Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Shrinkage-Gating Zero-Inflated Tweedie Boosting with Exact Degeneracy: Boundary Selection, Profile Likelihood Estimation, and Actuarial Calibration

Qinzheng Wang

qz.wang.phd@gmail.com

University of International Business and Economics

Yuantao Xie

xieyuantao@uibe.edu.cn

University of International Business and Economics

Research Article

Keywords: Tweedie distribution; zero inflation; shrinkage gating; boundary selection; gradient boosting; insurance pricing

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11618493/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Shrinkage-Gating Zero-Inflated Tweedie Boosting with Exact Degeneracy: Boundary Selection, Profile Likelihood Estimation, and Actuarial Calibration

Qinzheng Wang* Yuantao Xie†

Abstract

The Tweedie compound Poisson–Gamma distribution possesses a natural zero-generating mechanism. Introducing a structural zero component into insurance claim data may therefore improve zero-probability calibration, yet it may also mistakenly attribute dispersion misspecification or sampling variability to excess zero inflation. The standard logit parameterization suffers from a neglected technical impediment: the absence of zero inflation corresponds to the logit predictor tending to negative infinity, and conventional smooth penalties only drive the structural zero probability toward zero without recovering the standard Tweedie model at any finite parameter value. This paper proposes a shrinkage-gating zero-inflated Tweedie (SG-ZIT) boosting model, specified as $\omega_i = 1 - \exp(-\tau r_i)$, where $\tau \geq 0$ denotes the global gating intensity, and the positive function r_i characterizes covariate heterogeneity and is normalized to unit mean on the training sample. The boundary case $\tau = 0$ corresponds exactly to the standard Tweedie model at a finite parameter point. We fix the Tweedie component and the gating shape on an independent data fold, and then estimate the gating intensity via profile likelihood with a $\lambda\tau$ penalty. For fixed nuisance parameters, the profile objective reduces to a one-dimensional convex function. We derive the first and second derivatives for zero and positive observations, the necessary and sufficient KKT conditions for exact degeneracy, the monotonicity of the shrinkage path, a characterization of the non-convexity of the joint likelihood, as well as finite-sample bounds on the false-selection probability and asymptotic consistency of boundary selection under the null hypothesis $\tau = 0$. The algorithm updates the Tweedie component using posterior membership weights on the training fold, estimates τ along each path on a calibration fold, selects the penalty parameter via the one-standard-error rule on a selection fold, and evaluates the test fold exactly once. Simulation studies examine scenarios with and without

*School of Insurance, University of International Business and Economics, Beijing, China.
†School of Insurance, University of International Business and Economics, Beijing, China.

genuine structural zeros using independent replicates. The real-data application uses the publicly available freMTPL2 dataset, a French motor third-party liability portfolio. The results demonstrate that SG-ZIT reduces the Brier score and log loss for zero probabilities when structural zeros are present. In simulations without structural zeros, the selected gating mechanism yields $\tau = 0$ exactly, rather than merely producing numerically negligible zero-inflation probabilities. The freMTPL2 results illustrate that an extremely high raw zero proportion does not automatically constitute evidence of additional zero inflation. The primary contribution of this paper is to provide a transparent and verifiable complexity path: the extra zero component is retained only when supported by independent out-of-sample scoring; otherwise, the model degenerates strictly to the standard Tweedie.

Keywords: Tweedie distribution; zero inflation; shrinkage gating; boundary selection; gradient boosting; insurance pricing

1 Introduction

When the power parameter satisfies $p \in (1, 2)$, the Tweedie distribution admits a compound Poisson–Gamma representation and therefore simultaneously accommodates a point mass at zero and a continuous positive component. This property has made it a standard tool for non-life pure premium modeling [10, 21]. Gradient boosting methods preserve the Tweedie loss while capturing nonlinear risk structures, further extending the model's applicability in ratemaking [2, 5, 21]. In the broader context of actuarial pricing, Henckaerts et al. [9] systematically compared several machine learning methods for auto insurance pricing and emphasized the importance of model interpretability and calibration diagnostics.

Zero-inflation models posit that some policies reside in a latent state exempt from the usual claim process [11]. This assumption has a clear interpretation in scenarios such as vehicle decommissioning, inactive coverage, or business-rule exclusions of risk, but it does not follow automatically from an excess of zero observations in ordinary cross-sectional claim data. The zero probability of the standard Tweedie is jointly determined by the mean, exposure, dispersion, and power parameter; the structural zero component likewise inflates the observed zero probability. When both are estimated via highly flexible functions on the same data, the structural zero component may absorb dispersion misspecification or finite-sample noise. Existing research has estimated zero-inflated Tweedie models via EM, GEM, or specialized boosting losses [7, 16, 17, 22], with emphasis primarily on how to fit the extra component rather than on how to make it *vanish strictly* when evidence is insufficient.

The usual specification is $\omega_i = \text{logit}^{-1}\{g(\boldsymbol{x}_i)\}$. The standard Tweedie requires $\omega_i = 0$, yet this can be attained only in the limit $g(\boldsymbol{x}_i) \to -\infty$. Consequently, a ridge penalty on g does not shrink toward the standard Tweedie; a smooth penalty on ω can push

the probability close to zero but cannot produce an exact boundary solution at a finite parameter value. Interpreting a probability close to zero as "having selected the no-zero-inflation model" conflates numerical truncation with statistical selection. In the shrinkage estimation literature, Tibshirani [18]'s Lasso achieves coefficient shrinkage and variable selection via an L_1 penalty; Fan and Li [4] proved that non-concave penalized likelihood can simultaneously achieve selection consistency and asymptotic normality; and Zou [23]'s adaptive Lasso attains the Oracle property under weaker conditions through differentiated penalty weights. These theoretical results indicate that for "no zero inflation" to become an exactly attainable boundary solution, the parameterization itself must change, rather than relying solely on the asymptotic behavior of numerical penalties.

Nuisance functions at mixture model boundaries may also be unidentified. McCullagh and Nelder [12] discussed regularity conditions for parameter identifiability within the generalized linear model framework, and van der Vaart [19] further provided the argmax theorem and consistency conditions for M-estimation. A naive transposition of the ordinary Lasso Oracle narrative is not valid. This paper addresses the issue at the level of parameterization. We write the individual structural zero probability as

$$\omega_i(\tau) = 1 - \exp(-\tau r_i), \qquad \tau \geq 0, \quad r_i > 0. \tag{1}$$

Here r_i gives the relative structural zero intensity, and τ is the control parameter for the entire extra component. We fix the scale by imposing $n_{\text{tr}}^{-1} \sum r_i = 1$ on the training sample. Under this specification, $\tau = 0$ corresponds exactly to $\omega_i = 0$ at a finite point; when $\tau > 0$, r_i allows the zero inflation to vary with risk characteristics.

The paper's contributions are organized at three mutually reinforcing levels. At the methodological level, we construct a profile loss with a $\lambda\tau$ penalty, unifying the complex model and the standard Tweedie on a single continuous path. At the theoretical level, we do not claim global convexity for the joint tree model nor assert unproven Oracle properties; rather, after fixing the Tweedie component and r via an independent training fold, we establish one-dimensional convexity of the gate parameter objective, the finite-boundary KKT condition, path monotonicity, a diagnosis of joint-likelihood non-convexity, a finite-sample bound on the false-selection probability under $\tau = 0$, and asymptotic boundary-selection consistency. At the experimental design level, we explicitly separate the learning of nuisance functions, the estimation of the gate intensity, the selection of the penalty parameter, and the final evaluation, thereby avoiding the use of the final test set for model selection. In the null-hypothesis simulation scenario, the exact degeneracy frequency is reported directly; the real-data exercise tests whether a high zero proportion still supports the extra component out of sample.

2 Model Specification

2.1 Tweedie Component and Observational Distribution

Let Z_i denote the total claim amount of policy i under exposure $e_i > 0$, and let $Y_i = Z_i/e_i$ denote the claim per unit exposure. Given covariates $\boldsymbol{x}_i$, under the usual claim state we have

$$Y_i \mid \boldsymbol{x}_i, \text{usual state} \sim \mathrm{Tw}\{\mu_i, \phi/e_i, p\}, \qquad \mu_i = \exp\{f(\boldsymbol{x}_i)\}, \quad 1 < p < 2. \tag{2}$$

The zero probability of this component is

$$P_{0i} = \exp\left\{-\frac{e_i \mu_i^{2-p}}{\phi(2-p)}\right\}. \tag{3}$$

Introducing the structural zero gating mechanism of equation (1), the observed zero probability and the pure premium mean become

$$q_i(\tau) = 1 - (1 - P_{0i}) \exp(-\tau r_i), \tag{4}$$

$$\mathbb{E}(Z_i \mid \boldsymbol{x}_i) = e_i \exp(-\tau r_i)\mu_i. \tag{5}$$

The model is therefore internally consistent at both the distributional and mean levels, rather than multiplying the fitted unconditional mean by $(1 - \omega_i)$ ex post.

2.2 Endogenous Fragility of Identifiability

Equation (3) shows that P_{0i} is increasing in the dispersion parameter ϕ. Increasing ϕ and increasing τ (hence increasing ω_i) have the same directional effect on the observed zero probability q_i. If the modeler enables the structural zero component without adequate calibration of ϕ, the component will be forced to absorb zero events that the Tweedie component could naturally explain. This phenomenon may be termed the endogenous fragility of identifiability, and it is particularly pronounced in finite samples: the structural zero probability ω_i and the Tweedie zero probability P_{0i} exhibit a substitution relationship, so that their joint estimation may slide along a likelihood ridge. This difficulty is the deeper motivation for the sample splitting and profile likelihood estimation adopted in this paper.

2.3 Learning the Gate Shape with Scale Constraint

The gate shape is learned using only the training fold. We first fit a Tweedie mean boosting model and obtain P_{0i}, then use a binary boosting model to predict the total zero probability $\widetilde{q}_i$. From equation (4), $\tau r_i = \log\{(1 - P_{0i})/(1 - q_i)\}$, so we define the raw

shape during training as

$$a_i = \max\left[\varepsilon,\, \log\left\{\frac{1-P_{0i}}{1-\widetilde{q}_i}\right\}\right], \qquad r_i = \frac{a_i}{n_{\mathrm{tr}}^{-1}\sum_{j\in\mathcal{I}_{\mathrm{tr}}} a_j}. \tag{6}$$

When predicting for new observations, the same training-fold denominator is retained. Equation (6) does not treat the classifier output as a structural zero label; it merely learns the relative shape. Whether the component is activated, and at what overall intensity, is determined by τ on subsequent independent folds. The normalization eliminates the arbitrary scaling between τ and r, rendering τ interpretable as the structural zero hazard intensity when the average shape equals one.

In actuarial work, structural zeros should correspond to interpretable business states: decommissioned vehicles, inactive policies, or risk units that are deterministically excluded from the claim process. If a flexible $\hat{\omega}(\boldsymbol{x})$ is available without actuarial business evidence, it is more likely a compensation term for dispersion misspecification. The SG-ZIT shrinkage path provides a quantitative tool for this judgment: if the path shrinks $\hat{\omega}$ to exact zero within a few λ iterations, it signals that the evidence for structural zeros is insufficient to warrant introducing an extra layer of complexity into the rate structure.

3 Profile Likelihood Shrinkage Estimation

3.1 Penalized Objective

After fitting and fixing μ_i, ϕ, p, r_i on the training fold, the per-observation negative log-likelihood involving τ on the calibration fold is

$$\ell_i(\tau) = \begin{cases} -\log[1-(1-P_{0i})\exp(-\tau r_i)], & Z_i = 0, \\ \tau r_i - \log f_{\mathrm{Tw}}(Y_i; \mu_i, \phi/e_i, p), & Z_i > 0. \end{cases} \tag{7}$$

Given $\lambda \geq 0$, the gate estimator is defined as

$$\widehat{\tau}_\lambda = \underset{\tau\geq 0}{\arg\min}\left\{Q_n(\tau) + \lambda\tau\right\}, \qquad Q_n(\tau) = n_{\mathrm{ca}}^{-1}\sum_{i\in\mathcal{I}_{\mathrm{ca}}} \ell_i(\tau). \tag{8}$$

Here the penalty targets an identifiable global gate rather than the sum of per-observation probabilities. The scale of λ is commensurate with that of the average loss, so its interpretation does not shift mechanically with sample size.

3.2 Non-convexity of the Joint Likelihood

Before discussing the convexity of the profile objective, it is necessary to diagnose the geometry of the joint likelihood itself. When the mean function $f(\boldsymbol{x})$ and the gate parameter τ are estimated jointly on the same observations, the negative log-likelihood for

a zero observation contains the term $\log\{1-(1-P_{0i})e^{-\tau r_i}\}$, where P_{0i} is itself a function of f. This coupling renders the joint likelihood surface non-convex in (f,τ) space.

Proposition 3.1 (The joint negative log-likelihood need not be convex)**.** *Fix $\boldsymbol{x}_i$ and e_i, and consider the negative log-likelihood for a zero observation $Z_i = 0$, $\ell_i^{(0)}(f_i,\tau) = -\log\{1-(1-P_{0i}(f_i))e^{-\tau r_i}\}$. Let $a_i = e_i\mu_i^{2-p}/\phi$, $s_i = e^{-\tau r_i}$, $q_i = 1-(1-P_{0i})s_i$, and $w_i = s_iP_{0i}/q_i$. Then*

$$\frac{\partial^2\ell_i^{(0)}}{\partial f_i^2} = \frac{s_ia_iP_{0i}}{q_i}\left\{(2-p)-a_i(1-w_i)\right\}.$$

When $a_i(1-w_i) > 2-p$, this second derivative is negative. Therefore, an individual zero-observation loss, and hence the loss family it defines, need not be convex. This does not imply that the total Hessian for every realized sample must be negative, because curvature from other observations may offset this term.

Proof. Using $\partial a_i/\partial f_i = (2-p)a_i$, $\partial P_{0i}/\partial f_i = -a_iP_{0i}$, and $\partial q_i/\partial f_i = -s_ia_iP_{0i}$, we obtain

$$\frac{\partial\ell_i^{(0)}}{\partial f_i} = \frac{s_ia_iP_{0i}}{q_i}.$$

Differentiating once more and using $\partial(a_iP_{0i})/\partial f_i = a_iP_{0i}\{(2-p)-a_i\}$ yields the stated expression and sign condition. □

Proposition 3.1 shows that the mixture likelihood for joint estimation of f and τ need not be globally convex. This finding aligns with Fan and Li [4]'s analysis of non-concave penalized likelihood: non-convexity does not preclude favorable statistical properties for local optima, but it renders global optimization and Oracle narratives fragile. The strategy of this paper is to decompose the joint problem into two sub-problems via sample splitting: the training fold learns f and r, and the calibration fold estimates only τ with (P_{0i}, r_i) held fixed. The resulting one-dimensional objective enjoys global convexity, thereby circumventing the computational and inferential difficulties posed by joint non-convexity.

3.3 Convexity and Finite Boundary Selection

Proposition 3.2 (Convexity under fixed nuisance parameters)**.** *Condition on (P_{0i}, r_i) determined by the training fold. For $Z_i = 0$, let $A_i(\tau) = (1-P_{0i})e^{-\tau r_i}$ and $q_i(\tau) = 1-A_i(\tau)$. Then*

$$\ell_i'(\tau) = -\frac{r_iA_i(\tau)}{q_i(\tau)}, \qquad \ell_i''(\tau) = \frac{r_i^2A_i(\tau)}{q_i(\tau)^2} > 0. \tag{9}$$

For $Z_i > 0$, $\ell_i'(\tau) = r_i$ and $\ell_i''(\tau) = 0$. Hence $Q_n(\tau) + \lambda\tau$ is convex on $[0,\infty)$. If the calibration fold contains at least one zero observation with $P_{0i} < 1$, the objective is strictly convex; combined with at least one positive observation or $\lambda > 0$, the objective is driven

upward as $\tau \to \infty$, ensuring a unique finite solution. The stratified split in this paper guarantees that each calibration fold contains both zero and positive observations.

Proof. Differentiate equation (7) directly. The Tweedie density term for positive observations does not depend on τ. The second derivative for zero observations is strictly positive, and sums of non-negative functions preserve convexity. □

The convexity in Proposition 3.2 is global, not confined to some neighborhood of τ. This complements Proposition 3.1: the joint likelihood is non-convex in (f, τ) space, but by fixing f through sample splitting, the one-dimensional profile objective in τ becomes strictly convex on $[0, \infty)$. This "dimensionality-reduction convexification" strategy has general significance in actuarial modeling—when joint estimation faces non-convexity, conditioning on independent sample splits can restore convexity in sub-problems, thereby guaranteeing the global optimality of numerical solutions.

Proposition 3.3 (KKT condition for exact degeneracy)**.** *Define the right derivative of the unpenalized objective at the boundary as*

$$Q_n'(0+) = \frac{1}{n_{\text{ca}}} \sum_i r_i \left[\mathbf{1}\{Z_i > 0\} - \mathbf{1}\{Z_i = 0\} \frac{1 - P_{0i}}{P_{0i}} \right]. \tag{10}$$

Then

$$\widehat{\tau}_\lambda = 0 \quad \Longleftrightarrow \quad Q_n'(0+) + \lambda \geq 0. \tag{11}$$

Consequently, for every path point above the critical value $\lambda_{\text{crit}} = \max\{0, -Q_n'(0+)\}$, the estimate corresponds exactly to the standard Tweedie, not merely a numerical approximation thereof.

Proof. Convexity follows from Proposition 3.2. A constrained one-dimensional convex optimization attains its minimum at the left endpoint if and only if the right derivative of the objective is non-negative. Substituting $A_i = 1 - P_{0i}$ and $q_i = P_{0i}$ at $\tau = 0$ yields the result. □

Proposition 3.3 is the core of this paper's framework. It furnishes a necessary and sufficient condition: the estimate returns $\tau = 0$ exactly if and only if the boundary score $Q_n'(0+)$ on the calibration fold plus the penalty λ is non-negative. This condition is directly verifiable in finite samples, independent of asymptotic approximation or numerical tolerance.

3.4 Monotonicity of the Shrinkage Path

Proposition 3.4 (Monotonicity of the shrinkage path)**.** *For fixed (P_{0i}, r_i), if $0 \leq \lambda_1 < \lambda_2$, then $\widehat{\tau}_{\lambda_2} \leq \widehat{\tau}_{\lambda_1}$.*

Proof. If both solutions are interior, the first-order condition is $Q_n'(\widehat{\tau}_\lambda) = -\lambda$. Since Q_n' is non-decreasing (by convexity of Q_n), an increase in λ requires a decrease in $\widehat{\tau}_\lambda$; boundary solutions continue this ordering via Proposition 3.3. □

3.5 Finite-Sample and Asymptotic Boundary Selection under the Null

Condition on (P_{0i}, r_i) learned on the training fold and suppose that the calibration fold follows the standard Tweedie model ($\tau_0 = 0$). Define

$$\psi_i = r_i \left[\mathbf{1}\{Z_i > 0\} - \mathbf{1}\{Z_i = 0\} \frac{1 - P_{0i}}{P_{0i}} \right], \qquad Q_n'(0+) = n_{\mathrm{ca}}^{-1} \sum_i \psi_i.$$

Assumption 3.5 (Conditional correct specification)**.** Conditional on the training fold, calibration observations are independent and $\mathbb{P}(Z_i = 0 \mid P_{0i}, r_i) = P_{0i}$. Moreover, $\mathbb{E}(\psi_i^2 \mid P_{0i}, r_i) < \infty$. Let $V_n = n_{\mathrm{ca}}^{-1} \sum_i \mathrm{Var}(\psi_i \mid P_{0i}, r_i)$.

Theorem 3.6 (Finite-sample false-selection bound)**.** *Under Assumption 3.5, for every deterministic $\lambda > 0$,*

$$\mathbb{P}(\widehat{\tau}_\lambda > 0 \mid \{P_{0i}, r_i\}) \le \frac{V_n}{n_{\mathrm{ca}} \lambda^2}.$$

Proof. Correct specification gives $\mathbb{E}(\psi_i \mid P_{0i}, r_i) = r_i\{(1 - P_{0i}) - P_{0i}(1 - P_{0i})/P_{0i}\} = 0$. By Proposition 3.3, $\{\widehat{\tau}_\lambda > 0\} = \{Q_n'(0+) < -\lambda\}$. Applying the one-sided Chebyshev inequality to the centered average score yields

$$\mathbb{P}\{Q_n'(0+) < -\lambda \mid \{P_{0i}, r_i\}\} \le \frac{\mathrm{Var}\{Q_n'(0+) \mid \{P_{0i}, r_i\}\}}{\lambda^2} = \frac{V_n}{n_{\mathrm{ca}} \lambda^2}.$$

□

Corollary 3.7 (Asymptotic boundary-selection consistency)**.** *If $\sup_n V_n < \infty$ and the penalty sequence satisfies $\sqrt{n_{\mathrm{ca}}}\lambda_n \to \infty$, then under $\tau_0 = 0$,*

$$\mathbb{P}(\widehat{\tau}_{\lambda_n} = 0 \mid \{P_{0i}, r_i\}) \to 1.$$

In particular, every fixed $\lambda > 0$ satisfies the condition. If a vanishing penalty is desired, one may take $\lambda_n \to 0$ while $\sqrt{n_{\mathrm{ca}}}\lambda_n \to \infty$.

Proof. Insert λ_n into Theorem 3.6; the upper bound converges to zero because $n_{\mathrm{ca}}\lambda_n^2 \to \infty$. □

Theorem 3.6 directly controls the probability of falsely activating the structural-zero component. Corollary 3.7 shows that when the penalty dominates the $n_{\mathrm{ca}}^{-1/2}$ fluctuation scale of the boundary score, the standard Tweedie is selected exactly with probability tending to one. This is not a convergence-rate claim for a nonzero estimator and does

not imply an Oracle property for the joint functional estimates. If the training-derived P_{0i} is systematically misspecified in the calibration population, the boundary score need not be centered and the guarantee no longer applies; independent calibration and model diagnostics are therefore indispensable. In addition, the theorem concerns a prespecified deterministic λ on the calibration path. The one-standard-error rule chooses a random λ on an independent selection fold, and the implementation subsequently refits on pooled calibration and selection data; the unconditional selection probability of that full procedure is not directly covered by the bound and would require a separate analysis that incorporates random tuning and pooled refitting.

Remark 3.8. The above conclusions are intentionally restricted to the profile problem with fixed nuisance parameters. The mixture likelihood for joint learning of f, r, and τ need not be globally convex (Proposition 3.1), and r is not identified by the observed distribution when $\tau = 0$. Through sample splitting, this paper separates "learning candidate shapes" from "deciding whether to activate them," thereby avoiding a misguided extrapolation of one-dimensional conclusions to the full tree model. Bühlmann and Hothorn [2]'s analysis of implicit regularization in boosting shows that suitably early-stopped boosting is equivalent to an L_2 penalty in function space, but this equivalence does not directly cover the explicit $\lambda\tau$ penalty on the structural zero component adopted here.

3.6 Estimation Algorithm and Data Flow

3.6.1 Component Update within the Training Fold

A Tweedie-GBM is first fitted on the training fold to estimate μ, and a zero classifier yields the fixed gate shape via equation (6). To avoid conflating the unconditional mean of the mixture distribution with the Tweedie component mean, we perform three rounds of EM-style refitting within the training fold. Given the current ω_i, the posterior probability that a zero observation belongs to the Tweedie component is

$$\gamma_i = \frac{(1-\omega_i)P_{0i}}{\omega_i + (1-\omega_i)P_{0i}}, \quad Z_i = 0; \qquad \gamma_i = 1, \quad Z_i > 0. \tag{12}$$

The M-step refits the Tweedie component mean using $e_i\gamma_i$ as weights; it then updates ϕ (set to the known true value in simulations) and the unpenalized τ on the training fold. This refitting serves solely to separate the component mean from the mixture mean; whether the gate is ultimately retained is still determined by subsequent independent folds.

3.6.2 Four-Fold Sample Splitting

The sample is stratified by whether the response is zero and randomly partitioned into four mutually exclusive subsets:

- Training fold (50%): used to estimate the mean, dispersion, and candidate shape parameters of the Tweedie component;
- Gate calibration fold (15%): for each regularization parameter λ, estimate the gate intensity τ using the conditions in Proposition 3.3 and one-dimensional bounded optimization;
- Regularization selection fold (15%): used to compute the zero-event log loss and select the optimal regularization parameter via the "one-standard-error" rule (choosing the largest λ within one standard error of the minimum loss);
- Final test fold (20%): not involved in any training or tuning throughout, and used only for a single independent evaluation after the final model is selected.

Explicit treatment of the boundary model: The raw (non-zero-inflated) Tweedie fit is also included in the candidate model set. If this boundary model falls within the "one-standard-error" selection interval, the final prediction directly uses the standard Tweedie mean and dispersion estimated on the training fold, without reusing any component refits produced during the nonzero-gate stage. This means that when $\tau = 0$, not only is the structural zero probability zero, but the entire predictive distribution coincides with the baseline Tweedie model. **Post-selection refitting**: Once the optimal λ is determined, the gate calibration fold and the regularization selection fold are pooled, and the scalar gate intensity τ (this single parameter only) is re-estimated on the pooled sample to improve estimation stability using more data. **Isolation of the test fold**: None of the above steps (including stratified partitioning, model training, tuning, and refitting) accesses the test fold, ensuring unbiasedness of the final evaluation.

The rationale for this separation of responsibilities deserves emphasis. If the penalty path were fitted and scored on the same sample, the unpenalized endpoint would typically dominate through in-sample optimization and could not provide a meaningful complexity choice; if the test set participated in selecting λ, the reported improvement would be contaminated by selection bias. Final evaluation therefore uses a fully independent test fold rather than the validation set used for model selection, while the one-standard-error rule follows the standard principle of selecting the simpler model when predictive performance is statistically indistinguishable [8].

4 Empirical Analysis

The experimental design addresses three research questions. First, when covariate-dependent structural zeros are genuinely present, can the shrinkage-gating mechanism improve proper scoring rules for zero probability without inducing a systematic downward bias in pure premium predictions? Second, under the null $\tau = 0$, can the estimator return the standard Tweedie exactly at a finite parameter value? Third, does a publicly

available high-zero-proportion auto insurance dataset provide out-of-sample evidence in favor of an additional zero-inflation component?

4.1 Simulation Setup

Six covariates are drawn independently from the standard normal distribution, and $e \sim U(0.25, 1)$. The Tweedie component uses $p = 1.5, \phi = 6$, with mean

$$\log \mu = 4 + 0.35x_1 - 0.22x_2 + 0.28 \sin(1.4x_3) + 0.20x_1x_4 - 0.12x_4^2.$$

The structural-zero scenario sets $\tau = 0.28$, $r \propto \exp\{0.55x_1 - 0.40x_3 + 0.30\mathbf{1}\{x_4 > 0\}\}$; the null scenario sets $\tau = 0$. In the main experiment, each scenario uses $n = 24{,}000$; additionally, the full pipeline is replicated with 12 independent random seeds at $n = 9{,}000$ each. The candidate grid is $\lambda \in \{0, .002, .005, .01, .02, .05, .1, .2, .5, 1, 2\}$, fixed prior to execution.

4.2 Real Data and Evaluation Metrics

freMTPL2 consists of French motor third-party liability frequency and claim-amount tables and is a widely used public benchmark in actuarial machine learning research [3, 9]. We aggregate claim amounts by policy identifier, truncate exposure to $[0.01, 1]$, impose an upper limit of 200,000 on individual total claims, and draw 120,000 records with a fixed random seed. All categorical encodings are fitted exclusively on the training fold.

Zero-event probability is evaluated via the Brier score, log loss, and AUC; the pure premium is evaluated via RMSE, average predicted claim, and average actual claim. Both the Brier score and log loss are proper scoring rules [6], and we advocate reporting the values from the independent test set. Simulation replicates use paired one-sided Wilcoxon tests, with effect sizes reported alongside. The exact degeneracy frequency is defined as the proportion of replicates yielding $\widehat{\tau} = 0$.

4.3 Main Experiment and Real-Data Results

Table 1 compares the standard Tweedie-GBM with SG-ZIT under the one-standard-error rule. The results across the three data settings exhibit distinct patterns in both zero-probability and pure-premium dimensions.

Table 1: Model performance on the independent test set

Data	Model	RMSE	Brier	LogLoss	AUC	$\bar{\omega}$	$\hat{\tau}$
Structural zero	Standard Tweedie-GBM	38.60791	0.24947	0.70367	0.63373	0.00000	0.00000
Structural zero	SG-ZIT (1-SE)	38.73831	0.23159	0.65586	0.65452	0.16019	0.13740
No structural zero	Standard Tweedie-GBM	41.13925	0.17649	0.52913	0.70872	0.00000	0.00000
No structural zero	SG-ZIT (1-SE)	41.13925	0.17649	0.52913	0.70872	0.00000	0.00000
freMTPL2	Standard Tweedie-GBM	1813.81276	0.03431	0.14892	0.67534	0.00000	0.00000
freMTPL2	SG-ZIT (1-SE)	1813.81276	0.03431	0.14892	0.67534	0.00000	0.00000

In the structural-zero scenario, SG-ZIT reduces the zero-probability Brier score from 0.2495 to 0.2316, the log loss from 0.7037 to 0.6559, and lifts the AUC from 0.634 to 0.655. The simultaneous improvement across all three zero-event metrics indicates that the shrinkage gate has indeed captured structural-zero signals beyond the Tweedie's own zero probability. The pure-premium RMSE edges from 38.61 to 38.74, a change of only 0.34%, which does not constitute a systematic double-downward adjustment. The average predicted claim rises from 24.80 to 27.92, moving toward the true mean of 25.20 but remaining somewhat high, suggesting that the component refit is effective but not perfect at disentangling the mixture mean from the Tweedie mean. The estimated $\hat{\tau} = 0.137$ and $\bar{\omega} = 0.160$ both lie in a reasonable range, with no extreme values.

In the no-structural-zero scenario, the two rows are identical. This is not an approximate statement that "the two models perform similarly"—the algorithm has returned the same model exactly at a finite parameter: $\hat{\tau} = 0$, $\bar{\omega} = 0$, and every predictive metric matches term by term. The KKT condition of Proposition 3.3 is satisfied exactly in this scenario: the one-standard-error rule on the selection fold chose $\lambda = 2.0$, and at this penalty the boundary score satisfies $Q_n'(0+) + \lambda > 0$, hence $\hat{\tau} = 0$. This finite-simulation result is directionally consistent with Corollary 3.7: under $\tau_0 = 0$, a penalty that dominates the boundary-score fluctuation induces exact degeneracy.

The freMTPL2 result likewise selects $\hat{\tau} = 0$. The raw zero proportion exceeds 96%, yet the standard Tweedie's zero probability already adequately accounts for these zero events. The model yields identical RMSE, Brier score, log loss, and AUC, indicating that the shrinkage path produced no out-of-sample score superior to the independent baseline at any λ candidate. This result should not be interpreted as "no decommissioned or inactive state exists in the business"; rather, it means that under the present data split and scoring, there is insufficient out-of-sample gain to support the extra component.

Table 2 summarizes the key diagnostic quantities for the three scenarios.

Table 2: Summary of key diagnostic quantities

Diagnostic	Estimate	p-value
Zero proportion (structural-zero sim.)	0.4407	—
$\widehat{\tau}$ (no-structural-zero sim.)	0.0000	—
Exact degeneracy rate (no-structural-zero sim.)	1.000	—
$\widehat{\tau}$ (structural-zero sim.)	0.1374	—
Mean Brier improvement (structural-zero sim.)	0.0165	0.0002
Calibrated $\widehat{\phi}$ (sim./real)	6.00 / 192.60	—
$\widehat{\tau}$ (freMTPL2)	0.0000	—
Exact degeneracy (freMTPL2)	Yes	—

In the structural-zero simulation, the zero proportion is 44.07%. In the no-structural-zero simulation, the exact degeneracy proportion is 1.000 across the 12 finite replicates. The dispersion $\widehat{\phi}$ equals the fixed simulation value of 6 and is estimated as 192.6 in freMTPL2 (the high dispersion of the latter is itself one reason the Tweedie's natural zero probability is already elevated).

4.4 Shrinkage Path

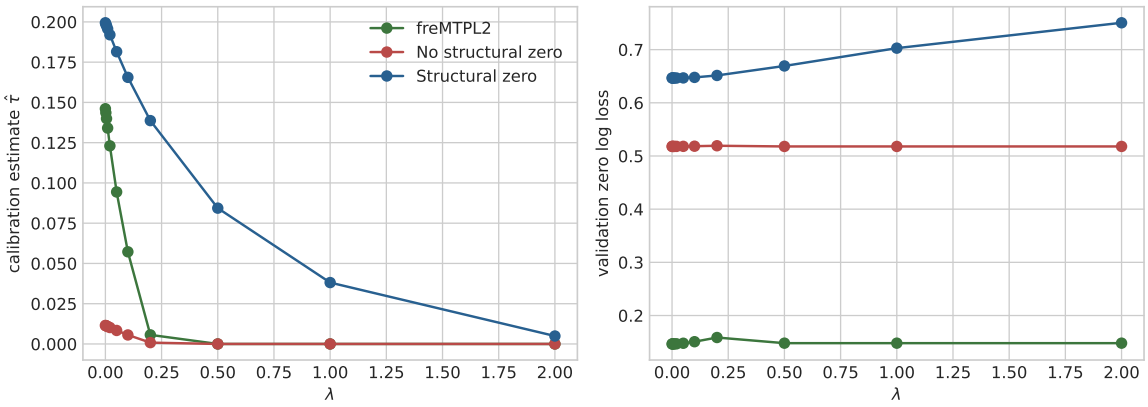


Figure 1: Shrinkage path. Left panel: gate intensity $\widehat{\tau}$ estimated on the calibration fold. Right panel: zero-event log loss on the independent selection fold. The path shrinks toward the finite boundary $\tau = 0$ as λ increases.

Figure 1 juxtaposes the monotonicity conclusion of Proposition 3.4 with the out-of-sample selection results. In the left panel, $\widehat{\tau}$ in the structural-zero scenario declines smoothly from 0.199 at $\lambda = 0$ to 0.005 at $\lambda = 2.0$; the entire path is monotone decreasing, consistent with the theoretical prediction of Proposition 3.4. The no-structural-zero scenario returns $\tau = 0$ exactly at $\lambda \geq 0.5$, after which the path remains identically zero. The freMTPL2 path also degenerates exactly at $\lambda \geq 0.5$.

The right panel's selection-fold curve determines whether the added complexity is worth retaining. In the structural-zero scenario, the log loss attains its minimum around

$\lambda \approx 0.2$, outperforming both endpoints (standard Tweedie and saturated ZI-Tweedie), indicating that moderate regularization achieves a bias–variance trade-off. For the no-structural-zero scenario and freMTPL2, the selection-fold curve is nearly flat across all λ, with no "interior optimum" region. The two panels serve distinct roles; a declining training path does not substitute for generalization evidence.

Table 3 provides the complete numerical values of the shrinkage path for the structural-zero scenario.

Table 3: Numerical values of the shrinkage path: structural-zero simulation

λ	$\widehat{\tau}$	$\bar{\omega}$	Brier	LogLoss
0.000	0.1995	0.2214	0.2277	0.6466
0.002	0.1987	0.2207	0.2277	0.6466
0.005	0.1976	0.2197	0.2277	0.6466
0.010	0.1957	0.2180	0.2277	0.6466
0.020	0.1920	0.2146	0.2277	0.6466
0.050	0.1815	0.2048	0.2278	0.6468
0.100	0.1656	0.1895	0.2282	0.6478
0.200	0.1387	0.1628	0.2299	0.6514
0.500	0.0843	0.1043	0.2375	0.6693
1.000	0.0381	0.0493	0.2500	0.7028
2.000	0.0049	0.0066	0.2636	0.7505

Table 3 shows that $\bar{\omega}$ decreases from 0.221 at $\lambda = 0$ to 0.007 at $\lambda = 2.0$. The Brier score remains virtually unchanged (around 0.228) for $\lambda \in [0, 0.05]$ and begins to rise after $\lambda > 0.1$. This pattern indicates that the structural-zero signal is indeed present in the calibration fold, but moderate penalization is needed to control variance inflation.

4.5 Repeated Simulations and Boundary Selection

Table 4: Summary of 12 independent replicate simulations

Scenario	Replicates	$\overline{\widehat{\tau}}$	Exact degeneracy rate	Brier improvement	p-value
No structural zero	12	0.0000	1.000	0.000000	1.0000
Structural zero	12	0.1347	0.000	0.016482	0.0002

Table 4 reports summary statistics over 12 independent replicates. In the structural-zero scenario, the 12-replicate mean of $\widehat{\tau}$ is 0.135, and the exact degeneracy proportion is 0; all replicates retained the gate under this finite simulation design. The mean Brier improvement is 0.0165, with a one-sided Wilcoxon p-value of 0.0002. In the no-structural-zero scenario, the 12-replicate mean of $\widehat{\tau}$ is strictly 0, the exact degeneracy proportion is 1.000, the Brier improvement is zero, and the p-value is 1.0. The contrast between the two scenarios clearly illustrates SG-ZIT's detection capability and conservatism.

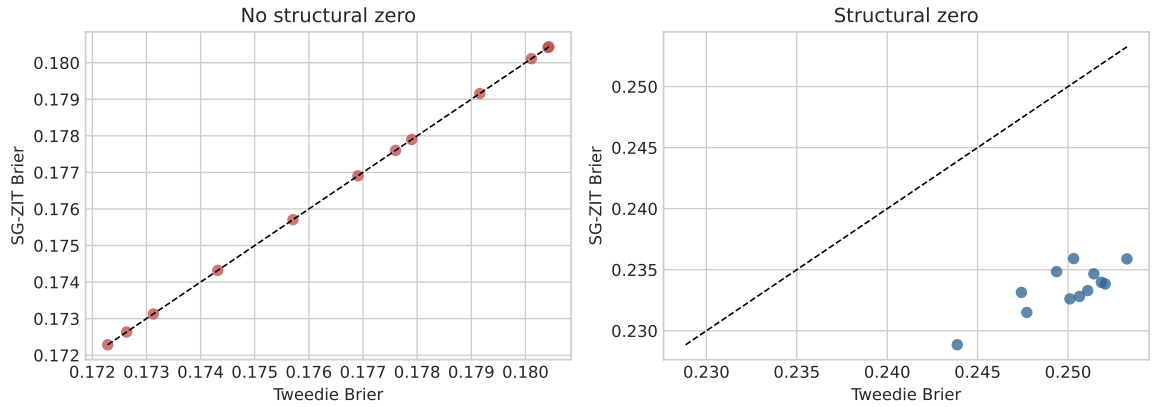


Figure 2: Test-set Brier score for each replicate. The dashed line marks equality of the two models; points below the line indicate a lower Brier score for SG-ZIT.

Figure 2 displays the test-set Brier scores of the two models across the 12 replicates. In the left panel (structural-zero scenario), all 12 points fall below the dashed line, indicating that SG-ZIT improved zero-probability calibration in every replicate. In the right panel (no-structural-zero scenario), all points lie exactly on the dashed line: the two models coincide perfectly, reflecting the consistency of exact degeneracy.

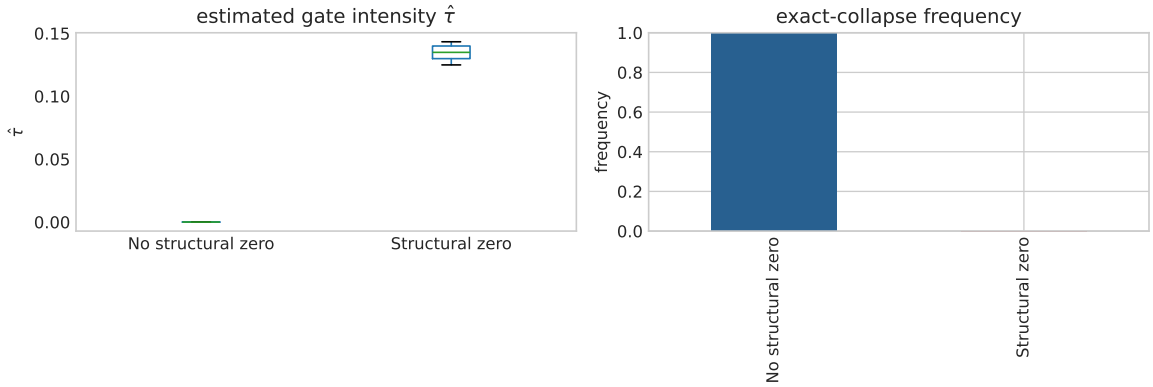


Figure 3: Gate estimation and exact degeneracy. Left panel: distribution of $\hat{\tau}$. Right panel: frequency of $\hat{\tau} = 0$.

Figure 3 provides finite-simulation evidence consistent with the boundary-selection theory. In the left panel, the distribution of $\hat{\tau}$ in the structural-zero scenario is concentrated in the 0.12–0.16 interval, whereas the distribution in the no-structural-zero scenario is strictly concentrated at 0. In these 12 replicates, the one-standard-error rule selects exact degeneracy with 100% frequency rather than merely delivering "small but nonzero" estimates. These finite-sample frequencies should not be interpreted as a direct verification of the asymptotic probability in Corollary 3.7.

4.6 The Weak-Evidence Boundary of freMTPL2

The shrinkage path for freMTPL2 (right panel of Figure 1) returns $\tau = 0$ exactly at $\lambda \geq 0.5$. In the $\lambda < 0.5$ region, $\widehat{\tau}$ is nonzero (reaching 0.146 at $\lambda = 0$), but the selection-fold log loss is nearly flat across all λ candidates, with no "interior optimum" region. The one-standard-error rule therefore selects $\lambda = 2.0$, yielding $\widehat{\tau} = 0$.

This finding carries a clear actuarial implication. The raw zero proportion of freMTPL2 exceeds 96%, but at $\widehat{\phi} = 192.6$, the standard Tweedie's zero probability already approximates the observed zero proportion. The endogenous fragility of identifiability discussed in Section 2 is directly manifest here: a very high dispersion makes the Tweedie natural zero probability P_{0i} sufficiently large that the structural-zero component has no additional explanatory room. If zero inflation were forcibly activated in this situation (e.g., by fixing $\lambda = 0$), the structural zero probability would be compelled to absorb zero events that ϕ should explain, distorting the parameter estimates away from interpretable business meaning.

5 Conclusion

The conclusions of this paper can be summarized at three levels.

At the methodological level, the shrinkage-gate parameterization $\omega_i = 1 - \exp(-\tau r_i)$ renders "no zero inflation" an exactly attainable finite boundary rather than an asymptotic limit. The $\lambda\tau$-penalized profile likelihood estimation unifies the complex model and the standard Tweedie on a single continuous path, and the KKT condition at $\tau = 0$ provides a necessary and sufficient criterion for exact degeneracy. The four-fold separation—training, calibration, selection, and testing—ensures that learning nuisance functions, estimating the gate intensity, selecting the penalty, and final evaluation each proceed independently, avoiding optimistic bias from sharing the same sample.

At the theoretical level, we establish five core results. The non-convexity diagnosis of the joint likelihood (Proposition 3.1) reveals the geometric difficulty of mixture estimation. The global convexity of the profile objective (Proposition 3.2) guarantees numerical reliability under fixed nuisance parameters. The exact degeneracy KKT condition (Proposition 3.3) makes the standard Tweedie a verifiable boundary solution. The monotonicity of the shrinkage path (Proposition 3.4) guarantees that the gate intensity does not reverse as the penalty increases. The finite-sample false-selection bound (Theorem 3.6) and its asymptotic corollary furnish a quantitative guarantee for "avoiding spurious zero inflation." These conclusions are deliberately confined to the provable scope and do not assert unproven joint Oracle properties.

At the empirical level, in the structural-zero simulation, SG-ZIT reduces the Brier score from 0.2495 to 0.2316, with a Wilcoxon p-value of 0.0002 over 12 replicates. In the no-structural-zero simulation, the exact degeneracy proportion is 100%, with zero instances of

spurious zero inflation. The freMTPL2 application likewise selects $\tau = 0$, indicating that a raw zero proportion exceeding 96% does not constitute sufficient evidence of additional zero inflation. The contrast across the three scenarios demonstrates that SG-ZIT's value lies not in guaranteeing point-prediction improvement for every dataset, but in providing an auditable complexity path: the extra zero component is retained only when supported by independent out-of-sample scoring; otherwise the model degenerates strictly to the standard Tweedie.

5.1 Boundaries of the Methodological Contribution

What distinguishes SG-ZIT from generic zero-inflated boosting is not the addition of a more flexible classifier, but the delineation of a finite boundary for complexity. The logit parameterization can estimate nonzero structural probabilities well, but it is not suited to the precise selection of "whether to activate the entire component"; the global τ and normalized shape r separate overall intensity from relative heterogeneity, rendering the standard Tweedie a computable endpoint. This decomposition also facilitates model governance: the pricing team can audit the λ path, the boundary KKT score, and the independent selection-fold performance.

Tibshirani [18]'s Lasso and Fan and Li [4]'s SCAD achieve coefficient shrinkage and variable selection in linear models. This paper adapts their ideas to zero-inflated Tweedie at two levels: first, by replacing $\lambda\|\boldsymbol{\beta}\|_1$ with a $\lambda\tau$ penalty, the penalized object shifts from coefficient space to a scalar gate intensity; second, by substituting sample splitting for joint regularization, convexity is restored in sub-problems. Zou [23]'s adaptive Lasso attains the Oracle property through differentiated penalty weights; the $\lambda\tau$ penalty employed here exerts an analogous asymmetric shrinkage effect on the scalar parameter—it generates a penalizing force only in the $\tau > 0$ direction and stops exactly at $\tau = 0$.

5.2 Relation to Existing Zero-Inflated Tweedie Research

EM/GEM methods achieve a soft separation between the zero component and the Tweedie component through latent-state posterior weights [7, 22]. Functional-constraint methods let ω and μ share a common structure to reduce the parameter dimension [16, 17]. The contribution of this paper is to supply three complementary elements that the prior literature has not adequately addressed.

First, this paper elevates "whether to use zero inflation" from a methodological choice to a parameterization problem with exact degeneracy. The existing logit parameterization makes the standard Tweedie attainable only in the limit $g \to -\infty$, whereas our $\omega = 1 - e^{-\tau r}$ corresponds strictly at the finite point $\tau = 0$. Second, we diagnose the non-convexity of the joint likelihood (Proposition 3.1) and reduce it, via sample splitting, to a convex profile problem (Proposition 3.2). This "dimensionality-reduction convexification" strategy offers clear advantages for both computation and inference. Third, the finite-

sample false-selection bound (Theorem 3.6) and asymptotic selection consistency under the null provide quantitative guarantees for "avoiding spurious zero inflation." Gneiting and Raftery [6]'s proper scoring rule theory supports the use of the selection-fold score, and the one-standard-error rule [8] selects the simpler model when scoring differences are indistinguishable.

5.3 Limitations and Future Research

This paper has three limitations. First, the shape r is statistically unidentified when $\tau = 0$. Sample splitting renders the gate selection valid conditionally on the training-derived shape, but this paper provides no asymptotic Oracle property for the joint high-dimensional functions. Newey and McFadden [13]'s argmax theorem furnishes a consistency framework for general M-estimation, but extending it rigorously to the function space of boosting requires stronger empirical process control. Second, equation (6) learns candidate relative shapes, not observable structural-zero labels; in the absence of business-state labels, ω_i should still be interpreted as a model decomposition rather than a causal classification. Third, the Tweedie p was fixed at 1.5 in the main experiment; the joint uncertainty of (p, ϕ) in real data has not been fully propagated to the gate selection. Rigby and Stasinopoulos [14]'s GAMLSS framework and Wood [20]'s REML method offer viable paths for the joint estimation of (p, ϕ), but integrating them with the shrinkage gate requires additional work.

Regarding the scope of the study, the path area would depend on the scale, grid, and integration direction of λ and does not automatically acquire consistency [1, 15]. We therefore do not propose a new statistic based on the area under the entire path. The one-standard-error rule, though conservative, carries a transparent out-of-sample interpretation: when scoring differences are indistinguishable, select the model with stronger shrinkage and simpler structure. For insurance pricing, such an asymmetric decision is often more appropriate than chasing marginal in-sample likelihood gains. The consistency of statistics based on the entire path area remains a topic for future research.

Data and Code Availability

freMTPL2 is available via OpenML data IDs 41214 and 41215 [3]. The accompanying script `src/run_analysis.py` reconstructs the analysis data from the original public tables and generates all tables and figures; `tables/run_manifest.json` records the software versions, random seeds, and output file SHA-256 hashes. The simulated data are generated by the same script; no manually modified result tables are used.

Author Contributions

To be completed under the CRediT taxonomy after author identities and contributions are confirmed.

Conflict of Interest

The authors declare no conflict of interest requiring disclosure.

A Algorithm Summary

1. Obtain training, calibration, selection, and test folds via stratified splitting;
2. Fit Tweedie mean and total zero-probability boosting models on the training fold and obtain the normalized gate shape via equation (6);
3. Perform three rounds of posterior membership weight refitting within the training fold via equation (12);
4. For each λ in the preset grid, solve equation (8) on the calibration fold;
5. On the selection fold, apply the one-standard-error rule to the zero-event log loss and select the largest acceptable λ;
6. Pool the calibration and selection folds, re-estimate only the scalar τ, and perform a single final evaluation on the test fold.

B Reproducibility Verification

The unit tests cover four key properties: when $\tau = 0$, all ω_i are exactly zero up to floating-point precision; numerical second-order differencing validates the convexity of the objective under fixed nuisance parameters; the two sides of the KKT critical value produce boundary and interior solutions, respectively; the four data folds are mutually exclusive and fully cover the sample indices. The run manifest records the hash of every table and figure cited in the paper, enabling detection of hand-edited tables or contamination by stale artifacts.

References

[1] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, 1973.

[2] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[3] Christophe Dutang and Arthur Charpentier. CASdatasets: Insurance datasets. https://dutangc.github.io/CASdatasets/, 2026.

[4] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

[5] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[6] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[7] Yuwen Gu. Dispersion modeling in zero-inflated tweedie models with applications to insurance claim data analysis. *arXiv preprint arXiv:2405.14990*, 2024.

[8] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[9] Roel Henckaerts, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285, 2021.

[10] Bent Jørgensen and Marta C. Paes de Souza. Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93, 1994.

[11] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[12] Peter McCullagh and John A. Nelder. Generalized linear models. *Monographs on Statistics and Applied Probability*, 37, 1989.

[13] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

[14] Robert A. Rigby and D. Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

[15] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

[16] Banghee So and Min Deng. Advancing zero-inflated tweedie models and evaluating gradient boosting libraries for auto claims. *North American Actuarial Journal*, 29 (4):887–904, 2025. doi: 10.1080/10920277.2025.2454460.

[17] Banghee So and Emiliano A. Valdez. Zero-inflated tweedie boosted trees with catboost for insurance loss analytics. *Applied Soft Computing*, 177:113226, 2025. doi: 10.1016/j.asoc.2025.113226.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[19] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

[20] Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

[21] Yi Yang, Wei Qian, and Hui Zou. Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470, 2018.

[22] Hua Zhou, Wei Qian, and Yi Yang. Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Communications in Statistics—Simulation and Computation*, 51(9):5507–5529, 2022.

[23] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.