split choices on the target variable. This increased depth arises because random splits reduce the impurity reduction at each node, thus requiring additional splits to achieve sufficient purity. Formally, this can be expressed as: $\mathbb{E}[d_{\text{random}}] > \mathbb{E}[d_{\text{optimal}}]$, where $\mathbb{E}[d_{\text{random}}]$ and $\mathbb{E}[d_{\text{optimal}}]$ represent the expected tree depths for randomized and optimal splits, respectively. Our approach and experiments reveal comparable effects, with zero-variance features increasing tree depth and encouraging more variation among splits.

Corollary 5.1 (Increased Expected Tree Depth)**.** *Consider a Random Forest whose effective feature set is $n_{\text{eff}} = n + n_{\text{ZV}}$, with n informative features ($r(X_j) > 0$) and n_{ZV} zero-variance features ($r_{\text{ZV}} \approx 0$). Let d denote the depth of a decision tree grown under a fixed impurity-based stopping rule. Then, holding all other training hyperparameters constant,*

$$\mathbb{E}\big[d(n_{\text{eff}})\big] \; > \; \mathbb{E}\big[d(n)\big]. \tag{62}$$

Proof. Theorem 5.1 shows that adding zero-variance features compresses the odds of high- versus mid-ranked variables:

$$\frac{P_h^{(\text{eff})}(m)}{P_\ell^{(\text{eff})}(m)} \; < \; \frac{P_h(m)}{P_\ell(m)}. \tag{63}$$

Consequently, top-scoring features win fewer splits relative to before, and more mid-ranked features are selected. Because those mid-ranked features achieve smaller impurity reductions ($r_\ell < b < r_h$), the expected impurity drop per internal node is lower. A lower per-split reduction means the chosen impurity threshold is reached later in the recursive partitioning process, so additional levels are needed before termination. Hence the expected depth increases: $\mathbb{E}[d(n_{\text{eff}})] > \mathbb{E}[d(n)]$. □

Increased randomness in split selection explicitly reduces the correlation among trees (Geurts et al., 2006), expressed mathematically as:

$$\sigma^2 = \rho\,\frac{\operatorname{Var}(h(x))}{B}, \tag{64}$$

where lower correlation ρ directly reduces ensemble variance σ^2. Our theoretical analysis and empirical results confirm this assertion.

Corollary 5.2 (Reduced Ensemble Correlation)**.** *Let ρ denote the pair-wise correlation between base learners in a Random Forest. Adding n_{ZV} zero-variance features lowers that correlation and hence the variance term $\rho\operatorname{Var}(h)/B$ in Breiman's bias–variance decomposition.*

Proof. Without zero-variance features, the highest-ranked variables win a large fraction of candidate splits; many trees therefore grow similar decision paths, inflating ρ. Theorem 5.1 shows that after augmentation the odds ratio $P_h^{(\mathrm{eff})}(m)/P_\ell^{(\mathrm{eff})}(m)$ shrinks for every pair of scores $a > b > 0$. Consequently, top-ranked variables win fewer splits relative to mid-ranked ones, and different features now have a greater chance of initiating branches. This increased heterogeneity of split choices makes the predictions of individual trees less correlated, so ρ decreases; the factor $\rho \operatorname{Var}(h)/B$ is therefore reduced. □

5.2. Expanding the regularization search space

Prior work by (Mentch and Zhou, 2020) shows that tuning the m (`max_features`) hyperparameter, the number of features considered at each split, regularizes Random Forests by altering the chance that a feature is selected. Because m is an integer, their scheme moves in discrete steps and yields a finite set of selection probabilities. Our approach adds constant-value features, which changes the total feature count and yields a near continuum of expected feature-selection probabilities. The theorem below formalizes that introducing n_{ZV} constant features can approximate any target probability in a broad interval by randomized interpolation between adjacent choices of n_{ZV}, thereby "filling in the gaps" left by discrete m adjustments.

Theorem 5.2 (Continuous Approximation via Zero-Variance Dilution)**.** *Let n denote the number of truly informative features, and fix an integer m such that $1 \le m \le n$. Let n_{ZV} be the number of constant (zero-variance) features added. In the absence of constant features, the effective probability of selecting an informative feature at a split is*

$$\gamma = \frac{m}{n}. \tag{65}$$

If we add $n_{\mathrm{ZV}} \ge 0$ constant (zero-variance) features, then the total number of features is $n + n_{\mathrm{ZV}}$, and the effective selection probability of an informative feature becomes

$$\gamma' = \frac{m}{n + n_{\mathrm{ZV}}}. \tag{66}$$

The set

$$S_L = \left\{ \tfrac{m}{n+n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0 \right\} \tag{67}$$

is a countable, monotone grid spanning $\left(0, \frac{m}{n}\right]$ with 0 as its only accumulation point. Moreover, for any desired probability

$$0 < \gamma^* \le \frac{m}{n}, \tag{68}$$

and any $\epsilon > 0$, there exist adjacent integers $k, k+1$ and a mixing weight $p \in [0, 1]$ such that the randomized scheme that uses $n_{\mathrm{ZV}} = k$ with probability p and $n_{\mathrm{ZV}} = k + 1$ with probability $1 - p$ achieves

$$\left| \mathbb{E}[\gamma'] - \gamma^* \right| < \epsilon, \tag{69}$$

where $\mathbb{E}[\gamma'] = p \, \frac{m}{n+k} + (1 - p) \, \frac{m}{n+k+1}$. Hence, by randomized interpolation between adjacent grid points, the expected selection probability can be tuned arbitrarily finely over $\left(0, \frac{m}{n}\right]$.

Proof. **Discrete set without constant features:** Following (Mentch and Zhou, 2020), let n be the number of informative features and let m be the chosen subset size at each split. The probability that any one informative feature appears in a candidate set is then $\gamma = \frac{m}{n}$. Because m must be an integer with $1 \le m \le n$, the set of possible probabilities (as m varies) is

$$S_{MZ} = \left\{ \tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{n}{n} \right\}. \tag{70}$$

This set is finite and discrete.

Zero-variance dilution and the grid S_L: Fix m. Instead of varying m itself, we add n_{ZV} constant (zero-variance) features to the existing n informative ones, so the total feature count is $n + n_{\mathrm{ZV}}$. As a result, the effective probability of picking an informative feature becomes

$$\gamma' = \frac{m}{n + n_{\mathrm{ZV}}}.$$

Hence, each nonnegative integer n_{ZV} in $\{0, 1, 2, \ldots\}$ produces one element of the set

$$S_L = \left\{\tfrac{m}{n+n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0\right\}.$$

Because n_{ZV} can grow arbitrarily large, the values of γ' can get arbitrarily close to 0. Also, when $n_{\mathrm{ZV}} = 0$, $\gamma' = \frac{m}{n} = \gamma$. Thus, S_L spans probabilities in $(0, \frac{m}{n}]$ and forms a countable, monotone grid with 0 as its only accumulation point.

Randomized interpolation to approximate any target γ^*: Take any target probability γ^* satisfying $0 < \gamma^* \leq \frac{m}{n}$ and any $\epsilon > 0$. Choose $k = \left\lfloor \frac{m}{\gamma^*} - n \right\rfloor \vee 0$. Then $\frac{m}{n+k} \geq \gamma^* \geq \frac{m}{n+k+1}$, so γ^* lies between the adjacent grid points. Define

$$p = \frac{\gamma^* - \frac{m}{n+k+1}}{\frac{m}{n+k} - \frac{m}{n+k+1}} \in [0, 1]. \tag{71}$$

If we select $n_{\mathrm{ZV}} = k$ with probability p and $n_{\mathrm{ZV}} = k + 1$ with probability $1 - p$, then

$$\mathbb{E}[\gamma'] = p \, \tfrac{m}{n+k} + (1 - p) \, \tfrac{m}{n+k+1} = \gamma^*, \tag{72}$$

and thus $\left|\mathbb{E}[\gamma'] - \gamma^*\right| = 0 < \epsilon$. When γ^* is very small, increasing k makes the adjacent grid spacing arbitrarily fine; the same interpolation then achieves any prescribed $\epsilon > 0$. This establishes the claimed approximation. □

Corollary 5.3 (Continuous Accuracy Expansion via Selection Probability)**.** *Let*

$$\gamma' = \frac{m}{n + n_{\mathrm{ZV}}}$$

be the effective probability of selecting an informative feature when the original n features are augmented with n_{ZV} constant (zero-variance) features. Assume that the mapping from γ' to the classifier's accuracy ν is continuous, and let $(\nu_{\min}, \nu_{\max}]$ denote the interval of achievable accuracy values under the original discrete scheme. Then, for any target accuracy ν^* satisfying

$$\nu_{\min} < \nu^* \leq \nu_{\max}, \tag{73}$$

and for any $\epsilon > 0$, there exist adjacent integers $k, k+1$ and a mixing weight $p \in [0, 1]$ such that the randomized scheme using $n_{\mathrm{ZV}} \in \{k, k+1\}$ with probabilities $(p, 1 - p)$ yields an expected accuracy $\mathbb{E}[\nu']$ (as a function of $\mathbb{E}[\gamma']$) satisfying

$$\left|\mathbb{E}[\nu'] - \nu^*\right| < \epsilon. \tag{74}$$

In other words, the set of expected achievable accuracies is dense in $(\nu_{\min}, \nu_{\max}]$, providing near-continuous control over the model's performance by fine-tuning the effective selection probability.

Adjusting m influences a Random Forest's effective complexity (Mentch and Zhou, 2020), yet the integer nature of m limits how finely one can tune selection probabilities. Expanding the feature set from n to $n + n_{\mathrm{ZV}}$ with zero-variance features yields γ' values for any integer $n_{\mathrm{ZV}} \geq 0$. These values form a monotone grid on $\left(0, \frac{m}{n}\right]$ and can be randomly interpolated to match any target in expectation, giving near continuous control over the effective regularization level. This can mimic or surpass the effect of a small m while refining probabilities beyond what integer steps allow. Related continuous approximations from discrete spaces appear in hyperparameter tuning, where refined grids approximate continuous optimization (Cironis et al., 2022), broadening the applicability of discrete-choice methods in machine learning.

5.3. Connections to penalty methods, functional data analysis, and related areas

Our approach connects fundamentally to classic regularization methods, such as ridge regression (Hoerl and Kennard, 1970; Tikhonov, 1943), where explicit quadratic penalties emerge naturally from Gaussian priors:

$$x_{MAP} = \arg\min_x \{\|Ax - b\|^2 + \lambda \|x\|^2\}. \tag{75}$$

Extending beyond foundational work, the feature probability reweighting structurally parallels recent advancements in regularization across various domains. For instance, in functional data analysis (FDA), recent methods such as roughness penalization in free-knot spline estimation (De Magistris et al., 2024) redistribute information to avoid over-concentration on specific knots, maintaining balanced representations. Similarly, our implicit regularization dynamically adjusts feature selection probabilities, preventing dominance by specific features:

$$p_i' = \frac{p_i}{1 + \lambda \sum_j p_j}, \tag{76}$$

This formulation resembles penalized optimization used in FDA:

$$C = \arg\min_C \|Y - \Phi^T C\|^2 + \lambda C^T R C, \tag{77}$$

which explicitly penalizes abrupt variations to enforce smoothness.

Furthermore, our implicit feature-selection regularization is related to another penalty-based approach, the inverse-problem hyperparameter optimization framework introduced by (Dunbar et al., 2025), whose formulation includes a log-determinant regularization:

$$L_M^{(EKI)}(u) = \|\Gamma(u)^{-1/2}(z - G(u))\|^2 - 2 \log P(u). \tag{78}$$

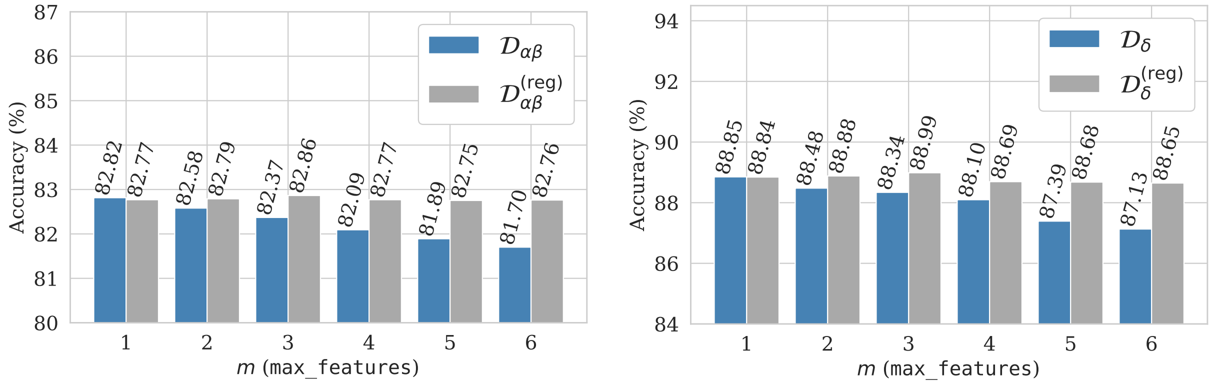
The interplay between implicit and explicit regularization frameworks presents an intriguing intersection of theoretical and applied perspectives.

5.4. Analogous effects in ticket pricing and digit classification

The same phenomena appear in both (i) the ticket pricing dataset with distribution-summary features and (ii) handwritten digit classification. In these settings, highly ranked features (e.g., strong distributional predictors or informative pixel locations) dominate splits and suppress weaker but useful signals. Introducing zero-variance variables reduces this dominance, increasing the selection frequency of subtle features and enriching the model. These variables function as implicit regularization similar to setting $m = 1$: they lower the effective weight of very high-ranked features, raise the probability of choosing secondary ones, and increase splitting variety across the ensemble, which can improve performance.

5.5. Experimental results

This section applies the same pairwise Random Forest classification methodology from the $\mathcal{D}_{\alpha\beta}$ experiments to test zero-variance features as an implicit regularizer in two domains: a new concert ticket pricing dataset and the UCI handwritten digits dataset (Alpaydin and Kaynak, 1998). We compare $\mathcal{D}_{\alpha\beta}$ to $\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})}$ on 5,000 artist-pair models spanning 954 artists, each trained and tested on an 80/20 split with ≈ 40 training events and ≈ 10 test events per model. We also examine $\mathcal{D}_\delta$ versus $\mathcal{D}_\delta^{(\mathrm{reg})}$ on 90 digit-pair classifiers across 10 digits with a similar 80/20 split providing ≈ 287 training and ≈ 72 test comparisons per pair. Regularized models use $n_{\mathrm{ZV}} = 20$. For $\mathcal{D}_\delta$ we select a random subset of $n = 6$ representative features for consistency with $\mathcal{D}_{\alpha\beta}$. These complementary experiments show the consistently beneficial effect of implicit regularization via zero-variance features in both real-world secondary ticket pricing and a classic benchmark dataset.



(a) Percent accuracy across feature selection size (m) iterations for artist event training datasets $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})}$. The impact of constant features is evident, with $\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})}$ showing declining accuracy as m increases. The highest accuracy is achieved using zero-variance feature regularization, exceeding what standard hyperparameter tuning alone can reach.

(b) Percent accuracy across feature selection size (m) iterations for handwritten digit training datasets $\mathcal{D}_{\delta}$ and $\mathcal{D}_{\delta}^{(\mathrm{reg})}$. The impact of zero-variance features is again evident, with $\mathcal{D}_{\delta}^{(\mathrm{reg})}$ showing declining accuracy as m increases. Zero-variance feature regularization achieves the highest accuracy, unattainable via standard tuning alone.

Figure 6: Accuracy trends across feature selection sizes (m) for artist and digit datasets, highlighting the implicit regularization effects of constant-value features.

Accuracy and the selection size, m: Figs. 6a and 6b show accuracy trends for the concert pricing and handwritten digit datasets as m varies. Iterating m in a standard Random Forest alters the chance of selecting informative features, yet our experiments find peak accuracy only with implicit regularization. Theorem 5.2 shows that adding constant (zero-variance) features expands tuning from discrete m steps to a near continuum of selection probabilities, and corollary 5.3 confirms that this continuous space enables fine-grained performance adjustments. Selection probabilities reach extremes at the boundaries of m (e.g., $m = 1$ yields uniform selection, while $m = n$ favors highly ranked features), and our empirical results show that continuous tuning via implicit regularization adds flexibility. By incorporating constant features, the effective average rank of candidates is diluted, balancing dominant and less prominent predictors and achieving accuracies difficult to reach by iterating m alone.

Scope of model improvements: We analyze improvements at $m = 6$, the setting with the largest discrepancy, for both ticket pricing and digit classification. Figs. 7a and 7b show statistically significant gains from zero-variance regularization. For ticket pricing, $\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})}$ adds zero-variance features while $\mathcal{D}_{\alpha\beta}$ includes α and β only. A paired comparison gives $n_{\mathrm{better}} = 1084$, $n_{\mathrm{worse}} = 675$, and $N' = 1759$. Under H_0 with $p = 0.5$, a binomial sign test with the normal approximation and continuity correction yields $Z \approx 9.72$ and $p < 10^{-21}$. Fig. 7a confirms this improvement. For digit classification, $\mathcal{D}_{\delta}^{(\mathrm{reg})}$ includes zero-variance features while $\mathcal{D}_{\delta}$ excludes them. The paired counts are $n_{\mathrm{better}} = 52$, $n_{\mathrm{worse}} = 14$, and $N' = 66$. The same test gives $Z \approx 4.56$ and $p < 10^{-5}$, as highlighted in Fig. 7b. Table 2 summarizes the statistical significance for both datasets.

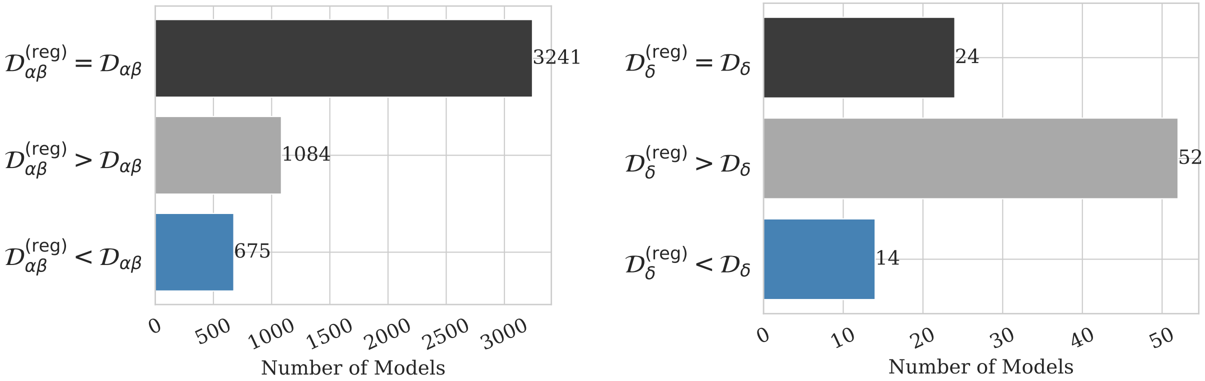
Feature re-ranking and usage in the models: In both $\mathcal{D}_{\alpha\beta}$ vs. $\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})}$ and $\mathcal{D}_{\delta}$ vs. $\mathcal{D}_{\delta}^{(\mathrm{reg})}$, introducing zero-variance features modifies the selection probabilities in the approximate formula

$$P(X_j) \;\approx\; \frac{m}{n_{\mathrm{eff}}} \cdot \frac{r(X_j)}{r(X_j) + (m-1)\,\bar{r},} \tag{79}$$

where $n_{\mathrm{eff}} = n + n_{\mathrm{ZV}}$ when constant features are present and equals n otherwise. By injecting low-scoring features into the ensemble, these datasets effectively dilute the dominance of highly ranked predictors X_j.

Table 2: Summary of statistical results for improvements with $m = 6$.

Statistic	**Tickets**	**Digits**
Effective Sample Size (N')	1759	66
n_{better}	1084	52
n_{worse}	675	14
Mean ($\mu = N'/2$)	879.5	33
Std. Dev. (σ)	20.98	4.06
Z-score	9.72	4.56
p-value	$< 10^{-21}$	$< 10^{-5}$



(a) Performance comparison for $m = 6$ across $N_{\text{models}} =$ 5,000 artist classification models. Bars indicate how often $\mathcal{D}^{(\text{reg})}_{\alpha\beta}$ performed the same, better, or worse than $\mathcal{D}_{\alpha\beta}$ (see Fig. 6a).

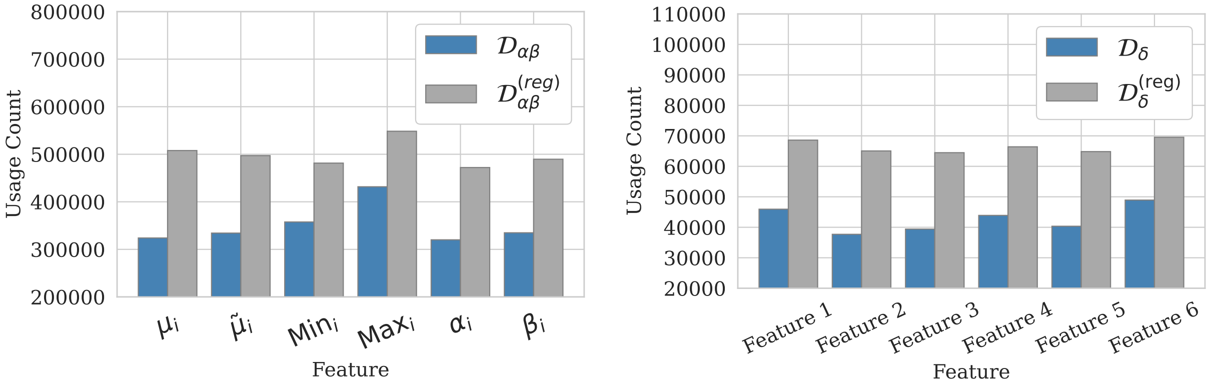
(b) Performance comparison for $m = 6$ across $N_{\text{models}} =$ 90 digit classification models. Bars indicate how often $\mathcal{D}^{(\text{reg})}_{\delta}$ performed the same, better, or worse than $\mathcal{D}_{\delta}$ (see Fig. 6b).

Figure 7: Effect of constant-value feature regularization at $m = 6$, showing the distribution of model performance changes for artist and digit classification tasks.

Consequently, the final usage distribution, aggregated across all base learners, becomes more balanced, giving subtle but informative features more opportunities at split nodes. This re-ranking serves as a form of implicit regularization, stabilizing the Random Forest. Figs. 8a and 8b show the redistribution of feature usage, highlighting the increased prominence of moderately-ranked predictors. Notably, the shifts shown in these figures underscore how implicit regularization effectively promotes model robustness through enhanced feature breadth.

Increased tree depth with implicit regularization: We further examine zero-variance features by computing the average tree depth across all models. Models with zero-variance features ($\mathcal{D}^{(\text{reg})}_{\delta}$ and $\mathcal{D}^{(\text{reg})}_{\alpha\beta}$) grow deeper trees than non-regularized counterparts, consistent with corollary 5.1 and supporting the view that implicit regularization encourages the use of more nuanced feature representations. The results validate our expected-depth analysis and extend (Geurts et al., 2006), where randomized split selection increases depth through reduced impurity gains. Figs. 9a and 9b show the depth increases. For ticket pricing, the median depth rose from 3.0 to 4.0 and the average from 3.18 to 4.16. For handwritten digits, the median rose from 8.0 to 10.0 and the average from 8.35 to 10.42. These consistent and significant increases highlight a stabilizing effect of implicit regularization across distinct datasets.

Tree variety as measured by feature count distance: We quantify the ensemble "variety" by examining the pairwise Euclidean distance between trees' feature usage vectors, $\mathbf{v}_i \in \mathbb{R}^d$. Defining the distance between



(a) Average feature usage counts across $N_{\text{models}} = 5{,}000$ for $\mathcal{D}_{\alpha\beta}$ versus $\mathcal{D}^{(\text{reg})}_{\alpha\beta}$.

(b) Average feature usage counts across $N_{\text{models}} = 90$ for $\mathcal{D}_{\delta}$ versus $\mathcal{D}^{(\text{reg})}_{\delta}$.

Figure 8: Comparison of average feature usage between unregularized and regularized models, across both artist and digit datasets. Regularization via zero-variance features leads to more varied and balanced feature selection.

trees i and j as $\|\mathbf{v}_i - \mathbf{v}_j\|_2$, we compute the sum over all pairs:

$$V(m) = \sum_{1 \le i < j \le n} \|\mathbf{v}_i - \mathbf{v}_j\|_2. \tag{80}$$

This measure is computed per model for both regularized and non-regularized datasets (Figs. 10a and 10b). A higher average $V(m)$ indicates more varied feature usage among trees, consistent with the effect of zero-variance features. The figures show that zero-variance features increase variety, reduce correlation among ensemble members, and stabilize models across scenarios. For ticket pricing, the median variety increased from 2.00 to 2.83 and the average from 2.26 to 2.99. For handwritten digits, the median increased from 4.90 to 7.35 and the average from 5.27 to 7.67.

(Geurts et al., 2006) show that the fully randomized split selection in Extra-Trees reduces correlation among trees, which we quantify via the average cosine similarity of their normalized feature usage vectors. In particular, for any two trees with vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ (with $\|\mathbf{v}_i\|_2 = \|\mathbf{v}_j\|_2 = 1$), we have

$$\mathbf{v}_i^{\top}\mathbf{v}_j = 1 - \frac{1}{2}\|\mathbf{v}_i - \mathbf{v}_j\|_2^2. \tag{81}$$

Defining the average correlation p as

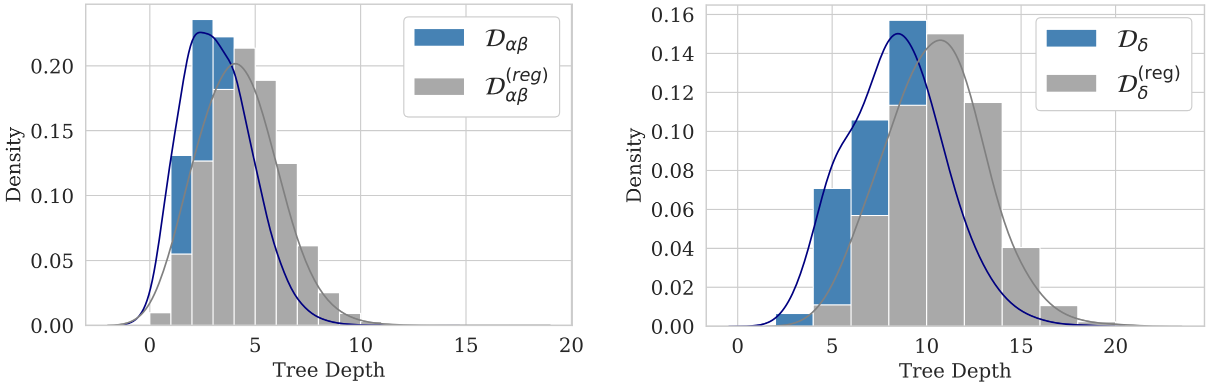
$$p = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \mathbf{v}_i^{\top}\mathbf{v}_j, \tag{82}$$

we obtain

$$p = 1 - \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2. \tag{83}$$

Thus, as our variety measure $V(m)$ increases, the average correlation p decreases, demonstrating that greater tree variety leads to reduced inter-tree correlation. This mathematical relationship aligns with our analysis, showing that implicit regularization via zero-variance features promotes a more decorrelated ensemble.

Error decomposition and implicit regularization effects: Recalling the bias-variance decomposition, zero-variance features dilute top-ranked predictors and increase tree variety, which modifies ensemble



(a) Average tree depth for $N_{\text{models}} = 5{,}000$ models: $\mathcal{D}_{\alpha\beta}$ vs. $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$. Median depth increased from 3.0 to 4.0, and average depth from 3.18 to 4.16.

(b) Average tree depth for $N_{\text{models}} = 90$ models: $\mathcal{D}_{\delta}$ vs. $\mathcal{D}_{\delta}^{(\text{reg})}$. Median depth increased from 8.0 to 10.0, and average depth from 8.35 to 10.42.

Figure 9: Effect of zero-variance feature regularization on tree depth. Both artist and digit models trained with regularized datasets grow deeper trees on average, suggesting increased robustness and feature utilization.

variance. If ρ is the correlation among tree predictions, then

$$\operatorname{Var}(f(x; n_{\text{ZV}})) = \frac{\rho(n_{\text{eff}}) \operatorname{Var}(h(x; n_{\text{ZV}}))}{B}. \tag{84}$$

This aligns with (Geurts et al., 2006), who show that added randomness lowers ensemble correlation. Zero-variance features decrease $\rho(n_{\text{eff}})$ (Corollary 5.2), thus reducing variance. Bias may rise slightly as moderate features are used more, yet our experiments show a net generalization gain.

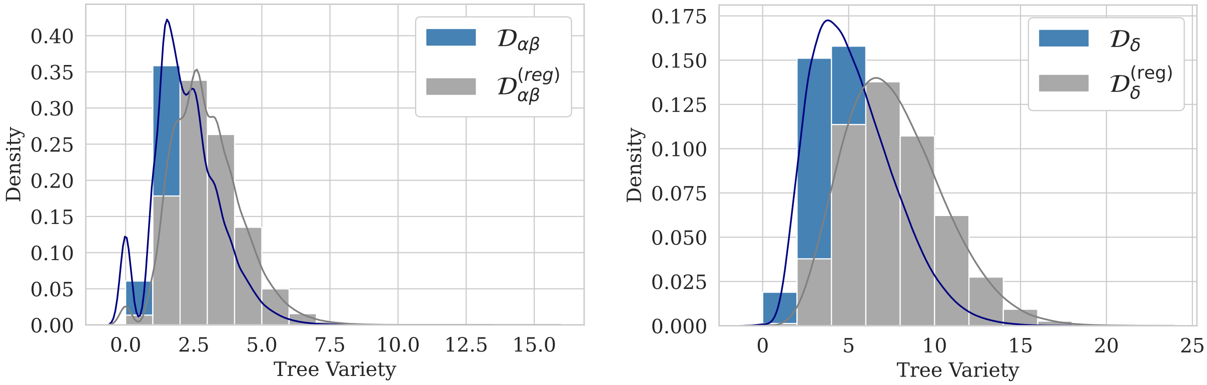
The regularization view is reinforced by (Wyner et al., 2017), who attribute AdaBoost's success to averaging interpolating classifiers that yield "spiked-smooth" boundaries, indicating implicit regularization through averaging rather than explicit penalties. This perspective is consistent with the correlation-controlled variance expression above and with classic variance-shrinkage from averaging; see (Wyner et al., 2017).

Trust in recovered distributions: In the SeatGeek ticket-pricing use case, the experimental results show that implicit regularization not only improves classification accuracy but also expands the effective search space of the Random Forest ensemble. This dual effect enhances trust in the recovered distributions, since accuracy gains more directly validate fidelity to the underlying pricing dynamics. At the same time, varying the number of zero-variance features provides an additional knob for tuning that trust, giving fine-grained control over the balance between distributional fidelity and ensemble robustness.

Case study, Dropkick Murphys vs. the Avett Brothers: A Dropkick Murphys concert on 10/28/2023 is initially mislabelled by the baseline Random Forest as an Avett Brothers show. In the unregularized model the location statistics μ_i, $\tilde{\mu}_i$, Min_i, and Max_i hold the highest empirical ranks and dominate split lotteries across trees (Fig. 12a, blue bars). The artists' feature distributions overlap substantially (Fig. 11). Driven by location parameters, the ensemble overlooks the visual match between the event price density (black curve) and the Dropkick template (blue dashed) in Fig. 12b.

Section 5 adds n_{ZV} zero-variance columns, diluting the pool from n to $n + n_{\text{ZV}}$ and lowering expected ranks of genuine predictors. Theorem 5.1 (Zero-Variance Dilution Effect) shows this increases sampling of mid-ranked features, especially β_i. Empirically β_i gains split frequency while location statistics decline (Fig. 12a, grey). The amplified shape signal flips a majority of trees and the ensemble classifies the concert correctly.

Previously correct predictions remain. Regularization preserves the decision boundary and informative location cues, gives shape parameters more opportunities, and trims tree correlation for a modest bias variance



(a) Average tree variety for $N_{\text{models}} = 5{,}000$ models: $\mathcal{D}_{\alpha\beta}$ vs. $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$. Median variety increased from 2.00 to 2.83, and average variety from 2.26 to 2.99.

(b) Average tree variety for $N_{\text{models}} = 90$ models: $\mathcal{D}_{\delta}$ vs. $\mathcal{D}_{\delta}^{(\text{reg})}$. Median variety increased from 4.90 to 7.35, and average variety from 5.27 to 7.67.

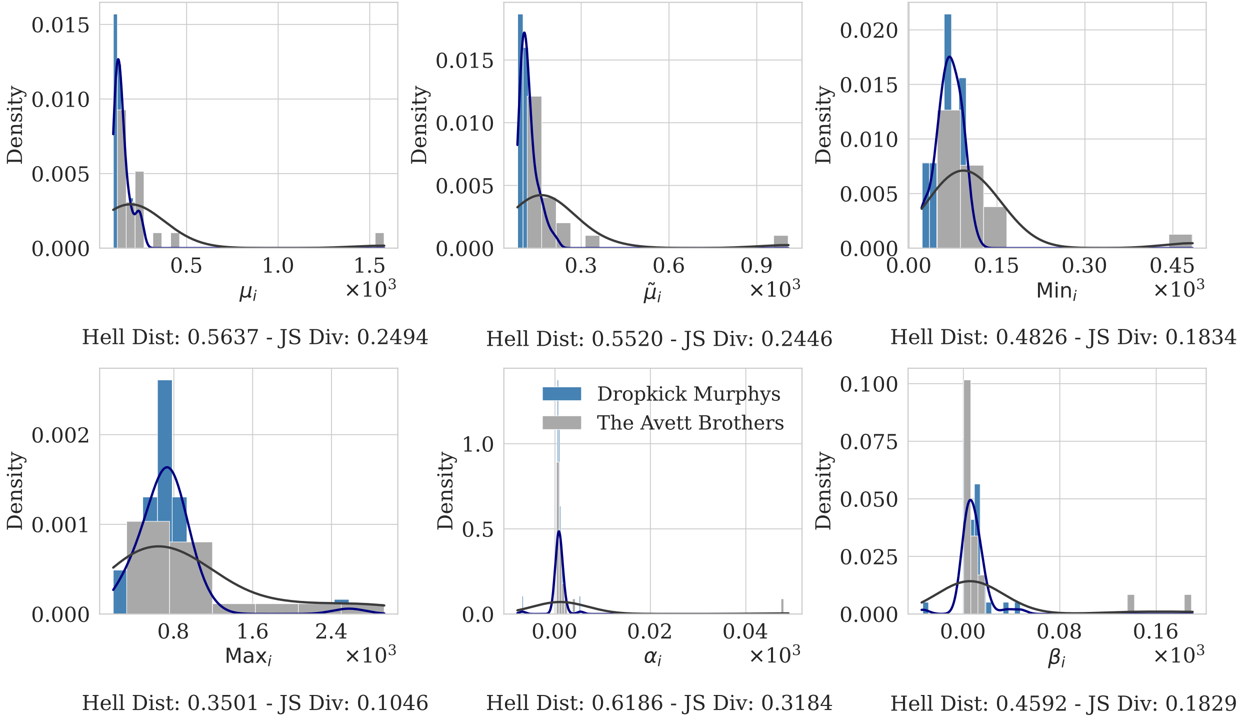
Figure 10: Effect of zero-variance feature regularization on Random Forest tree variety. Regularization increases both the median and average distances of tree structures in the ensemble, improving generalization capacity.

gain. The effect requires both scaled-Beta parameter estimation and zero-variance regularization. Section 5 anticipates this and Figs. 6a and 6b show how regularization unlocks a near continuous space of feature weightings beyond the baseline and standard hyperparameter search.

6. Conclusions

We conclude by restating and reinforcing the three established contributions.

1. **Closed-form distribution recovery from limited statistics.** We developed a composite quantile-and-moment matching estimator that reconstructs scaled Beta distributions from the minimum, maximum, mean, and median $(\text{Min}_i, \text{Max}_i, \mu_i, \tilde{\mu}_i)$, yielding parameters (α_i, β_i) that capture shape beyond location summaries. The approach builds on classical Beta results (Krishnamoorthy, 2016) and connects to recent work on quantile and moment based estimation under information constraints (Zhang et al., 2024; Dempster et al., 2024; Wei et al., 2024). In our application setting, these statistics were retrieved through the SeatGeek API (SeatGeek), making the estimator practical at scale. Relatedly, black-box parameter estimation methods (Lenzi and Rue, 2025) and reliability inference from record values (Saini, 2025) highlight the broader importance of recovering distributions from compressed or incomplete data. Unlike iterative solvers, the proposed estimator achieves distributional recovery in a single analytical step, emphasizing both scalability and computational economy.
2. **Accuracy–fidelity theory.** We established a link from predictive accuracy to distributional fidelity, using Total Variation Distance and Jensen–Shannon divergence (Lin, 1991; Devroye et al., 1996; Tsybakov, 2004). The analysis shows that improvements in artist classification accuracy correspond to increasingly precise estimates of the underlying scaled Beta parameters, and that convergence of the information-theoretic discrepancy is quadratic in the accuracy margin. This provides stability guarantees for estimation in sparse and noisy environments, a regime that is typical for market snapshots.
3. **Implicit regularization via zero-variance features.** We showed that augmenting Random Forests with zero-variance (constant-value) features can serve as an implicit regularizer that reduces the dominance of highly ranked variables, encourages variety, and deepens trees. The effect is consistent with the literature on bagging, random subspaces, and randomized trees (Breiman, 1996, 2001; Ho, 1998; Geurts



Hell Dist: 0.5637 - JS Div: 0.2494

Hell Dist: 0.5520 - JS Div: 0.2446

Hell Dist: 0.4826 - JS Div: 0.1834

Hell Dist: 0.3501 - JS Div: 0.1046

Hell Dist: 0.6186 - JS Div: 0.3184

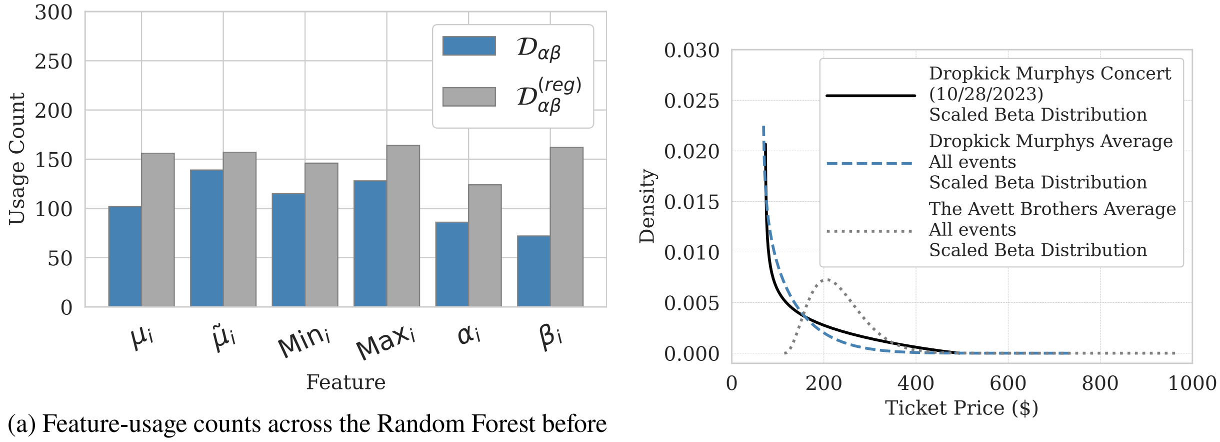
Hell Dist: 0.4592 - JS Div: 0.1829

Figure 11: The plots show the distributions of each feature across all events for artists Dropkick Murphys and The Avett Brothers. In each case, there is considerable overlap, making classification as determined by location statistics alone difficult. With a satisfactory re-ranking of feature importances by the zero-variance implicit regularization mechanism, the Random Forest model can classify more effectively based on the estimated distributional shape as shown explicitly in Fig. 12.

et al., 2006), and resonates with stability and ensemble correlation perspectives (Bousquet and Elisseeff, 2002; Wyner et al., 2017; Wager and Athey, 2018). It complements hyperparameter oriented strategies (Mentch and Zhou, 2020; Dunbar et al., 2025) by shifting split-selection probabilities through small structural changes to the feature space. Comparable to prior work on sparsity-inducing priors (Bai and Sun, 2023), our contribution emphasizes how structural constraints can implicitly regularize complex ensembles, tuning trust in distributional adherence.

Applied impact and the SeatGeek dataset: The secondary ticket resale market provides a natural testbed for distributional reconstruction and feature based classification. Platforms such as SeatGeek, StubHub, and Ticketmaster surface highly dynamic signals at event and artist resolution. Working with the curated SeatGeek dataset and daily snapshots, we transformed subsequences of pricing activity into distributional snapshots and then into learned features that expose artist-specific economic signatures. In the time-series classification literature, targeted feature representations have proven effective for high dimensional problems (Middlehurst et al., 2024; Lubba et al., 2019; Christ et al., 2018). Our results contribute a distribution based feature pathway that is tractable from minimal statistics and operational in data settings where full empirical distributions are not retained (SeatGeek).

Theoretical guarantees and empirical validation: The accuracy–fidelity bridge formalized here grounds distribution recovery in risk bounds and information measures (Lin, 1991; Devroye et al., 1996; Tsybakov, 2004). Within this framework, the modest but consistent classification gains from injecting (α_i, β_i) into the feature set have outsized meaning for distributional integrity due to the quadratic Jensen–Shannon rate. In practice, this allows classifier accuracy to serve as an operational stand-in for distributional ground truth when only summary statistics are available. The same perspective clarifies the role of constant-value features



(a) Feature-usage counts across the Random Forest before (blue) and after (grey) the zero-variance regularization. Location statistics (Min_i, Max_i, μ_i, $\tilde{\mu}_i$) remain frequent but lose relative dominance, while the shape parameter β_i increases its appearances in split decisions significantly, an explicit illustration of the Zero-Variance Dilution Effect that corrects the misclassification and leaves earlier correct calls intact.

(b) Scaled-Beta ticket-price densities: the fitted distribution for the 10/28/2023 concert (solid) versus the mean Dropkick Murphys profile (dashed) and The Avett Brothers profile (dotted). The signature matches the Dropkick template, but the signal is muted when the model relies chiefly on statistics such as Min_i, Max_i, μ_i, $\tilde{\mu}_i$.

Figure 12: Illustration of how summary statistics can mislead model classification when the true distribution shape is more informative. Dropkick Murphys and Avett Brothers show overlapping location descriptors despite nuanced and differing density shapes. Regularization shifts the focus of the Random Forest to shape.

as a mechanism that rebalances split selection, lowers correlation across trees, and improves generalization for ensembles (Breiman, 1996, 2001; Ho, 1998; Geurts et al., 2006; Bousquet and Elisseeff, 2002; Wyner et al., 2017; Wager and Athey, 2018; Mentch and Zhou, 2020; Dunbar et al., 2025). Empirically, we validated the approach on a newly curated SeatGeek pricing dataset and on the UCI handwritten digits benchmark (Alpaydin and Kaynak, 1998), confirming generality beyond ticket pricing and showing that the implicit regularization effect is not domain restricted.

Outlook: The arc demonstrated here is as follows: sparse distributional snapshots of time series → closed-form scaled Beta recovery (α, β) → Random Forest accuracy gains, with fidelity, and therefore trust, amplified by implicit regularization from zero-variance features. It is broadly applicable when dynamic systems are observed through compressed summaries. For live-market analytics, healthcare operations, demand forecasting, and energy systems, the same constraints on data access recur. The methods presented are simple to instrument, amenable to scale, and compatible with existing ensemble workflows. In settings where data arrives as aggregated snapshots rather than full samples, this narrative offers a principled route to reconstruct informative distributions, improve classification, and support decision making with clear theoretical guarantees.

Acknowledgment

Code and data availability: Derived statistical features, methodology code, and figure generation code are available at the following GitHub repository for reproducibility:
`github.com/jonland82/seatgeek-beta-modeling`

Methods: The author used standard computational tools and programming libraries, including Python packages and a large language model (OpenAI), to assist with code snippets, algebraic manipulation, and editorial refinement. All content and interpretations were conceptualized, reviewed and finalized by the author.

Declarations: The author used publicly available event data accessed via the SeatGeek API (SeatGeek, Inc.) in accordance with SeatGeek's API Terms of Use. SeatGeek is not affiliated with this research and does not endorse it. All trademarks and content remain the property of their respective owners. Proper attribution is provided at seatgeek.com as required. Raw API data is not redistributed per licensing requirements.

The author reports no conflicts of interest. No funding was received for this research.

References

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits [dataset], 1998. URL https://doi.org/10.24432/C50P49. Accessed: February 11, 2025.

Zonglong Bai and Jinwei Sun. Sparse bayesian learning with automatic-weighting laplace priors for sparse signal recovery. *Computational Statistics*, 38:2053–2074, 2023. doi: 10.1007/s00180-023-01354-4.

Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pages 359–370, 1994.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. doi: 10.1162/153244302760200704.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi: 10.1023/A:1018054314350.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018. doi: 10.1016/j.neucom.2018.03.067.

Lukas Cironis, Jan Palczewski, and Georgios Aivaliotis. Automatic model training under restrictive time constraints. *Statistics and Computing*, 33(1), 2022. doi: 10.1007/s11222-022-10166-3.

A. De Magistris, V. De Simone, E. Romano, and G. Toraldo. Roughness regularization for functional data analysis with free knots spline estimation. *Statistics and Computing*, 34(5), 2024. doi: 10.1007/s11222-024-10474-w. URL https://doi.org/10.1007/s11222-024-10474-w.

A. Dempster, D. F. Schmidt, and G. I. Webb. quant: a minimalist interval method for time series classification. *Data Mining and Knowledge Discovery*, 38:2377–2402, 2024. doi: 10.1007/s10618-024-01036-9.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. 1996. ISBN 978-1-4612-6877-2. doi: 10.1007/978-1-4612-0711-5.

Oliver R. A. Dunbar, Nicholas H. Nelsen, and Maya Mutic. Hyperparameter optimization for randomized algorithms: a case study on random features. *Statistics and Computing*, 35(3), 2025. ISSN 0960-3174. doi: 10.1007/s11222-025-10587-w.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1): 3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. doi: 10.1109/34.709601.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. 2nd edition, 2016. doi: 10.1201/b19191. URL https://doi.org/10.1201/b19191.

Learfield. Seatgeek partners with paciolan, the largest ticketing company in college athletics, February 2023. URL https://www.learfield.com/2023/02/seatgeek-partners-with-paciolan-the-largest-ticketing-company-in-college-athletics/. Accessed: February 11, 2025.

Amanda Lenzi and Håvard Rue. Towards black-box parameter estimation. *Computational Statistics*, 2025. doi: 10.1007/s00180-025-01623-4.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. *Data Mining and Knowledge Discovery*, 33(6):1821–1852, 2019. ISSN 1384-5810. doi: 10.1007/s10618-019-00647-x.

Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020. URL http://jmlr.org/papers/v21/19-905.html.

Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38: 1958–2031, 2024. doi: 10.1007/s10618-024-01022-1.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

PR Newswire. Secondary tickets market size is set to grow by usd 132.1 billion from 2024-2028 – rising popularity of sports events to boost the revenue. November 2024. URL https://www.prnewswire.com/news-releases/secondary-tickets-market-size-is-set-to-grow-by-usd-132-1-billion-from-2024-2028--rising-popularity-of-sports-events-to-boost-the-revenue--technavio-302315097.html. Accessed: February 11, 2025.

Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 262–270, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314626. doi: 10.1145/2339530.2339576.

Shubham Saini. Advancements in reliability estimation for the exponentiated pareto distribution: A comparison of classical and bayesian methods with lower record values. *Computational Statistics*, 40: 353–382, 2025. doi: 10.1007/s00180-024-01497-y.

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=C62d2nS3KO.

SeatGeek. SeatGeek API Documentation. https://platform.seatgeek.com. Accessed: 2025-07-02.

Andrey N. Tikhonov. On the stability of inverse problems. *Proceedings of the USSR Academy of Sciences*, 39:195–198, 1943. URL https://api.semanticscholar.org/CorpusID:202866372.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32 (1):135–166, 2004. doi: 10.1214/aos/1079120131.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Wei Wei, Tom De Schepper, and Kevin Mets. Dataset condensation with latent quantile matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7703–7712, 2024. doi: 10.1109/CVPRW63382.2024.00766.

Abraham J. Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017. URL `http://www.jmlr.org/papers/volume18/15-240/15-240.pdf`.

Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 947–956, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557122.

Dinghuai Zhang, Ling Pan, Ricky T. Q. Chen, Aaron Courville, and Yoshua Bengio. Distributional GFlownets with quantile flows. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=vFSsRYGpjW`.

Mingtian Zhang, Alex Hawkins-Hooker, Brooks Paige, and David Barber. Moment matching denoising gibbs sampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=NWrN6cMG2x`.

Mingyuan Zhou, Tianqi Chen, Zhendong Wang, and Huangjie Zheng. Beta diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=zTSlm4nmlH`.