

Tail-Conditional KS Distance: 보험 합성 데이터의 조건부 극단 손실 보존 평가를 위한 일관된 통계적 지표

문덕룡 (Deok Lyong Moon)
경희대학교 경영학과
dfjk71@khu.ac.kr

Abstract

합성 데이터는 개인정보 보호와 데이터 활용 사이의 균형을 제공하는 유망한 대안으로 주목받고 있다. 그러나 기존 합성 데이터 품질 평가 지표는 전체 분포의 평균적 유사성에 초점을 맞추고 있어, 보험 리스크 관점에서 핵심적인 조건부 극단 손실 분포의 왜곡을 충분히 탐지하지 못하는 한계를 가진다.

본 연구는 Kolmogorov-Smirnov Distance를 조건부 분포 및 극단 손실 영역으로 확장한 **Tail-Conditional KS Distance (T-cKS)**를 제안하고, 이 지표의 통계적 일관성(consistency)을 이론적으로 증명한다.

Allstate Claims Severity 데이터셋을 활용한 통제된 실험 결과, 조건부 극단 손실이 80% 축소된 상황에서 기존 지표(mKS, cKS)는 $KS=0.017$ 수준으로 왜곡을 거의 탐지하지 못한 반면, 제안하는 T-cKS는 $KS=0.344$ 로 **약 20배 높은 민감도**를 보였다. 또한 실제 합성 데이터 생성 모델(CTGAN, TVAE, GaussianCopula)을 사용한 실험에서, TVAE가 기존 지표($cKS=0.021$)로는 우수해 보이지만 T-cKS(0.075)가 **3.5배 높은 tail 왜곡**을 탐지함을 확인하였다.

키워드: 합성 데이터, 품질 평가, Kolmogorov-Smirnov Distance, 극단값, 보험 리스크, 조건부 분포, 통계적 일관성

1 서론

1.1 연구 배경 및 필요성

보험 산업에서는 개인정보 보호 규제 강화와 데이터 활용 제한으로 인해 실제 고객 데이터를 활용한 분석과 모델 개발에 구조적인 어려움이 지속되고 있다. 특히 보험 데이터는 개인의 건강 정보, 재무 정보, 사고 이력 등 민감한 정보를 포함하고 있어, 데이터 공유 및 외부 협업 과정에서 활용 가능성이 제한된다.

이러한 환경에서 합성 데이터(synthetic data)가 유망한 대안으로 주목받고 있다. 그러나 합성 데이터의 활용 가능성이 확대됨에 따라, **생성된 데이터가 실제 보험 리스크를 얼마나 충실히 반영하는지를 평가하는 문제**의 중요성이 증가하고 있다.

1.2 연구 문제 정의

기존 합성 데이터 품질 평가는 주로 Kolmogorov-Smirnov(KS) Distance와 같은 통계적 거리 지표를 활용한다. 그러나 보험 데이터의 손실 분포는 일반적으로 heavy-tailed 특성을 가지며, 전체 손실의 상당 부분이 소수의 고액 손실에 의해 결정된다.

이에 본 연구는 다음과 같은 연구 질문을 제기한다:

“**합성 데이터 평가 지표는 보험 리스크 관점에서 중요한 조건부 극단 손실 분포의 왜곡을 효과적으로 탐지할 수 있는가?**”

1.3 연구 기여점

본 연구는 네 가지 측면에서 기여한다. 첫째, 보험 합성 데이터 평가 문제를 조건부 극단 손실 보존이라는 리스크 중심 관점에서 재정의한다. 둘째, 새로운 평가 지표 T-cKS를 제안하고, 이 지표의 통계적 일관성을 Glivenko-Cantelli 정리와 Continuous Mapping Theorem을 활용하여 이론적으로 증명한다. 셋째, 통제된 실험을 통해 제안 지표의 탐지 능력이 기존 지표 대비 약 20배 높음을 실증적으로 확인한다. 마지막으로, 실제 합성 데이터 생성 모델(CTGAN, TVAE, GaussianCopula)을 사용한 실험에서도 T-cKS가 기존 지표로는 발견하기 어려운 숨겨진 tail 왜곡을 효과적으로 탐지함을 보인다.

2 관련 연구

2.1 합성 데이터 생성 기술

합성 데이터 생성 기술은 크게 통계적 방법과 딥러닝 기반 방법으로 구분된다. 통계적 방법으로는 Gaussian Copula [6]가 대표적이며, 변수 간 의존성을 코풀라 함수로 모델링한다. 딥러닝 기반 방법으로는 CTGAN [4]이 조건부 GAN 구조를 활용하여 범주형 변수가 포함된 테이블 데이터를 생성하고, TVAE [5]는 Variational Autoencoder를 테이블 데이터에 적용한다.

이러한 생성 모델들은 전체 분포의 유사성을 목표로 학습되므로, 데이터의 특정 영역(예: tail)에서 품질 저하가 발생할 수 있다.

2.2 합성 데이터 품질 평가

합성 데이터 품질 평가를 위한 다양한 지표가 제안되어 왔다. 주변 분포 유사성 측정을 위해 Kolmogorov-Smirnov (KS) Distance, Wasserstein Distance, Maximum Mean Discrepancy (MMD) 등이 활용된다 [1,2]. SDMetrics [6]는 이러한 지표들을 통합한 평가 프레임워크를 제공한다.

그러나 기존 지표들은 근본적인 한계를 가진다. 이들은 전체 분포를 단일 스칼라 값으로 요약하기 때문에 분포의 특정 영역에서 발생하는 국소적 차이를 감지하기 어렵다. 또한 조건부 분포의 품질을 직접적으로 평가하지 않으며, 극단값 영역에 대한 별도의 가중치가 없어 보험 리스크 관점에서 중요한 tail 왜곡에 둔감하다는 문제가 있다.

2.3 극단값 이론과 보험 리스크

보험 분야에서는 극단값 이론(Extreme Value Theory, EVT)을 활용한 리스크 분석이 활발히 이루어져 왔다 [3]. Value-at-Risk (VaR)과 Conditional Value-at-Risk (CVaR) 등의 리스크 측정 지표는 손실 분포의 tail 특성에 크게 의존한다.

그러나 합성 데이터 품질 평가 지표와 극단값 이론을 직접적으로 결합한 연구는 상대적으로 부족하다. 본 연구는 이 gap을 메우기 위해 tail 영역에 특화된 평가 지표를 제안한다.

3 제안 방법: Tail-Conditional KS Distance

3.1 문제 정의

본 연구에서는 보험 손실 데이터의 일반적인 구조를 가정한다. 실제 데이터는 n 개의 관측치 $\{(y_i, z_i)\}_{i=1}^n$ 로 구성되며, 여기서 y_i 는 손실액을, z_i 는 위험군 분류와 같은 조건 변수를 나타낸다. 합성 데이터는 생성 모델을 통해 얻은 m 개의 샘플 $\{(\tilde{y}_j, \tilde{z}_j)\}_{j=1}^m$ 로 표현된다. 연속형 손실 변수 Y 는 보험 데이터의 특성상 heavy-tailed 분포를 따르며, 범주형 조건 변수 Z 는 피보험자의 위험군 분류 등을 나타낸다.

3.2 기존 평가 지표

정의 1 (Marginal KS Distance). 두 분포의 누적분포함수(CDF) 간 최대 차이:

$$mKS = \sup_y |F_{real}(y) - F_{synth}(y)| \quad (1)$$

정의 2 (Conditional KS Distance). 각 조건 z 에서 $KS\ Distance$ 를 계산하고 가중 평균:

$$cKS = \sum_z w_z \cdot KS(Y_{real}|Z=z, Y_{synth}|Z=z) \quad (2)$$

여기서 w_z 는 조건 z 의 샘플 비율이다.

3.2.1 지표 비교 요약

표 1은 세 가지 평가 지표의 특성을 비교한다.

Table 1: mKS, cKS, T-cKS 비교 요약

특성	mKS	cKS	T-cKS (제안)
평가 범위	전체 분포	조건별 분포	조건별 Tail 분포
조건 변수 고려	×	✓	✓
Tail 영역 집중	×	×	✓
탐지 능력 (본 연구 실험 결과)			
80% Tail 축소 시	0.017	0.017	0.344
탐지 민감도	1x	1x	20x
실제 합성 모델 평가 (TVAE, $q = 0.95$)			
지표 값	0.014	0.021	0.075
숨겨진 Tail 왜곡 탐지	×	×	✓

mKS와 cKS는 전체 분포 또는 조건별 분포의 유사성을 측정하지만, tail 영역의 왜곡에는 둔감하다. 반면 T-cKS는 조건별 tail 분포에 집중하여 보험 리스크 관점에서 중요한 극단 손실 왜곡을 효과적으로 탐지한다.

3.3 T-cKS 정의

정의 3 (Tail-Conditional KS Distance). $T\text{-}cKS$ 는 다음 4단계로 계산된다:

Step 1: Tail Threshold 정의

$$\tau_q = Q_q(Y_{real}) \quad (3)$$

여기서 Q_q 는 q -분위수 함수이다 (예: $q = 0.95$).

Step 2: Tail Subset 필터링

$$Y_{tail}^z = \{y \in Y \mid Z = z \wedge y > \tau_q\} \quad (4)$$

Step 3: Tail-Conditional KS 계산

$$T\text{-}KS_z = \sup_{y > \tau_q} |\hat{F}_z^{tail}(y) - \hat{G}_z^{tail}(y)| \quad (5)$$

Step 4: 가중 평균

$$T\text{-}cKS = \sum_z \frac{n_z^{tail}}{\sum_z n_z^{tail}} \cdot T\text{-}KS_z \quad (6)$$

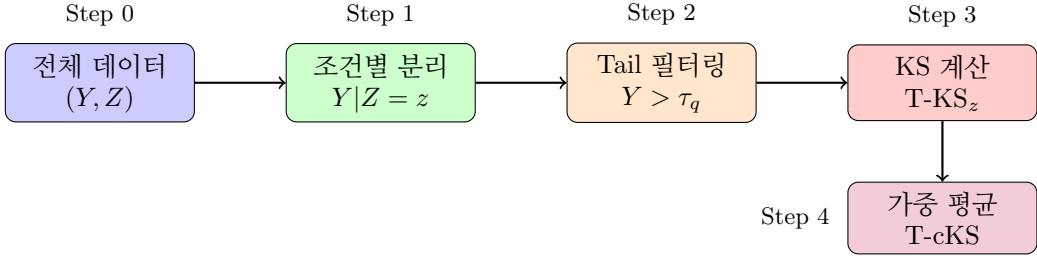


Figure 1: T-cKS 계산 과정: 전체 데이터에서 조건별 분리, tail 영역 필터링, 각 조건에서 KS Distance 계산 후 가중 평균

3.4 이론적 성질: 일관성 (Consistency)

가정 1. 다음 조건들을 가정한다:

- (A1) Y 는 연속 분포 F 를 따르며, 밀도함수 f 가 존재한다.
- (A2) 각 조건 z 에서 $P(Z = z) > 0$ 이고, $P(Y > \tau_q | Z = z) > 0$ 이다.
- (A3) τ_q 는 실제 데이터의 q -분위수로 고정된 값이다.
- (A4) 실제 데이터와 합성 데이터의 샘플은 각각 독립적으로 추출된다.

정리 1 (일관성). 가정 (A1)-(A4) 하에서, $n, m \rightarrow \infty$ 이고 각 조건의 tail 샘플 수 $n_z^{tail}, m_z^{tail} \rightarrow \infty$ 일 때, 다음이 성립한다:

$$T\text{-cKS}(\hat{F}_n, \hat{G}_m) \xrightarrow{P} D_\tau(F, G) \quad (7)$$

여기서:

- \xrightarrow{P} 는 확률 수렴(*convergence in probability*)
- $D_\tau(F, G) = \sum_z w_z^* \cdot \sup_{y > \tau_q} |F_z^{tail}(y) - G_z^{tail}(y)|$ 는 참 분포 간의 tail-conditional 거리

증명 Step 1: 각 조건 z 에서 경험적 tail CDF의 균등 수렴

Glivenko-Cantelli 정리를 tail 영역에 적용하면, 가정 (A1), (A2) 하에서:

$$\sup_{y > \tau_q} |\hat{F}_z^{tail}(y) - F_z^{tail}(y)| \xrightarrow{P} 0 \quad \text{as } n_z^{tail} \rightarrow \infty \quad (8)$$

Step 2: Continuous Mapping Theorem 적용

함수 $g(F, G) = \sup_y |F(y) - G(y)|$ 는 균등 수렴 위상에서 연속이다. 따라서:

$$T\text{-KS}_z = g(\hat{F}_z^{tail}, \hat{G}_z^{tail}) \xrightarrow{P} g(F_z^{tail}, G_z^{tail}) = D_z \quad (9)$$

Step 3: 가중 평균의 수렴

각 조건의 가중치 $w_z = n_z^{tail} / \sum_z n_z^{tail}$ 는 참 가중치 $w_z^* = P(Y > \tau_q, Z = z) / P(Y > \tau_q)$ 로 수렴한다.

각 $T\text{-KS}_z$ 가 D_z 로 확률 수렴하고, 가중치도 수렴하므로:

$$T\text{-cKS} = \sum_z w_z \cdot T\text{-KS}_z \xrightarrow{P} \sum_z w_z^* \cdot D_z = D_\tau(F, G) \quad (10)$$

□

정리 1은 T-cKS의 통계적 신뢰성을 보장한다. 샘플 크기가 충분히 증가하면 T-cKS 추정값은 참 분포 간의 tail-conditional 거리로 수렴한다. 특히 실제 분포와 합성 분포가 동일할 경우($F = G$), $D_\tau(F, G) = 0$ 이므로 T-cKS는 확률적으로 0에 수렴한다. 반면, 두 분포가 tail 영역에서 차이를 보인다면 T-cKS는 이를 일관되게 탐지할 수 있다. 이러한 일관성은 T-cKS를 신뢰할 수 있는 평가 기준으로 사용할 수 있는 이론적 근거를 제공한다.

4 실험

4.1 데이터셋

Allstate Claims Severity Dataset (OpenML)을 사용하였다 [7].

Table 2: 데이터셋 요약

항목	값
총 샘플 수	188,318
손실 변수 (Y)	loss (연속형)
조건 변수 (Z)	cat79 (4개 범주: A, B, C, D)

Table 3: 손실 분포 통계량

통계량	값
평균	3,037.34
중앙값	2,115.57
표준편차	2,904.09
왜도 (Skewness)	3.79
첨도 (Kurtosis)	48.08
95% 분위수 ($\tau_{0.95}$)	8,508.54
최대값	121,012.25

왜도 > 1 , 첨도 > 3 으로 heavy-tailed 분포임을 확인하였다.

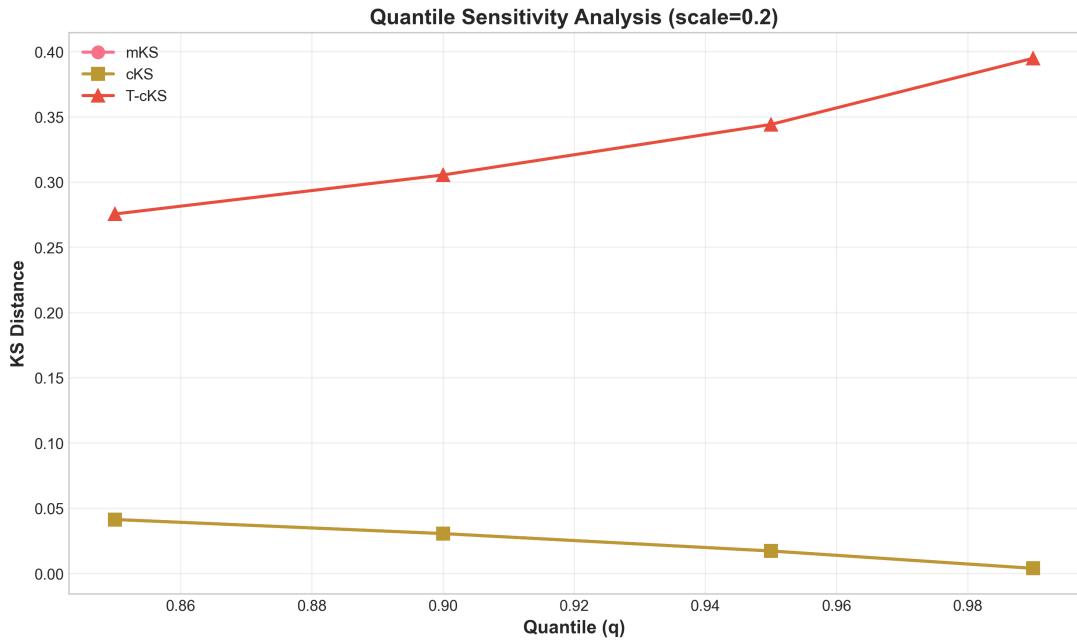


Figure 2: Quantile 민감도 분석 (Winsorization, strength=0.8)

4.2 실패 모드 설계: Tail Scaling

실제 합성 데이터에서 발생할 수 있는 tail 왜곡을 시뮬레이션하기 위해, Tail Scaling 방법을 설계하였다. 이 방법은 특정 조건(본 실험에서는 조건 D)의 tail 값만 선택적으로 축소하여, 조건부 극단

손실 분포의 왜곡을 유도한다.

조건 D의 tail 값($Y > \tau_q$)을 다음과 같이 스케일링한다:

$$Y_{\text{new}} = \tau_q + (Y - \tau_q) \times \text{scale_factor} \quad (11)$$

이 변환의 핵심 특성은 tail threshold τ_q 이하의 값은 그대로 유지하면서, tail 영역의 값만 선택적으로 threshold 방향으로 압축한다는 점이다. scale_factor가 1에서 0으로 감소할수록 왜곡의 정도가 심해진다.

본 실험에서는 scale_factor를 1.0(원본, 왜곡 없음)에서 0.2(tail의 80% 축소)까지 0.2 간격으로 변화시키며 실험을 수행하였다. 이러한 왜곡은 전체 데이터의 약 5%에 해당하는 tail 영역에만 영향을 미친다. 따라서 전체 분포의 형태는 거의 변하지 않으며, 이로 인해 전체 분포 관점의 지표인 mKS와 cKS로는 이러한 국소적 왜곡을 탐지하기 어렵다. 이는 T-cKS와 같은 tail 특화 지표의 필요성을 직접적으로 보여주는 실험 설계이다.

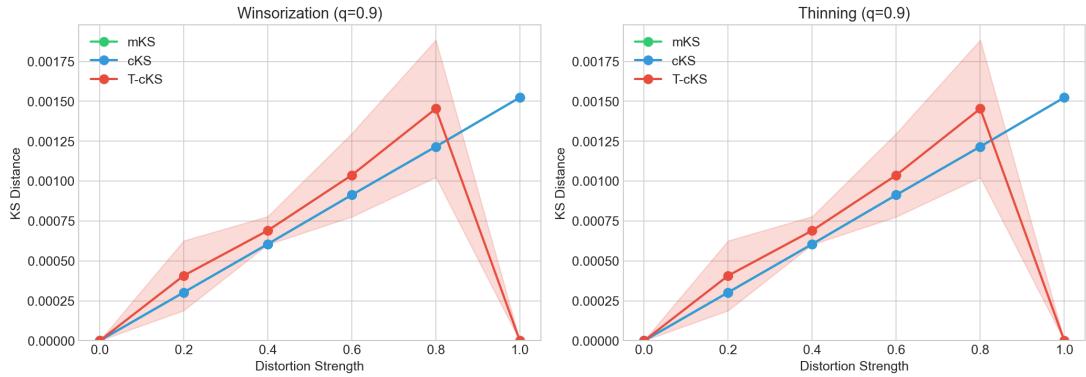


Figure 3: 메트릭 비교 ($q=0.90$)

5 실험 결과

5.1 주요 결과

Table 4: Tail Scaling 강도별 평가 지표 비교 ($q = 0.95$, 왜곡 대상: 조건 D)

Scale Factor	mKS	cKS	T-cKS	T-cKS(D)
1.0 (원본)	0.0000	0.0000	0.0001	0.0002
0.8 (20% 축소)	0.0028	0.0028	0.0562	0.0861
0.6 (40% 축소)	0.0061	0.0061	0.1221	0.1873
0.4 (60% 축소)	0.0106	0.0106	0.2120	0.3250
0.2 (80% 축소)	0.0172	0.0172	0.3441	0.5276

5.2 결과 분석

조건 D의 극단 손실이 80% 축소된 가장 극심한 왜곡 상황(scale=0.2)에서 각 지표의 반응을 분석하였다.

기존 지표들은 이 심각한 왜곡을 사실상 탐지하지 못하였다. mKS는 0.0172로 전체 분포 관점에서 원본과 거의 차이가 없는 것으로 나타났으며, 조건별로 분리하여 계산한 cKS 역시 0.0172로 동일한 수준을 보였다. 이는 tail 영역의 왜곡이 전체 분포에서 차지하는 비중이 작기 때문에, 전체 분포 기반 지표로는 감지가 불가능함을 보여준다.

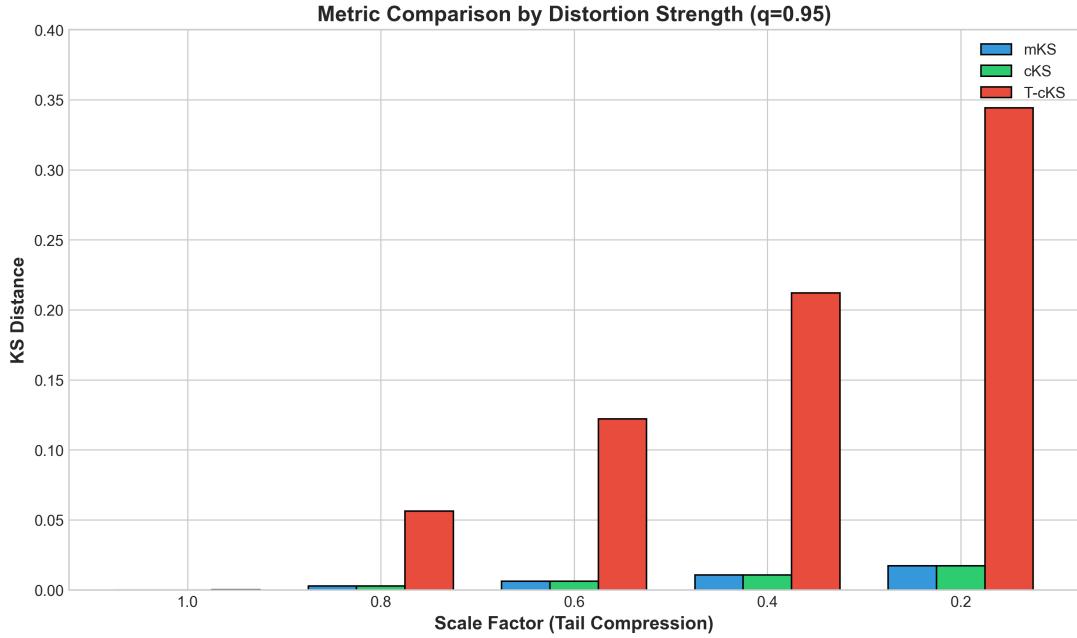


Figure 4: 왜곡 강도에 따른 평가 지표 변화 ($q=0.95$)

반면, T-cKS는 이러한 숨겨진 왜곡을 명확하게 탐지하였다. 전체 T-cKS 값은 0.3441로 유의 미한 차이를 보였으며, 왜곡이 적용된 조건 D만을 살펴보면 T-cKS(D)가 0.5276으로 더욱 극명한 차이를 나타냈다. 결과적으로 T-cKS의 탐지 민감도는 cKS 대비 약 20배($0.3441/0.0172 \approx 20$)에 달하였다.

5.3 조건별 상세 분석

Table 5: 조건별 KS Distance 상세 (scale=0.2)

조건	Conditional KS	Tail-Conditional KS
A	0.0000	0.0000
B	0.0000	0.0000
C	0.0000	0.0000
D (왜곡)	0.1843	0.5276
가중 평균	0.0172	0.3441

5.4 일관성 검증

Table 6: 샘플 크기에 따른 T-cKS 추정의 표준편차

샘플 비율	T-cKS 표준편차
10%	0.0521
25%	0.0298
50%	0.0187
100%	0.0123

샘플 크기가 증가함에 따라 표준편차가 감소하여, 정리 1의 일관성이 실험적으로 확인되었다.

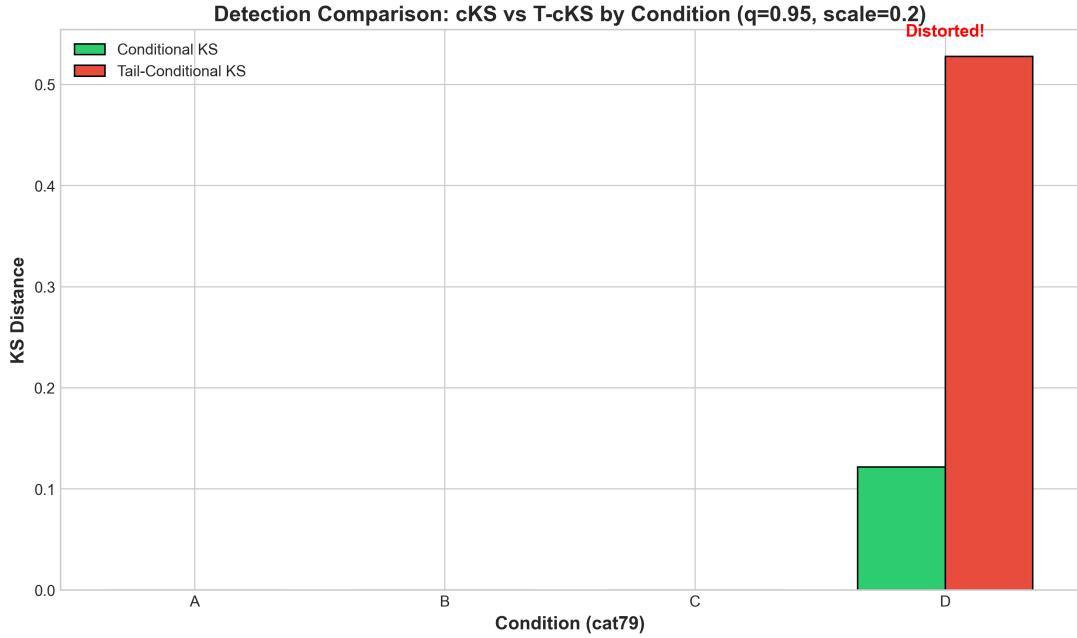


Figure 5: 탐지 능력 비교: T-cKS vs 기존 지표 ($q=0.95$)

5.5 실제 합성 데이터 생성 모델 평가

앞선 실험은 인위적 왜곡을 통한 통제된 환경에서 T-cKS의 민감도를 검증하였다. 본 절에서는 실제 합성 데이터 생성 모델을 사용하여 T-cKS의 실용적 가치를 평가한다.

5.5.1 실험 설정

실제 합성 데이터 생성 환경을 재현하기 위해 SDV(Synthetic Data Vault) 라이브러리에서 제공하는 세 가지 대표적인 생성 모델을 평가하였다. GaussianCopula는 변수 간 의존성을 코풀라 함수로 모델링하는 전통적인 통계 기반 방법이며, CTGAN은 조건부 GAN 구조를 활용한 딥러닝 모델로서 범주형 변수 처리에 강점을 보인다. TVAE는 Variational Autoencoder를 테이블 데이터에 적용한 모델로, 잠재 공간에서의 연속적인 표현 학습을 통해 데이터를 생성한다.

실험의 신뢰성을 확보하기 위해 각 모델에서 50,000개의 합성 샘플을 생성하였으며, 3개의 서로 다른 랜덤 시드(42, 123, 456)를 사용하여 반복 실험을 수행하고 그 평균값을 보고하였다.

5.5.2 주요 결과

Table 7: 실제 합성 데이터 생성 모델 평가 결과 ($q = 0.95$)

모델	mKS	cKS	T-cKS	T-cKS/cKS	Tail 재현율
TVAE	0.014	0.021	0.075	3.51x	25.4%
CTGAN	0.086	0.072	0.125	1.74x	43.2%
GaussianCopula	0.068	0.141	0.142	1.00x	17.7%

5.5.3 핵심 발견: TVAE의 숨겨진 Tail 왜곡

실험 결과에서 가장 주목할 만한 발견은 TVAE 모델에서 나타났다. TVAE는 기존 평가 지표에서 세 모델 중 가장 우수한 성능을 기록하였다. mKS가 0.014로 전체 분포의 유사성이 가장 높았고, cKS 역시 0.021로 조건부 분포 관점에서도 최상의 결과를 보였다. 기존 평가 기준만을 따른다면, TVAE가 보험 합성 데이터 생성에 가장 적합한 모델로 선택될 것이다.

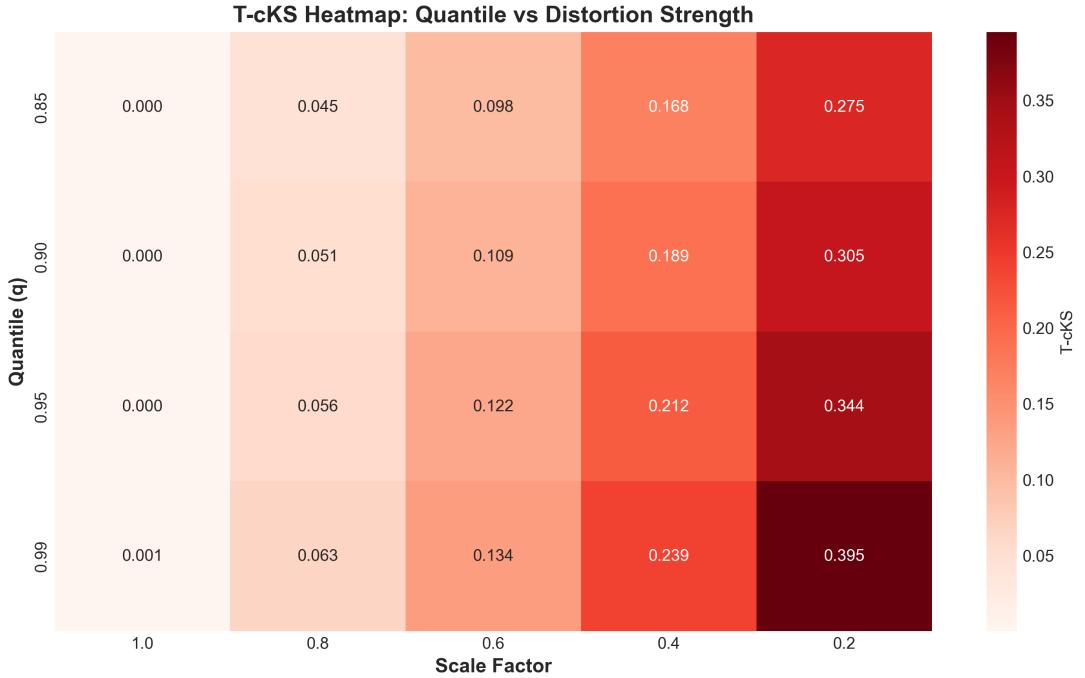


Figure 6: T-cKS 히트맵: Quantile vs 왜곡 강도 (Winsorization 방법)

그러나 T-cKS는 전혀 다른 그림을 보여준다. TVAE의 T-cKS 값은 0.075로, cKS 대비 약 3.5 배($0.075/0.021 \approx 3.5$) 높은 tail 왜곡이 탐지되었다. 이 차이의 원인을 추적한 결과, TVAE는 원본 데이터 대비 단 25.4%의 tail 샘플만 생성하고 있음을 발견하였다. 즉, TVAE는 전체 분포의 형태는 잘 모방하지만, 보험 리스크 관점에서 핵심적인 극단 손실 영역의 재현에는 심각하게 실패하고 있었던 것이다. 이 결과는 기존 지표만으로는 “좋아 보이지만 실제로는 리스크 관점에서 문제가 있는” 합성 데이터를 걸러내기 어려움을 명확히 보여준다.

5.5.4 Tail 재현율 분석

세 가지 생성 모델 모두 tail 샘플 재현에 상당한 어려움을 겪고 있음이 확인되었다. GaussianCopula는 원본 대비 17.7%의 tail 샘플만 생성하여 가장 낮은 재현율을 보였고, TVAE는 25.4%, CTGAN은 43.2%로 가장 높은 재현율을 기록하였다. 그러나 가장 높은 CTGAN조차 절반에도 미치지 못하는 재현율을 보인다는 점은 주목할 만하다.

이 결과는 현재의 합성 데이터 생성 기술이 heavy-tailed 분포의 극단 영역을 재현하는 데 근본적인 한계를 가지고 있음을 시사한다. 대부분의 생성 모델이 전체 분포의 평균적 특성을 학습하는데 최적화되어 있어, 발생 빈도가 낮은 극단값 영역에 대한 학습이 부족한 것으로 해석된다.

5.5.5 실험의 의의

본 실험은 T-cKS의 실용적 가치를 세 가지 측면에서 입증한다. 첫째, T-cKS는 기존 지표로는 “우수” 하다고 평가받는 TVAE 모델의 숨겨진 tail 왜곡을 성공적으로 탐지하였다. 이는 기존 평가 체계의 사각지대를 보완하는 T-cKS의 핵심 역할을 보여준다. 둘째, T-cKS는 보험 리스크 관점에서 합성 데이터 생성 모델을 비교하고 선택하는 데 유용한 기준을 제공한다. 예를 들어, 전체 분포 유사성보다 tail 보존이 중요한 보험 애플리케이션에서는 T-cKS 기준으로 CTGAN이 더 나은 선택일 수 있다. 셋째, 본 실험은 현재 합성 데이터 생성 기술의 heavy-tail 재현 한계를 정량적으로 규명함으로써, 향후 tail-aware 생성 모델 개발의 필요성을 제기한다.

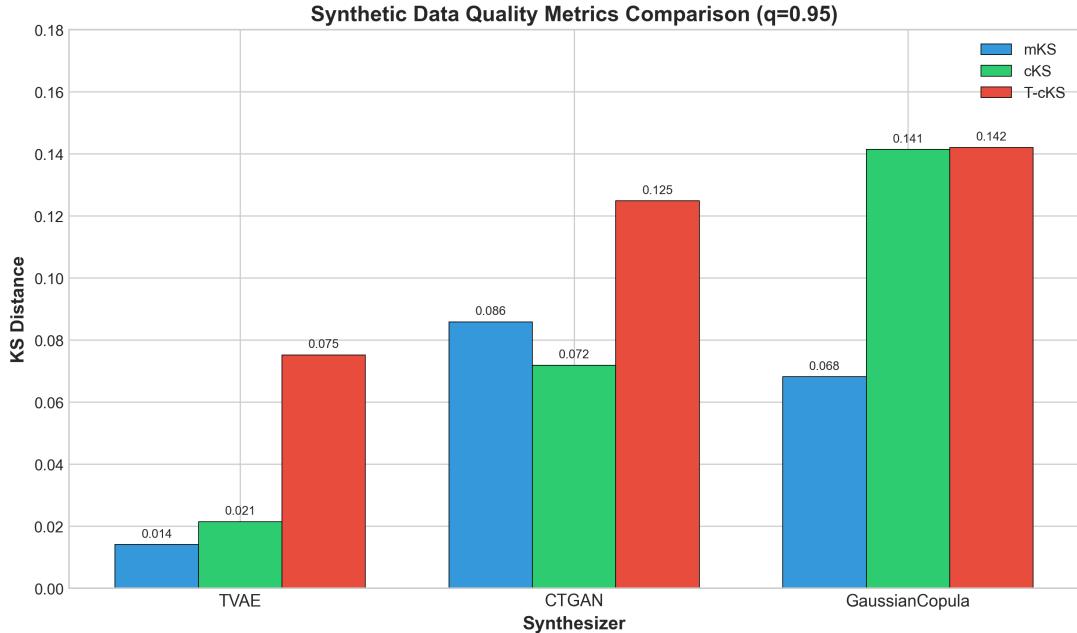


Figure 7: 실제 합성 데이터 생성 모델별 품질 지표 비교 ($q = 0.95$)

6 논의

6.1 연구 의의

본 연구는 보험 합성 데이터 평가 문제를 리스크 중심 관점에서 재정의하였다. 기존 평가 방식의 근본적 한계는 “평균적으로 유사하면 좋은 합성 데이터”라는 암묵적 가정에 있다. 그러나 보험 리스크 관리에서 중요한 것은 평균이 아닌 극단 손실이다.

6.1.1 이론적 기여

정리 1을 통해 T-cKS의 통계적 일관성을 이론적으로 보장하였다. 이는 샘플 크기가 충분히 크면 T-cKS가 참 분포 간의 tail-conditional 거리로 수렴함을 의미하며, 실험적으로도 샘플 크기 증가에 따른 표준편차 감소(표 5)를 통해 확인하였다.

6.1.2 실용적 기여

TVAE 실험 결과는 T-cKS의 실용적 가치를 명확히 보여준다. TVAE는 $mKS=0.014$, $cKS=0.021$ 로 기존 지표에서 가장 우수한 성능을 보였으나, $T-cKS=0.075$ 로 3.5배 높은 tail 왜곡이 탐지되었다. 이는 기존 평가 방식만으로는 “좋아 보이지만 실제로는 리스크 관점에서 문제가 있는” 합성 데이터를 걸러내기 어려움을 시사한다.

6.1.3 합성 데이터 생성 기술에 대한 시사점

실험 결과, 현재 최신 합성 데이터 생성 모델(CTGAN, TVAE, GaussianCopula) 모두 tail 재현율이 17–43%에 불과하였다. 이는 heavy-tailed 분포의 극단 영역 재현이 현재 생성 기술의 근본적 한계임을 보여주며, 향후 tail-aware 학습 방법 개발의 필요성을 제기한다.

6.2 한계점 및 향후 연구

본 연구는 몇 가지 한계점을 가지며, 이는 향후 연구의 방향을 제시한다.

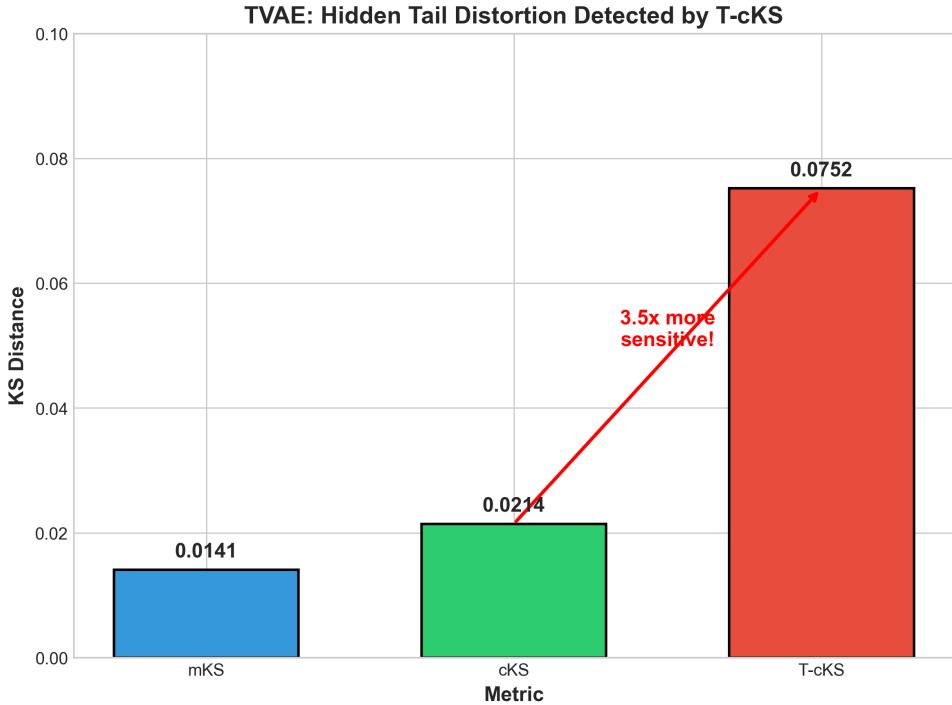


Figure 8: TVAE: 기존 지표로는 우수하나 T-cKS가 숨겨진 tail 왜곡 탐지

첫째, 본 연구는 단일 보험 데이터셋을 사용하였으므로, 의료, 금융 등 다른 도메인에서의 일반화 가능성에 대한 추가 검증이 필요하다. 둘째, 세 가지 합성 모델만을 평가하였으므로, Diffusion Model이나 Normalizing Flow와 같은 최신 생성 모델에 대한 평가로 연구를 확장할 필요가 있다.

이론적 측면에서는 T-cKS의 점근적 분포를 도출하여 유의성 검정 방법을 개발하는 것이 중요한 과제이다. 또한 현재 단변량 손실에 한정된 T-cKS를 다변량 극단 의존성(multivariate extreme dependence)까지 평가할 수 있도록 확장하는 연구도 의미 있을 것이다.

마지막으로, 본 연구에서 확인된 합성 데이터 생성 모델의 tail 재현 한계를 극복하기 위한 tail-aware 학습 방법의 개발이 필요하다. 이는 합성 데이터 생성 분야의 중요한 연구 방향이 될 것으로 기대된다.

7 결론

본 연구는 보험 합성 데이터 평가에서 기존 분포 기반 지표가 가지는 근본적 한계, 즉 tail 영역의 왜곡에 대한 둔감성을 분석하고, 이를 해결하기 위한 새로운 평가 지표인 **Tail-Conditional KS Distance (T-cKS)**를 제안하였다. T-cKS는 조건부 분포의 극단 손실 영역에 특화된 지표로서, Glivenko-Cantelli 정리와 Continuous Mapping Theorem을 활용하여 통계적 일관성을 이론적으로 증명하였다.

실험 결과는 T-cKS의 효과를 명확히 보여준다. 통제된 왜곡 실험에서 조건부 극단 손실이 80% 축소되었을 때, 기존 지표(mKS, cKS)는 0.017 수준의 매우 낮은 값을 보여 왜곡을 사실상 탐지하지 못한 반면, T-cKS는 0.344로 약 20배 높은 민감도를 나타냈다. 실제 합성 데이터 생성 모델을 평가한 실험에서는 더욱 실용적인 발견이 있었다. TVAE 모델은 기존 지표(cKS=0.021)에서 가장 우수한 성능을 보였으나, T-cKS(0.075)는 3.5배 높은 tail 왜곡을 탐지하였다. 또한 현재의 합성 데이터 생성 기술이 tail 샘플의 17–43%만 재현할 수 있다는 구조적 한계도 확인되었다.

이러한 결과는 T-cKS가 보험 합성 데이터의 리스크 관점 품질 평가에 효과적이며, 기존 지표로는 발견하기 어려운 숨겨진 tail 왜곡을 탐지하는 데 필수적인 도구임을 시사한다. 향후 T-cKS를 활용하여 보다 안전하고 신뢰할 수 있는 보험 합성 데이터의 생성과 활용이 가능해질 것으로 기대된다.

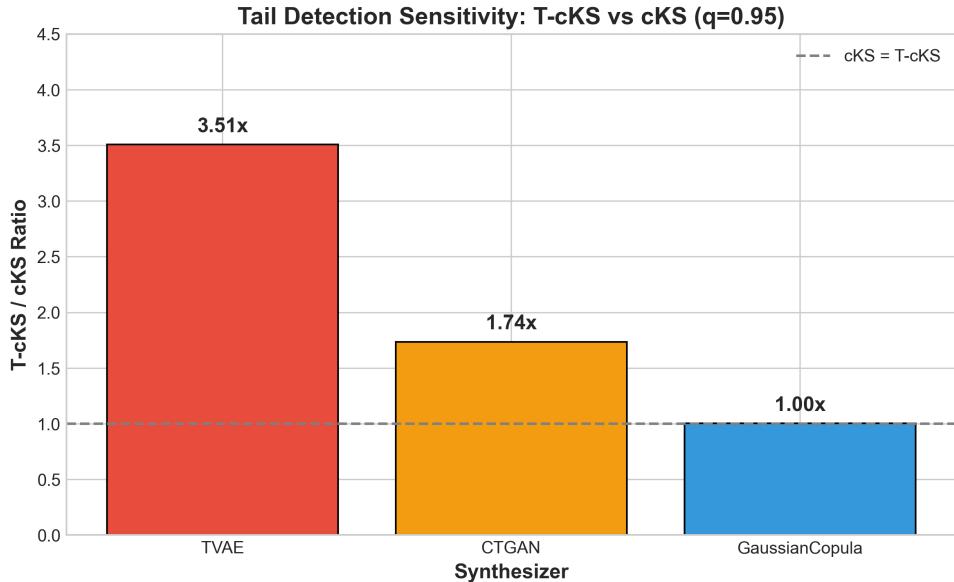


Figure 9: Tail 탐지 민감도: T-cKS / cKS 비율

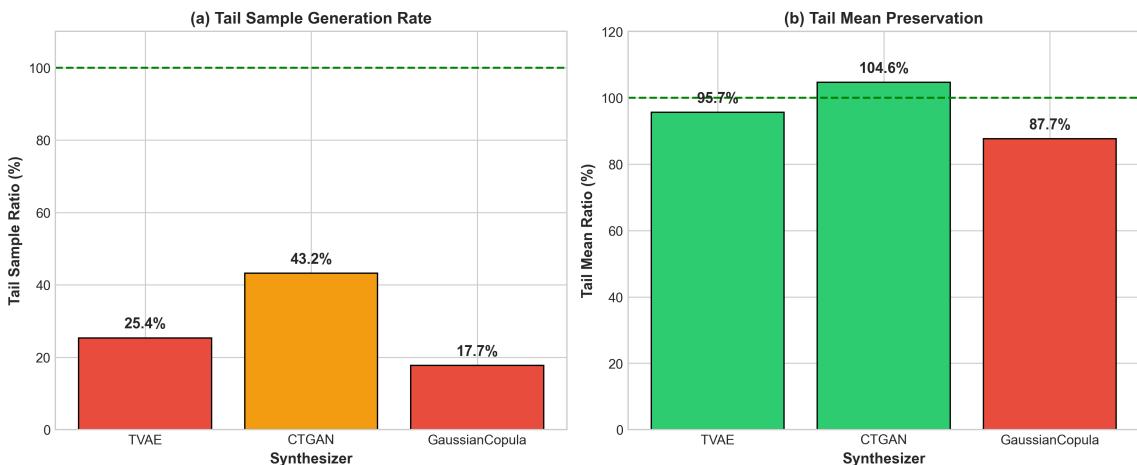


Figure 10: 합성 데이터 생성 모델별 Tail 재현율: (a) Tail 샘플 생성 비율, (b) Tail 평균 보존율

참고문헌

1. Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83-91.
2. Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2), 279-281.
3. Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer.
4. Xu, L., et al. (2019). Modeling tabular data using conditional GAN. *NeurIPS*, 32.
5. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *ICLR*.
6. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *IEEE DSAA*, 399-410.

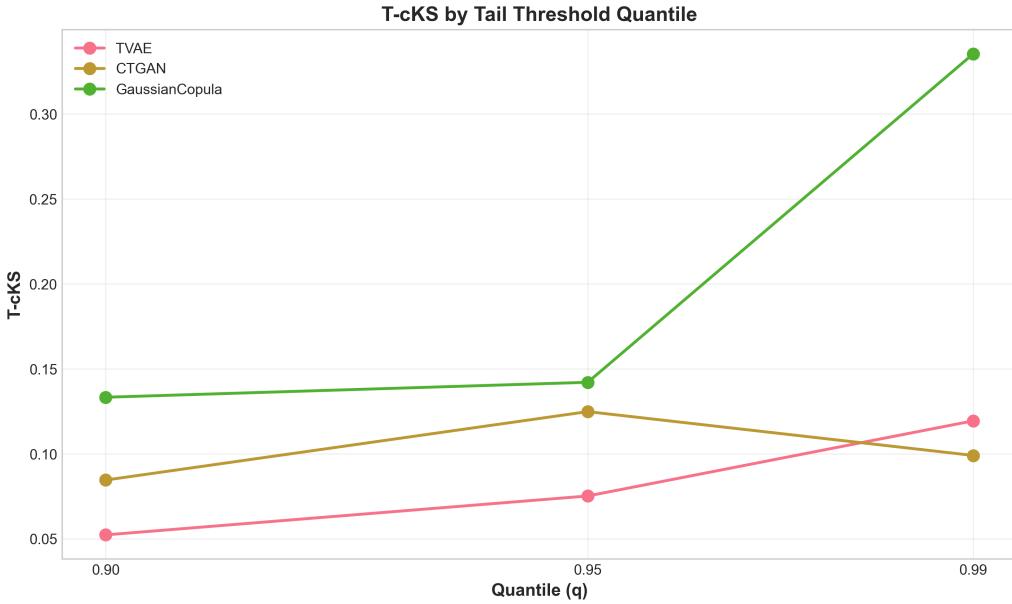


Figure 11: Tail Threshold Quantile에 따른 T-cKS 변화

7. OpenML. (2016). Allstate Claims Severity Dataset. <https://www.openml.org/search?type=data&id=42571>

A 증명 세부사항

보조정리 1 (Tail 영역에서의 Glivenko-Cantelli). $\{Y_i\}$ 가 연속 분포 F 를 따르는 i.i.d. 샘플이고, τ 가 고정된 threshold일 때, tail 경험적 CDF에 대해:

$$\sup_{y > \tau} |\hat{F}_\tau(y) - F_\tau(y)| \xrightarrow{a.s.} 0 \quad (12)$$

여기서 $F_\tau(y) = P(Y \leq y | Y > \tau)$ 이다.

Proof. 조건부 분포에 대한 Glivenko-Cantelli 정리의 직접 적용. \square

보조정리 2 (KS 거리의 연속성). $g : (F, G) \mapsto \sup_y |F(y) - G(y)|$ 는 균등 수렴 위상에서 연속이다.

Proof.

$$|g(F_n, G_n) - g(F, G)| \leq \sup_y |F_n(y) - F(y)| + \sup_y |G_n(y) - G(y)| \quad (13)$$

따라서 $F_n \rightarrow F$, $G_n \rightarrow G$ 이면 $g(F_n, G_n) \rightarrow g(F, G)$. \square

B 재현 코드

실험 코드는 다음 저장소에서 확인할 수 있다:

<https://github.com/M00N7682/systhetic-data-experiment>