Nonparametric Estimation of Splicing Points in Actuarial Loss Distributions via Data Transformation

Benedikt Funke[a,*], Masayuki Hirukawa[b]

[a] *Institute for Insurance Studies, TH Köln – University of Applied Sciences, Gustav-Heinemann-Ufer 54, 50968 Köln, Germany*
[b] *Faculty of Economics, Ryukoku University, 67 Tsukamoto-cho, Fukakusa, Fushimi-ku, Kyoto 612-8577, Japan*

Abstract

We propose a nonparametric method for estimating splicing points in actuarial loss distributions. Our approach transforms non-negative loss data onto the unit interval $[0, 1]$, which magnifies jump discontinuities in the density and facilitates detection of the splicing point even in the sparse tail region. The transformed data are smoothed using the asymmetric beta kernel, and the splicing point estimator is obtained by back-transforming the maximizer of a diagnostic function. We establish strong consistency and asymptotic normality with a convergence rate faster than the parametric one. Monte Carlo simulations and applications to non-life insurance loss data confirm the practical relevance of the approach.

Keywords: beta kernel, cross validation, data transformation, actuarial loss distributions, splicing point

*Corresponding author
Email addresses: benedikt.funke@th-koeln.de (Benedikt Funke), hirukawa@econ.ryukoku.ac.jp (Masayuki Hirukawa)

1. Introduction

In non-life insurance, the accurate modelling of loss distributions is essential for pricing, reserving, and solvency capital determination. A well-known challenge is that the right tail of such distributions, where infrequent but severe losses reside, typically follows a different pattern than the bulk of attritional claims. Actuaries therefore rely on splicing or composite models that combine separate distributions for the bulk and the tail, connected at a threshold or splicing point (e.g., Klugman et al., 2019). Misspecification of this splicing point can lead to distorted risk measures and inadequate reinsurance structures.

While many methods for threshold selection exist, ranging from heuristic quantile rules (e.g., DuMouchel, 1983) and graphical diagnostics (e.g., Scarrott and MacDonald, 2012) to automated procedures based on extreme value theory (EVT) (e.g., Clauset et al., 2009; Bader et al., 2018; Danielsson et al., 2019), most of them either depend on a specific tail model such as the generalized Pareto distribution (GPD) or leave substantial room for practitioners' discretion. Recently Funke and Hirukawa (2025) have proposed a model-free approach that interprets the splicing point as a jump location in the density and estimates it nonparametrically using the asymmetric gamma kernel on $\mathbb{R}_+$ by Chen (2000). Their estimator is strongly consistent and asymptotically normal with a convergence rate faster than the parametric one, but its finite-sample performance deteriorates when the jump size is small or the splicing point lies deep in the sparse tail.

In our contribution, we address precisely this limitation. We propose transforming the original loss data via a known monotone mapping $T\colon \mathbb{R}_+ \to [0,1]$ before estimating the splicing point. The transformation magnifies the relative jump size by a factor of $1/T^{(1)}(t_0) > 1$ and compresses inter-observation distances in the tail. In the transformed scale, we employ the asymmetric beta kernel by Chen (1999) to construct a diagnostic function whose maximizer yields the splicing point estimator after back-transformation. It is demonstrated that the estimator is strongly consistent and asymptotically normal with a convergence rate exceeding $\sqrt{n}$, where n is the sample size. Monte Carlo experiments confirm that the transformation-based estimator substantially outperforms both the original-scale estimator and several automated threshold detection methods. Four applications to real insurance loss datasets illustrate its practical value.

The remainder of this paper is organized as follows. Section 2 discusses the estimation procedure of the splicing point via data transformation. Large- and finite-sample properties of the proposed splicing point estimator are investigated in Sections 3 and 4, respectively. In Section 5, the estimator is applied to several real datasets. Section 6 concludes. The online Supplement provides all technical proofs and details of numerical analyses.

Throughout, '$a_n = O(b_n)$' means that a_n/b_n is bounded, and '$a.s.$' abbreviates "almost surely". For a function $h(x)$ and a point c, $h(c^-) = \lim_{x\uparrow c} h(x)$, $h(c^+) = \lim_{x\downarrow c} h(x)$ and $h^{(m)}(x) = d^m h(x)/dx^m$ denote the left and right limits, and the mth-order derivative, respectively.

2. Methodology

2.1. Setup and data transformation

Suppose the pdf $f_X(x)$ of an actuarial loss variable $X \in \mathbb{R}_+$ has a jump discontinuity at t_0 strictly inside a prespecified closed interval $I_0 := [\underline{t}, \bar{t}]$ with $0 < \underline{t} < \bar{t} < \infty$. The interval I_0 is assumed to lie in the right tail. Practitioners often have rough prior knowledge of its location through preliminary estimates, historical experience, policy limits, or empirical quantiles. We model the pdf locally as

$$f_X(x) = g_X(x) + d_0 \, \mathbf{1}\{x < t_0\},$$

where g_X is smooth and $d_0 := f_X(t_0^-) - f_X(t_0^+) \neq 0$ is the jump size.

When $|d_0|$ is small or data in the tail are sparse, detecting t_0 directly on $\mathbb{R}_+$ is difficult. Let $T \colon \mathbb{R}_+ \to [0,1]$ be a known smooth monotone transformation. Writing $Y_i = T(X_i) \in [0,1]$, the transformed pdf satisfies

$$f_Y(y) = g_Y(y) + d_T \, \mathbf{1}\{y < t_T\} \tag{1}$$

on $I_T := [T(\underline{t}), T(\bar{t})] \subset [0,1]$, with $t_T = T(t_0)$ and $d_T = f_Y(t_T^-) - f_Y(t_T^+)$. A direct calculation yields the key relationship

$$d_0 = d_T \, T^{(1)}(t_0).$$

If $0 < T^{(1)} < 1$ on I_0, then $|d_T| > |d_0|$ and as a consequence the jump

size is magnified in the transformed scale. Table 1 lists four transformations satisfying these requirements, with magnification factors up to 16 for realistic splicing point locations.

2.2. Estimation procedure

Following the jump-location detection strategy by Funke and Hirukawa (2025), we use shifted beta kernels. For a design point $y \in [0,1]$, the smoothing parameter $b > 0$ and the shift $\Delta > 0$, the shifted beta kernels are the densities of $Beta\{(y \pm \Delta)/b + 1, (1 - y \mp \Delta)/b + 1\}$ and take the forms of

$$K_y^{\pm}(u) = K_{B(y,b;\pm\Delta)}(u) = \frac{u^{(y\pm\Delta)/b}(1-u)^{(1-y\mp\Delta)/b}}{B\{(y \pm \Delta)/b + 1, (1 - y \mp \Delta)/b + 1\}} \mathbf{1}\{u \in [0,1]\}.$$

These kernels concentrate their mass slightly to the left or right of y, so that the difference between two shifted density estimators

$$\hat{f}_Y^{\pm}(y) := \frac{1}{n}\sum_{i=1}^{n} K_y^{\pm}(Y_i)$$

serves as a diagnostic for detecting discontinuities. By defining $\hat{J}(y) := \hat{f}_Y^{-}(y) - \hat{f}_Y^{+}(y)$ and $\hat{t}_T := \arg\max_{y \in I_T} |\hat{J}(y)|$, we finally obtain

$$\hat{t}_B := T^{-1}(\hat{t}_T)$$

as our proposed splicing point estimator in the original scale.

3. Large-sample properties

3.1. Regularity conditions

We state the assumptions needed for our convergence results. Assumptions 1, 2 and 4 are standard for strong uniform consistency of asymmetric kernel estimators (e.g., Hirukawa et al., 2022; Funke and Hirukawa, 2025). Assumption 3 refers to the transformation.

Assumption 1. $\{X_i\}_{i=1}^n \in \mathbb{R}_+$ are *i.i.d.* random variables.

Assumption 2. **(i)** $f_Y(y)$ is uniformly bounded on $[0,1]$. **(ii)** The local structure (1) holds, $g_Y^{(2)}$ is uniformly bounded on $[0,1]$, and $g_Y^{(3)}$ is Lipschitz continuous and bounded on I_T.

Assumption 3. **(i)** T is injective with $T(0)=0$ and $T(t_M)=1/2$, where $t_M := (\underline{t}+\overline{t})/2$ is the midpoint of I_0. **(ii)** $T^{(1)}$ is Lipschitz continuous on $\mathbb{R}_+$ with $0 < \underline{T}^{(1)} \leq 1/2 \leq \overline{T}^{(1)} < 1$ on I_0.

Assumption 4. Tuning parameters b and Δ satisfy $b, \Delta \to 0$,

$$\frac{b^{3/4}}{\Delta} + \frac{\Delta}{b^{1/2+\delta_1}} + \frac{b^{1/2-4\delta_1}}{n^{1-\delta_2}\Delta^2} \to 0$$

for some small $\delta_1, \delta_2 > 0$, and $\ln n/(nb^{3/2-\kappa}) = O(1)$ for some $\kappa \in [0,1)$, as $n \to \infty$.

It follows from Assumption 4 that $b = o(\Delta)$ and $\Delta = o(b^{1/2})$, ensuring that the shift parameter shrinks to zero between the smoothing parameter b

and $b^{1/2}$. Detailed discussion and the choice $\Delta = b^{\alpha}$ with $\alpha \in (1/2, 3/4)$ are provided in the Supplement.

3.2. Consistency and asymptotic normality

The theorems below provide strong consistency and asymptotic normality of the splicing point estimator $\hat{t}_B$. Their proofs, intermediate propositions and lemmata are provided in the Supplement.

Theorem 1 (strong consistency)**.** *Under Assumptions 1–4,* $|\hat{t}_B - t_0| = O(b^{1/2+\delta_1})$ *a.s. as* $n \to \infty$.

Theorem 2 (asymptotic normality)**.** *Under Assumptions 1–4,*

$$\sqrt{\frac{n}{b^{1/2}}}\left[\hat{t}_B - t_0 - \left\{\frac{1 - T(t_0)/2}{T^{(1)}(t_0)}\right\} b\{1 + o_p(1)\}\right] \xrightarrow{d} N(0, V_B)$$

as $n \to \infty$, *where*

$$V_B = V_B(T) := \frac{3\sqrt{\pi}\sqrt{T(t_0)\{1 - T(t_0)\}}}{4d_0^2\, T^{(1)}(t_0)} \left\{\frac{f_X(t_0^-) + f_X(t_0^+)}{2}\right\}.$$

Two practical observations follow from Theorem 2. First, the leading bias depends on the unknown t_0 through $T(t_0)$ and $T^{(1)}(t_0)$, and thus unlike the gamma kernel estimator $\hat{t}_G$ by Funke and Hirukawa (2025), a simple bias correction is not available for $\hat{t}_B$. Second, the estimator $\hat{t}_B$ is super-consistent in the sense that the leading variance term is of order $b^{1/2}/n$ and thus vanishes faster than $1/n$. This property is practically important: in two-step estimation of actuarial loss distributions (e.g., Reynkens et al., 2017), the

splicing point can be estimated first without distorting subsequent inference. Similarly, goodness-of-fit tests for the fitted tail model (e.g., Clauset et al., 2009) are unaffected by the threshold estimation error. Further theoretical results, including an approximation to the mean squared error and efficiency comparisons with $\hat{t}_G$, are given in the Supplement.

4. Finite-sample performance

4.1. Design

We simulate 1000 replications with $n \in \{250, 500\}$ from three distributions, each with a common splicing point $t_0 = 4$ and a small jump size $d_0 \approx 0.05$. Model A uses a log-normal-like density augmented by a quadratic step below t_0. Models B and C splice a Weibull bulk distribution with a GPD tail and a half-normal tail, respectively. Full model specifications and characteristic numbers are given in the Supplement. The transformed observations $\{Y_i = T(X_i)\}$ are computed for four transformations T_1–T_4 in Table 1. Three choices of I_0 are examined: (i) $[3, 5]$ ($t_0 = t_M$), (ii) $[3.5, 5.5]$ ($t_0 < t_M$), and (iii) $[2.5, 4.5]$ ($t_0 > t_M$).

We compare with five automated threshold detection methods: the minimum Kolmogorov-Smirnov distance procedure [KS] by Clauset et al. (2009); minimum quantile discrepancy criteria for the mean absolute deviation and sup-norm [Q-MAD, Q-SUP] and the automated Eye-Balling method [AEB] by Danielsson et al. (2019); and the Anderson-Darling sequential testing procedure [ADST] by Bader et al. (2018). The shifted gamma kernel estimator

and its bias-corrected version implemented through the modified likelihood cross-validation (CV) [SG-ML, SG-ML-BC] by Funke and Hirukawa (2025) serve as kernel-based benchmarks. Our shifted beta estimator is implemented through the least-squares CV with $\Delta = b^{\alpha}$ for $\alpha \in \{0.55, 0.60, 0.65, 0.70\}$ [SB-LS]. Implementation details and additional CV criteria are documented in the Supplement.

4.2. Results

Table 2 presents root mean squared error (RMSE) results for $n = 500$ and the best-performing configurations. Among automated methods, Q-MAD yields the lowest RMSE overall. SB estimators dominate both SG estimators in terms of variability and RMSE across all settings, confirming that data transformation is more effective than bias correction in the original scale when the jump size is small.

Among SB estimators, T_3 (rational function) consistently produces the lowest RMSE, which can be attributed to its largest magnification factor. Case (ii), where t_0 lies in the left part of I_0, tends to give the best results; this suggests setting the interval longer on the right-tail side. The exponent $\alpha = 0.70$ outperforms other choices. The recommended configuration SB-LS-T_3 with $\alpha = 0.70$ yields the smallest RMSE across all models and dominates all competing methods. Full Monte Carlo tables with all configurations, both sample sizes, and all performance measures are provided in the Supplement.

Table 1: Examples of transformations satisfying Assumption 3.

Transformation		Magnification Factor $d_T/d_0 = 1/T^{(1)}(t_0)$	(i)	(ii)	(iii)
Arctangent:	$T_1(x) = \frac{2\arctan(x/t_M)}{\pi}$	$\frac{\pi(t_0^2+t_M^2)}{2t_M}$	12.57	12.65	12.68
Exponential CDF:	$T_2(x) = 1 - \exp\left(-\frac{x\ln 2}{t_M}\right)$	$\frac{t_M\, 2^{t_0/t_M}}{\ln 2}$	11.54	12.02	11.15
Rational:	$T_3(x) = \frac{x}{t_M(1+x/t_M)}$	$\frac{(t_0+t_M)^2}{t_M}$	16.00	16.06	16.07
Hyp. tangent:	$T_4(x) = \tanh\left(\frac{x\ln 3}{2t_M}\right)$	$\frac{t_M\{3^{t_0/(2t_M)}+3^{-t_0/(2t_M)}\}^2}{2\ln 3}$	9.71	10.31	9.23

Note: $t_M = (\underline{t} + \bar{t})/2$. Cases (i)–(iii) correspond to $I_0 = [3, 5]$, $[3.5, 5.5]$, and $[2.5, 4.5]$ with $t_0 = 4$.

Table 2: Monte Carlo RMSE for $n = 500$ (selected configurations).

Estimator	Model A	Model B	Model C
Automated methods			
KS	0.889	1.174	0.907
Q-MAD	0.442	0.494	0.460
Q-SUP	0.992	1.568	1.078
AEB	0.712	1.893	0.642
ADST	1.475	1.269	1.535
Kernel-based, Case (ii): $I_0 = [3.5, 5.5]$			
SG-ML ($\alpha = 0.70$)	0.439	0.630	0.645
SG-ML-BC	0.400	0.609	0.618
SB-LS-T_3 ($\alpha = 0.70$)	**0.212**	**0.138**	**0.088**

Note: Bold values indicate the smallest RMSE in each column. Full tables with all configurations and sample sizes are in the Supplement.

5. Empirical applications

We apply SB-LS-T_3 with $\alpha = 0.70$ to four non-life insurance datasets, namely, (A) Danish fire insurance losses, which have been analyzed most popularly since the seminal work by McNeal (1997), (B) Norwegian fire losses, (C) Belgian motor losses, and (D) French motor losses; see the Supplement for data descriptions and summary statistics. Table 3 extracts the results alongside competing methods. Complete results are available in the Supplement.

Table 3: Real data: splicing point estimates.

Data	Method	$\hat{t}_0$	Method	I_0	$\hat{t}_0$
(A)	KS	1.375	SG-ML	$[1, 30]$	1.861
	Q-MAD	29.037	SG-ML-BC		2.096
	ADST	1.406	SB-LS-T_3	$[1, 30]$	1.808
(B)	KS	2.221	SG-ML	$[2, 50]$	2.756
	Q-MAD	35.794	SG-ML-BC		2.924
	ADST	2.121	SB-LS-T_3	$[2, 50]$	2.702
(C)	KS	3.233	SG-ML	$[2, 40]$	$40.000^{\ddagger}$
	Q-MAD	3.022	SB-LS-T_3	$[2, 40]$	2.435
	ADST	$29.451^{\dagger}$			
(D)	KS	1.318	SG-ML	$[1, 20]$	$20.000^{\ddagger}$
	Q-MAD	3.204	SB-LS-T_3	$[1, 20]$	11.247
	ADST	$37.149^{\dagger}$			

Note: † ADST rejects all candidates; ‡ SG-ML boundary failure.

For datasets (A) and (B), SB-LS-T_3 yields estimates close to those of KS, ADST, and SG-ML-BC, with no estimation problems. For datasets (C) and (D), however, ADST rejects all threshold candidates and SG-ML fails to find a splicing point inside I_0. In contrast, SB-LS-T_3 continues to produce stable estimates. These results illustrate that our method provides a reliable complement to existing approaches, particularly in challenging settings where automated or original-scale methods break down.

6. Conclusion

We have proposed a nonparametric method for estimating splicing points in actuarial loss distributions that transforms the original data onto $[0, 1]$ and applies the asymmetric beta kernel. The transformation magnifies jump discontinuities in the density, making threshold detection feasible even in sparse tail regions with small jumps. Strong consistency and asymptotic normality with a super-consistent convergence rate have been established.

Monte Carlo simulations confirm that the recommended configuration SB-LS-T_3 with $\alpha = 0.70$ outperforms both automated EVT-based methods and the gamma kernel estimator in the original scale. Applications to four non-life insurance datasets demonstrate the method's practical robustness, including settings where competing approaches fail.

Data availability

All datasets used are openly available; see the Supplement for details.

Declaration of interest

The authors report that there are no competing interests to declare.

Funding

This work was supported by the Japan Society for the Promotion of Science under the grant number 23K01340 (M. Hirukawa).

References

[1] Bader, B., Yan, J., Zhang, X., 2018. Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. Ann. Appl. Stat. 12, 310-329.

[2] Chen, S.X., 1999. Beta kernel estimators for density functions. Comput. Statist. Data Anal. 31, 131-145.

[3] Chen, S.X., 2000. Probability density function estimation using gamma kernels. Ann. Inst. Statist. Math. 52, 471-480.

[4] Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-law distributions in empirical data. SIAM Rev. 51, 661-703.

[5] Danielsson, J., Ergun, L.M., de Haan, L., de Vries, C.G., 2019. Tail index estimation: Quantile-driven threshold selection. Bank of Canada Staff Working Paper 2019-28.

[6] DuMouchel, W.H., 1983. Estimating the stable index α in order to measure tail thickness: A critique. Ann. Statist. 11, 1019-1031.

[7] Funke, B., Hirukawa, M., 2025. Nonparametric estimation of splicing points in skewed cost distributions: A kernel-based approach. J. Nonparametr. Stat., forthcoming.

[8] Hirukawa, M., Murtazashvili, I., Prokhorov, A., 2022. Uniform convergence rates for nonparametric estimators smoothed by the beta kernel. Scand. J. Stat. 49, 1353-1382.

[9] Klugman, S.A., Panjer, H.H., Willmot, G.E., 2019. Loss Models: From Data to Decisions, 5th ed. Wiley, Hoboken, NJ.

[10] McNeal, A.J., 1997. Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bull. 27, 117-137.

[11] Reynkens, T., Verbelen, R., Beirlant, J., Antonio, K., 2017. Modelling censored losses using splicing: A global fit strategy with mixed Erlang and extreme value distributions. Insurance Math. Econom. 77, 65-77.

[12] Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT Stat. J. 10, 33-60.