Operational Risk Capital Modeling for Extreme Loss Events by Semi-Nonparametric (SNP) Estimation

Heng Z. Chen[a], Stephen R. Cosslett[b], Gordon Liu[c], Rachel Wang[c]

Annual Computational Finance Conference
September 27-30, 2021, MathWorks, Inc.

a: HSBC Bank USA, corresponding author, email: heng.z.chen@us.hsbc.com. b: Ohio State University. c: HSBC Bank USA.

Please do not cite without authors' permission.

Table of Contents

- Executive summary and research objectives
- Challenges in the operational risk capital estimation
- The SNP model using the extreme value theory – point over threshold (EVT-POT) approach
- Example 1: The assessment of SNP model performance on simulation datasets with heavy tails
- Example 2: Operational risk capital estimation on the actual loss dataset with heavy tails

Executive Summary and Research Objectives

- Operational risk modeling using the parametric models in the EVT-POT approach can lead to a counter-intuitive estimate of value at risk at 99.9% as economic capital due to extreme events.
- This research proposes a flexible SNP model using the change of variables technique. The proposed SNP models enrich the family of distributions for modeling extreme events, thus overcoming the parametric model misspecifications.
- The SNP models are shown to have the same maximum domain of attraction (MDA) as the parametric kernels, and it follows that the SNP models are consistent with the EVT-POT approach but with different shape and scale parameters from the parametric models.
- When applied to the simulated datasets with heavy tails created by three different body-tail cutoff thresholds, the SNP models in the Fréchet and Gumbel MDAs are shown to perform satisfactorily by increasing the number of model parameters.
- When applied to an actual operational risk loss dataset from a major international bank, the SNP model capital estimate is more stable and intuitive, around 2 to 2.5 times as large as the single largest loss event.

Challengers in The Operational Risk Capital Estimation

The operational risk literature of AMA/LDA approach illustrates a lot of challenges. For examples,

- Moscadelli (Banca D'Italia discussion paper, 2004) found that the estimated Pareto tail shape parameter often exceeds one, yielding a counter-intuitive capital estimate.
- Cope et al. (Journal of Operational Risk, 2009) found that by removing the top three loss events from the modelling data sample, the quantile estimate at 99.9% reduces by a 65%, reflecting the significant impact of extreme loss events in the capital estimation.
- Colombo et al. (Journal of Operational Risk, 2015) used the weighted MLE by assuming the contaminated data points.
- Abdymomunov and Curti (Journal of Financial Services Research, 2019) proposed to rescale the bank loss by total assets to arrive at a more stable capital estimate through combining peer banks' data. However, the unobserved characteristics that might also influence the loss.
- Neslova *et al.* (Journal of Operational Risk, 2006) raised concern about naïve application of EVT-POT approach to operational risk capital modelling without a careful understanding of the loss dataset. It suggested that mixed true data generating processes can turn out to be difficult to detect if one does not look for them.
- Embrechts et al. (Modelling extremal events, 1997) commented on the challenges of making the body-tail cut off by following the EVT-POT approach.

The Semi-Nonparametric Estimation

- Gallant and Nychka (Journal of Econometrics, 1987) introduced the SNP methodology by combining a normal kernel with Hermite polynomials, and showed that the approximation errors can be made arbitrarily small by increasing the polynomial truncation point.
- Chen (Handbook of Econometrics, 2007) indicated that another attractive feature of the SNP methodology is its ease of implementation since the SNP distribution can often be characterized by a finite number of parameters, reduced to a parametric model, and thus estimated by maximum likelihood, generalized least squares, sieve minimum distance and other methods.
- Chen and Randall (Journal of Econometrics, 1997) introduced a semi-nonparametric estimation using the change of variables technique in the context of binary choice models and demonstrated its asymptotic statistical properties. The estimated willingness to pay is found to be substantially different from that of the initial parametric model.
- In this research, we extend the SNP estimation by change of variables to model tail events above the EVT-POT threshold without treating extreme events as contaminated data points. We found that the SNP distributions in the Fréchet and Gumbel MDA can be used to model the heavy tail loss events and result in a substantially different capital estimate from the parametric model.

The SNP Estimation Based on Jacobian Transformation

- Let $f(x)$ be the density function of any continuous variable x. Let $v = h(x)$ which has a known density function $g(v)$. Then we have

$$f(x) = g(h(x))\,\nabla_x h(x)$$

where $\partial v/\partial x = \partial h(x)/\partial x \equiv \nabla_x h(x) > 0$.

- For example, the following power series guarantees the gradient $\nabla_x h(x)$ to be non-negative.

$$\nabla_x h(x, \theta_0, \theta_1, \cdots, \theta_K) = \left(\sum_{k=0}^{K} \theta_k x^k\right)^2 \equiv \sum_{i=1}^{m} i\gamma_i x^{i-1} \geq 0$$

- Let $g(v)$ be the GPD density function. The approximated true density function has the mixed form with the weight that is a power function of degree m

$$f(x, c, \theta) = \left\{\sum_{j=1}^{m} jx^{j-1}\gamma_i\right\}\left\{1 + c\sum_{i=1}^{m} \gamma_i x^i\right\}^{-(1+\frac{1}{c})} \equiv \sum_{j=1}^{m} j\gamma_j x^{j-1} A_m(x)$$

where $A_m(x) = \left\{1 + c\sum_{i=1}^{m} \gamma_i x^i\right\}^{-(1+\frac{1}{c})}$.

Consistency With The EVT-POT Approach: An Example of SNPGPD Model

If the kernel distribution belongs to the Fréchet $\mathrm{MDA}(\Phi_{-1/c})$, as does the GPD with shape parameter c, then the SNPGPD also belongs to the Fréchet $\mathrm{MDA}(\Phi_{-1/\xi})$ with shape parameter $\xi = c/m$, where m is the degree of SNP polynomial $h(x)$.

Notes:

- The SNPGPD model with K additional parameters or order m has the shape parameter $c/(1+2K)$ or c/m.
- Instead of a constant scale parameter b for the GPD variable v, the SNPGPD model variable x is transformed or "rescaled" by the power series $h(x, \theta_0, \theta_1, \cdots, \theta_K)$ with parameters θ's or γ's.
- The SNPGPD model tail behavior $L(x)x^{-1/\xi}$ is more stable or has a smaller value than the GPD model as $x \to \infty$ asymptotically, where $L(x)$ is a slow varying function.
- As a result, the SNP distribution enriches the family of distributions that can be used to estimate the VaR model using the EVT-POT approach.
- The result can be proved by using the Gnedenko condition. In general, the SNP model under the continuous monotonic transformation $v = h(x)$ will not change the MDA of its kernel (Fréchet, Gumbel, and Weibull), and it follows that the SNP model is consistent with the EVT-POT approach.

The Representative SNP Model Log-Likelihood Functions

This research selects the following representative SNP model likelihood functions to model operational risk loss for the three MDAs in the Fisher-Tippett theorem.

1. Fréchet MDA: Generalized Pareto and Log-Logistic distributions
   - $\log L_{snpgpd}(c, \theta|x) = \log\left\{\sum_{j=1}^{m} j\gamma_j x^{j-1}\right\} - \left(1+\frac{1}{c}\right)\log\left\{1 + c\sum_{i=1}^{m}\gamma_i x^i\right\},$
   - $\log L_{snplgt}(c, \theta|x) =$

     $\log(c) + \log\left\{\sum_{j=1}^{m} j\gamma_j x^{j-1}\right\} + (c-1)\log\left\{\sum_{i=1}^{m}\gamma_i x^i\right\} - 2\log\left\{1 + \left(\sum_{i=1}^{m}\gamma_i x^i\right)^c\right\},$
2. Gumbel MDA: Lognormal distribution
   - $\log L_{snplgn}(c, \theta|x) =$

     $-\frac{1}{2}\log(2\pi c^2) - \log\left\{\sum_{i=1}^{m}\gamma_i x^i\right\} - \frac{1}{2c^2}\left\{\log\left(\sum_{i=1}^{m}\gamma_i x^i\right)\right\}^2 + \log\left\{\sum_{j=1}^{m} j\gamma_j x^{j-1}\right\},$
3. Weibull MDA for maximum: Weibull distribution
   - $\log L_{snpwbl}(c, \theta|x) =$

     $\log(c) + \log\left\{\sum_{j=1}^{m} j\gamma_j x^{j-1}\right\} + (c-1)\log\left\{\sum_{i=1}^{m}\gamma_i x^i\right\} - \left\{\sum_{i=1}^{m}\gamma_i x^i\right\}^c$

Example 1: Simulation Dataset

- Three datasets are generated, each with 1000 observations. Heavy tails mainly come from the Log-Logistic and Pareto distributions. The Weibull distribution has a light tail.

Summary Statistics	Shape	Scale	Sample Size	Minimum	Mean	Maximum	Standard Deviation	Skewness
Weibull	5/3	1/3	1000	0	12.67	978	55.66	11.02
Pareto	4/3	1/4	1000	0	14.57	7,856	259.13	28.15
Log-Logistic	2/3	1/20	1000	0	21.65	16,659	534.75	30.31

- Three mixed exceedance datasets with the body-tail cutoff at 50, 30, and 10 are created by using the Pareto, Log-Logistic, and Weibull distributed samples.

Modeling Dataset	Sample Size	Distribution	Count	Minimum	Mean	Maximum
Cut at 50 (2% Sample)	72	Weibull	49	4.96	127.06	928.28
		Pareto	13	20.56	957.43	7805.99
		Log-Logistic	10	3.1	2033.75	16609.29
Cut at 30 (4% Sample)	128	Weibull	82	0.87	91.53	948.28
		Pareto	21	0.89	608.02	7825.99
		Log-Logistic	15	1.74	1371.03	16629.29
Cut at 10 (9% Sample)	258	Weibull	179	0.45	55.3	968.28
		Pareto	47	0.09	284.15	7845.99
		Log-Logistic	32	0.23	654.89	16649.29

The Estimated Models on the Simulation Dataset

- In total, we estimated sixty-three models on the three simulation datasets to evaluate the model performance, including sensitivity to the body-tail threshold. Specifically,
- Six popular parametric distributions
  - Generalized Beta of Type 2 with four parameters (GB2)
  - Generalized Pareto, Log-Logistic, Log-Normal and Weibull with two parameters
  - Exponential with one parameter
- Five simple parametric distributions as the SNP kernels
  - Exponential, Generalized Pareto, Log-Logistic, Log-Normal, and Weibull.
- Fifteen SNP models on the five kernels with two, three, four additional parameters
  - SNPGPD2p, SNPLGT2p, SNPLGN2p, SNPWBL2p, SNPEXP2p
  - SNPGPD3p, SNPLGT3p, SNPLGN3p, SNPWBL3p, SNPEXP3p
  - SNPGPD4p, SNPLGT4p, SNPLGN4p, SNPWBL4p, SNPEXP4p

The SNP Model Performance Assessment

- The SNP model specification can be evaluated by gradually increasing the order of the polynomial to find a best fit model to the dataset.
- Since the SNP model nests the selected parametric model as a special case, traditional model specification tests can be carried out such as the nested LR test as UMP and Student t-test on the additional parameters.
- Q-Q plots will be evaluated to ensure that the model does not over-predict or under-predict the observed, especially for extreme events.
- Tail distribution $1 - F(x)$ and quantile estimate at 99.9% will be analyzed carefully due to its influence on the capital VaR at 99.9% which is more critically influenced by severity of extreme events than the count distribution.
- Sensitivity of different body-tail cutoff thresholds is assessed on the quantile estimate at 99.9%.

Model Performance: Parametric Models

- The Generalized Beta of Type 2 (GB2) model has the best performance among the parametric models due to its flexibility with four parameters.
- The GPD and LogLGT models perform better than the other two parameters models since the dataset is created using the mixture of Pareto, Log-Logistic, and Weibull distributions.
- However, the 99.9% quantile estimate for the parametric models differs significantly for the different models and across all three datasets with thresholds at 50, 30, and 10.

Performance Comparison	Sample Threshold	Six Parametric Models					
		GB2	GPD	LogLGT	LGN	WBL	EXP
Count of Parameters		4	2	2	2	2	1
Log-Likelihood Values	50	47.83	44.3	44.69	43.9	29.55	-27.9
	30	147.17	146.47	146.32	146.12	128.87	7.2
	10	533.15	532.96	532.81	530.51	491.23	197.12
Quantile Estimates at 99.9%	50	205,644	90,206	38,945	13,429	8,212	3,743
	30	44,804	95,713	54,351	13,032	5,493	2,391
	10	20,877	27,301	19,733	5,148	2,374	1,184

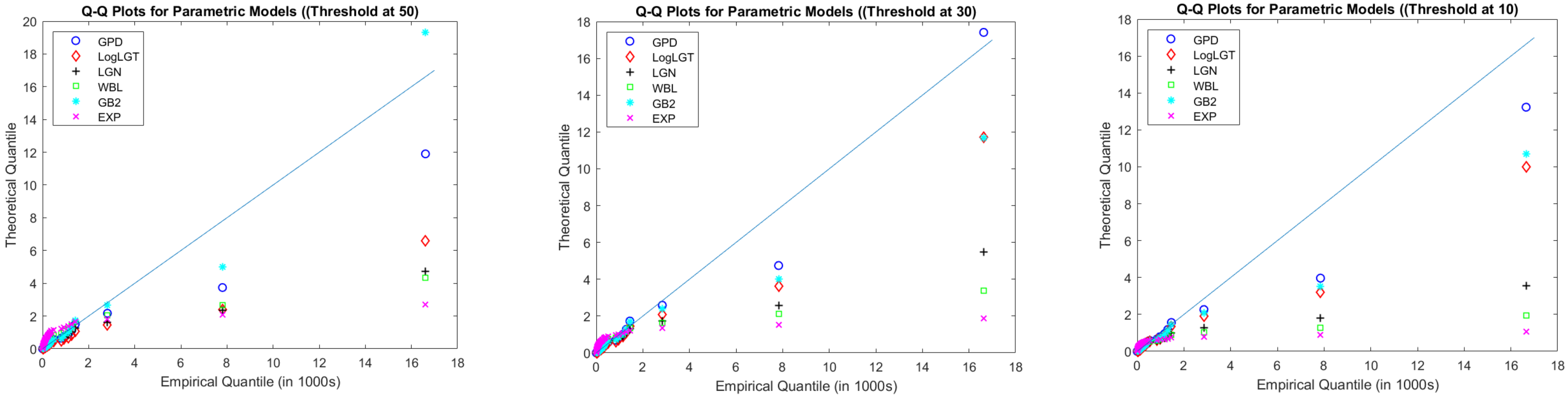
SNP Model Performance

- The SNPLGN3p performs the best with the five (or three additional) parameters.
- The quantile estimates at 99.9% by the SNPGPD, SNPLGT, and SNPLGT models are fairly stable, a key input component in the 99.9% VaR calculation.

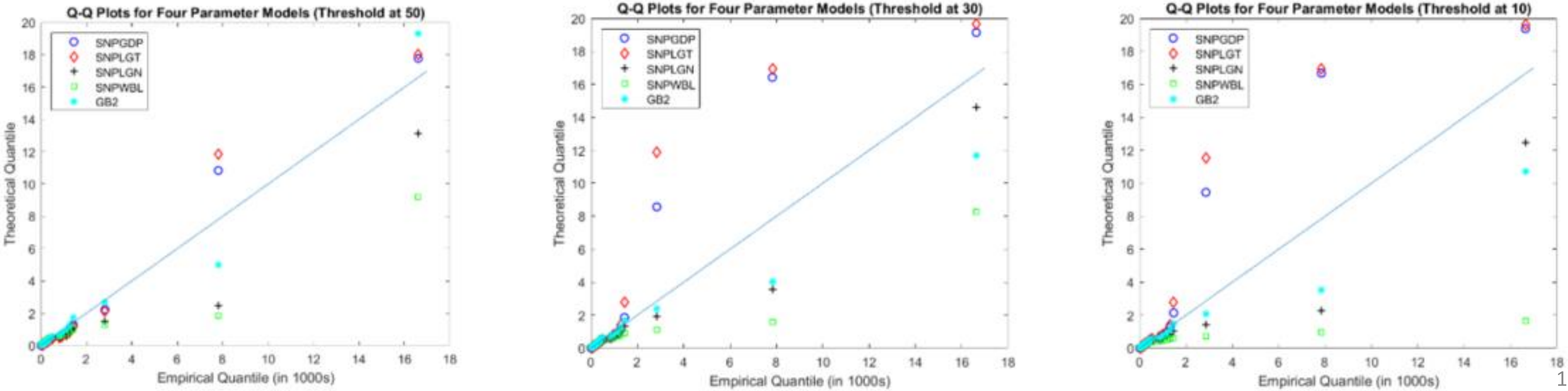
Performance Comparison	Sample Threshold	SNP Models with three additional parameters					SNP Models with four additional parameters				
		SNPGPD3p	SNPLGT3p	SNPLGN3p	SNPWBL3p	SNPEXP3p	SNPGPD4p	SNPLGT4p	SNPLGN4p	SNPWBL4p	SNPEXP4p
Parameters		5	5	5	5	4	6	6	6	6	5
Log-Likelihood Values	50	47.33	48.1	49.13	42.16	33.62	47.34	48.18	49.61	42.82	35.91
	30	149.36	149.27	150.32	143.19	110.84	149.36	149.28	150.35	143.87	115.98
	10	535.88	535.81	535.93	516.71	425.54	535.88	535.83	536.06	518.58	432.3
Quantile Estimates	50	18,993	18,973	17,735	30,370	14,378	18,636	18,325	17,727	17,438	16,970
	30	18,636	18,505	17,801	28,732	1,628	18,497	18,241	17,649	17,014	2,015
	10	17,370	17,374	16,940	9,841	578	17,215	17,152	16,850	10,087	568

- The Q-Q plots and tail distributions on the following pages reconfirm the observations across the estimated sixty-three models.

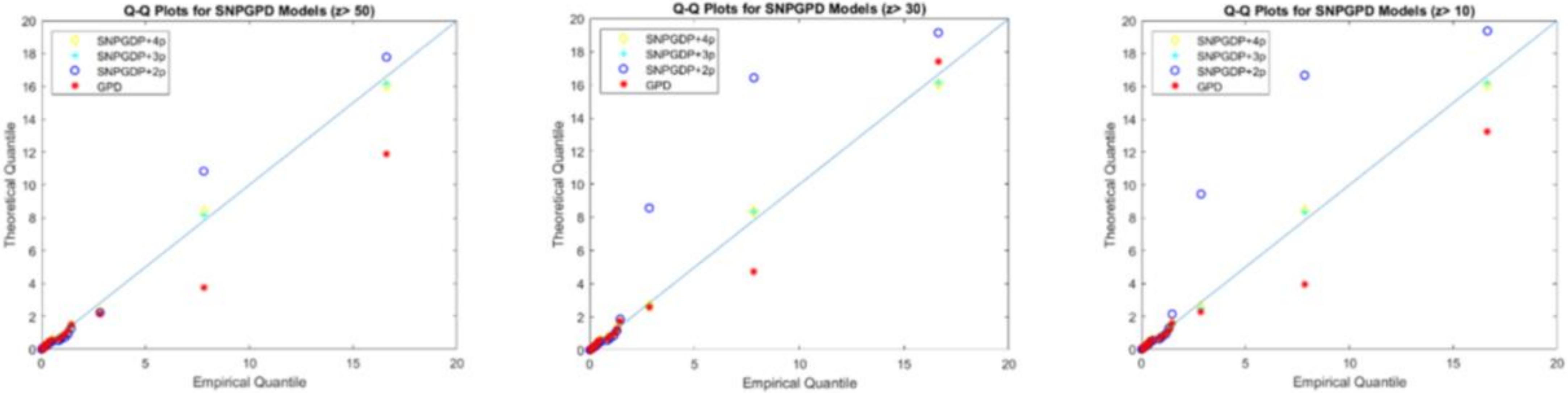
Q-Q plots for the parametric models at thresholds 50, 30, and 10.



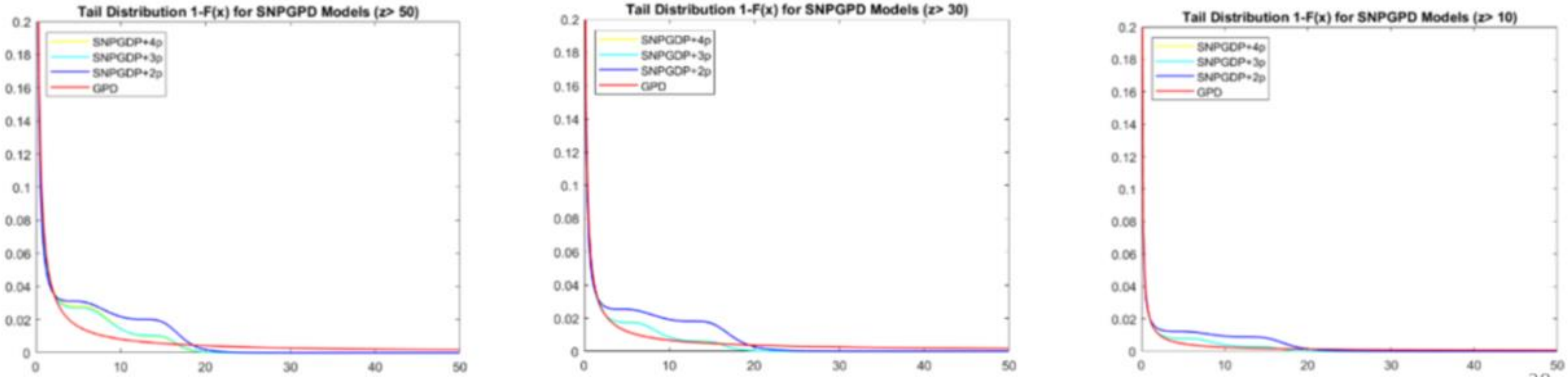
Q-Q plots for the SNP models with 2 additional parameters



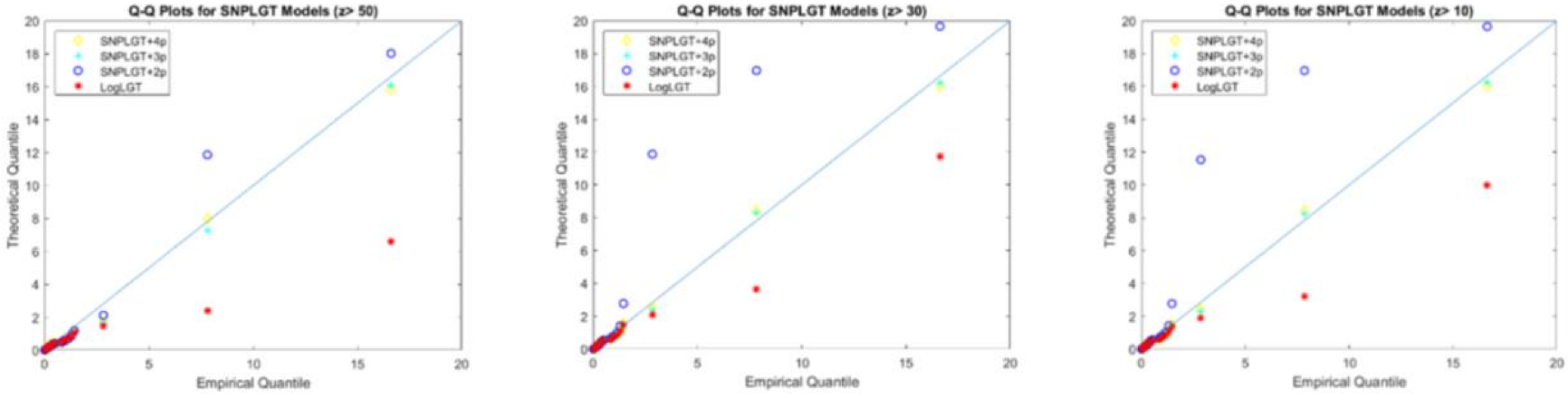
Q-Q plots for the GPD and SNPGPD models with 2, 3, and 4 additional parameters



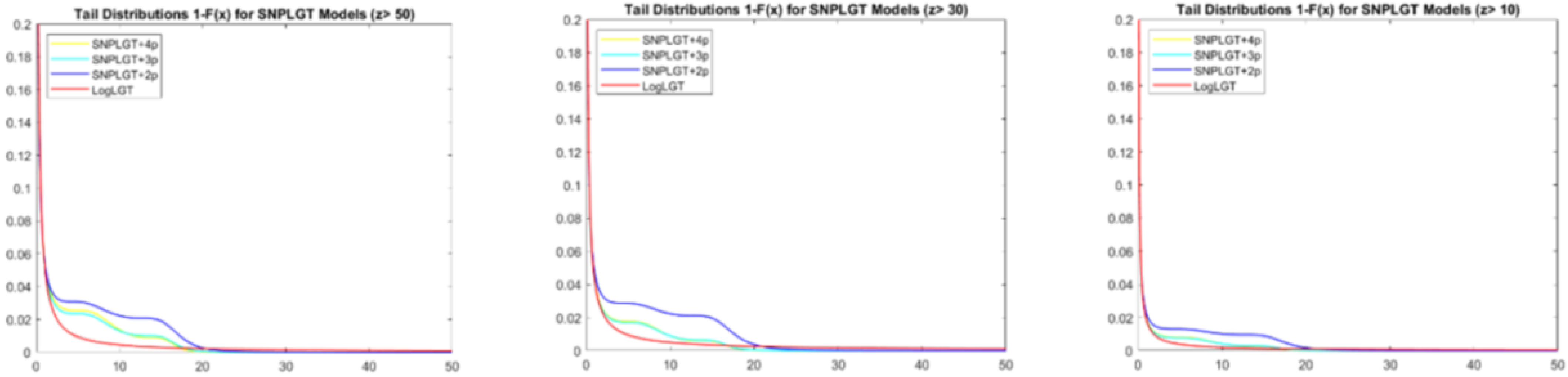
Tail distributions for the GPD and SNPGPD models with 2, 3, and 4 additional parameters



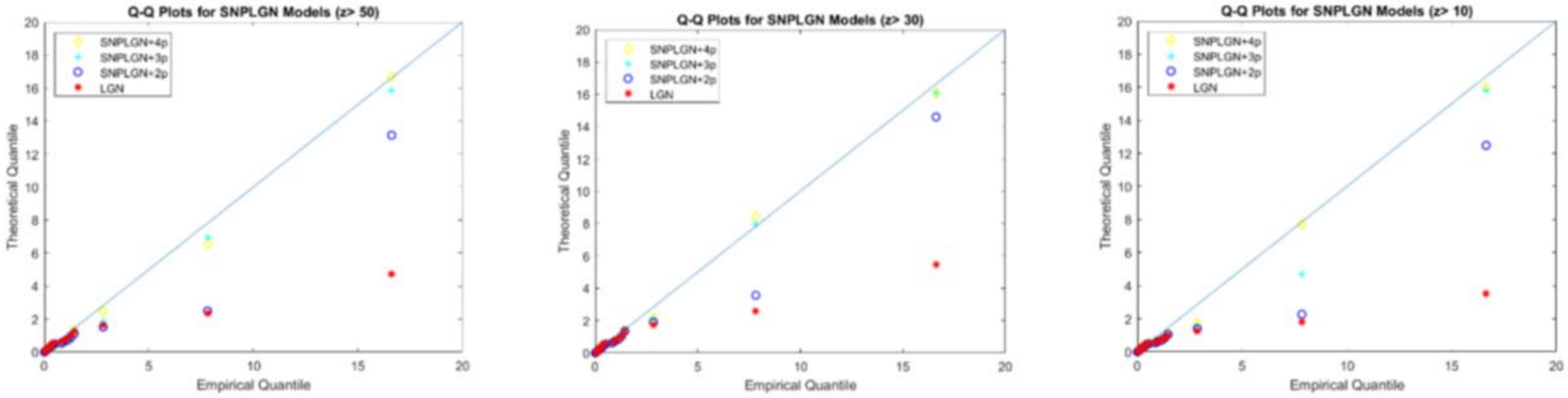
Q-Q plots for the LogLGT and SNPLGT models with 2, 3, and 4 additional parameters



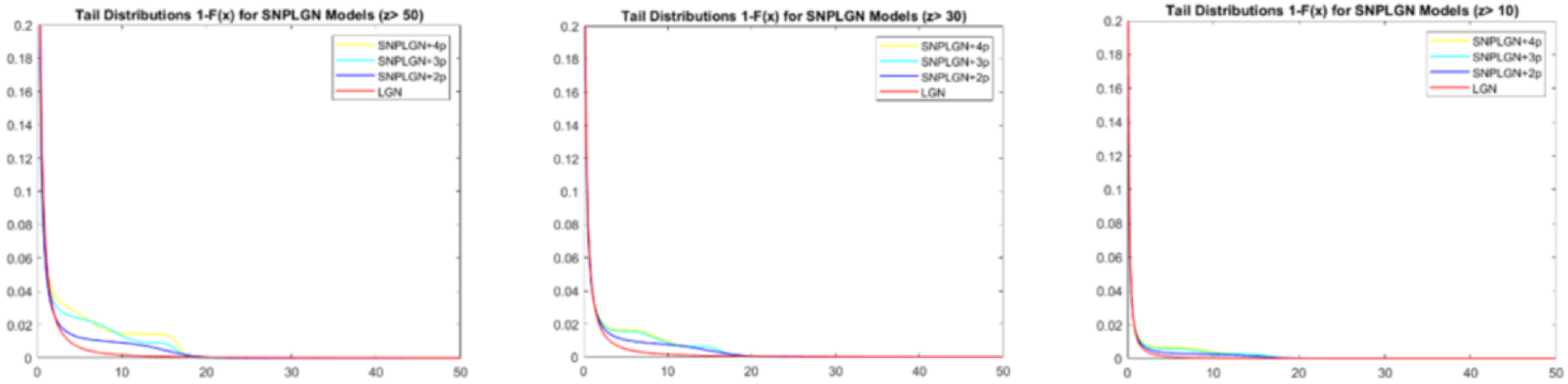
Tail distributions for the LogLGT and SNPLGT models with 2, 3, and 4 additional parameters



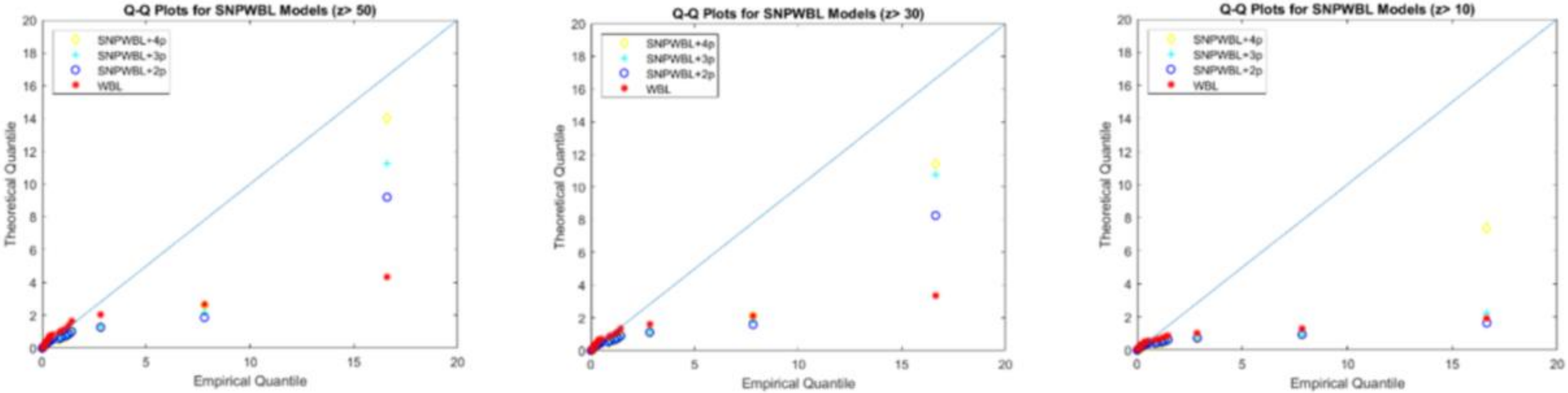
Q-Q plots for the LGN and SNPLGN models with 2, 3, and 4 additional parameters



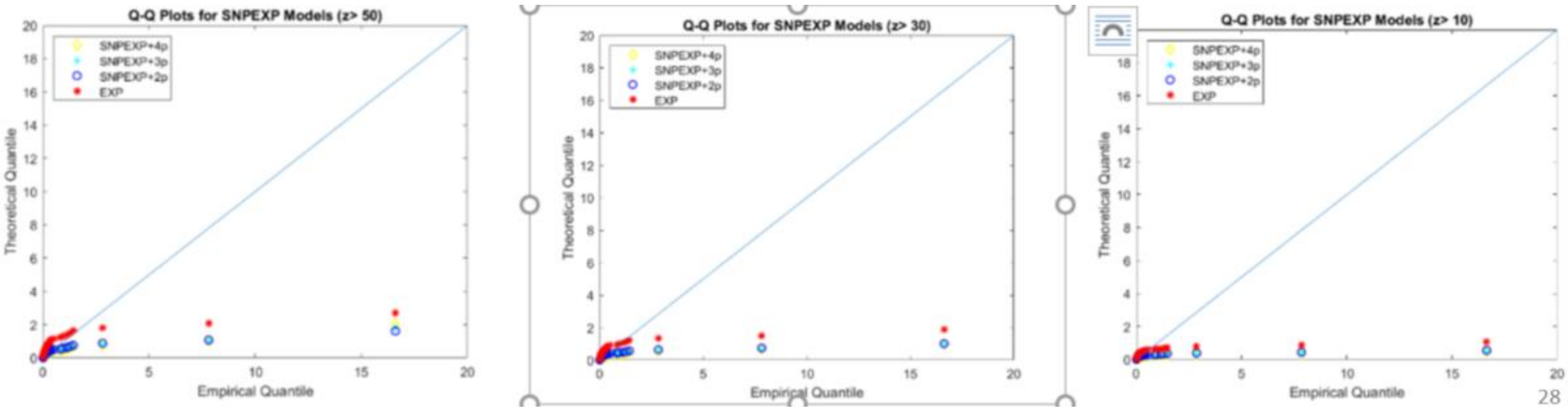
Tail distributions for the LGN and SNPLGN models with 2, 3, and 4 additional parameters



Q-Q plots for the WBL and SNPWBL models with 2, 3, and 4 additional parameters



Q-Q plots for the EXP and SNPEXP models with 2, 3, and 4 additional parameters



Example 2: The Operational Risk Loss Modeling Dataset

- The modeling dataset contains 324 CPBP* loss events from 2007Q1 to 2017Q1. It has the following loss distribution after normalization for data confidentiality.

Quantiles	Max	99%	95%	90%	75%	50%	25%	10%	5%	1%	Min
Estimates	17.0920	1.0742	0.0568	0.0196	0.0023	0.0007	0.0002	0.0001	0.0001	0.0001	0.0001

- Loss distribution has an extreme heavy tail: top 1% loss events constitute 90% of the total loss amount.
- The CPBP regulatory fines by misconducts: market manipulation, money laundering, antitrust violations, improper trade, product defects or mis-sells, fiduciary breaches, and account churning.
- The CPBP loss amount is typically in the magnitude of billions of dollars, and can also vary across Basel Units due to different business characteristics.

*The operational loss events are categorized into seven event types by the Basel Committee, namely Internal Fraud (IF); External Fraud (EF); Employment Practices and Workplace Safety (EPWS); Clients, Products, and Business Practice (CPBP); Damage to Physical Assets (DPA); Business Disruption and Systems Failures (BDSF); and Execution, Delivery, and Process Management (EDPM).

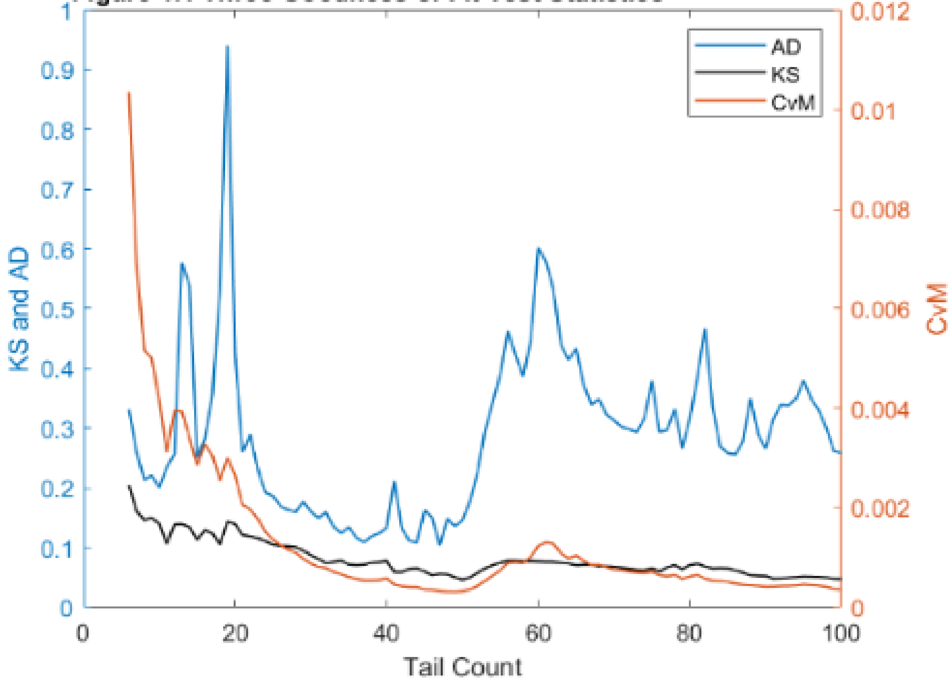
The SNP Models for Operational Risk Capital VaR

- Following industry practice, the Kolmogorov-Smirnoff (KS), Cramer-von Mises (CvM), and Anderson-Darling (AD) test statistics are calculated to determine the body-tail threshold.
- The 43 tail constitutes 99% of the total loss amount.

The GPD Model’s Goodness-of-Fit Tests				
Tests	Statistic		p-Value	
Kolmogorov-Smirnov	D	0.0646	Pr > D	0.888
Cramer-von Mises	W-Sq	0.0135	Pr > W-Sq	0.996
Anderson-Darling	A-Sq	0.11257	Pr > A-Sq	0.998

The GPD Model Estimates (LogL=42.39 and Quantile at 99.9% =1013.9)					
Parameter	Estimate	Std Err	DF	t-Value	Pr > |t|
shape	1.5858	0.3915	43	4.05	0.0002
scale	0.02811	0.00966	43	2.91	0.0057

Figure 4.1 Three Goodness of Fit Test Statistics



The SNP Models for Operational Risk Capital VaR at 99.9%

- The log-Logistic (LogLGT) and log-Normal (LGN) distributions are also estimated for the tail events

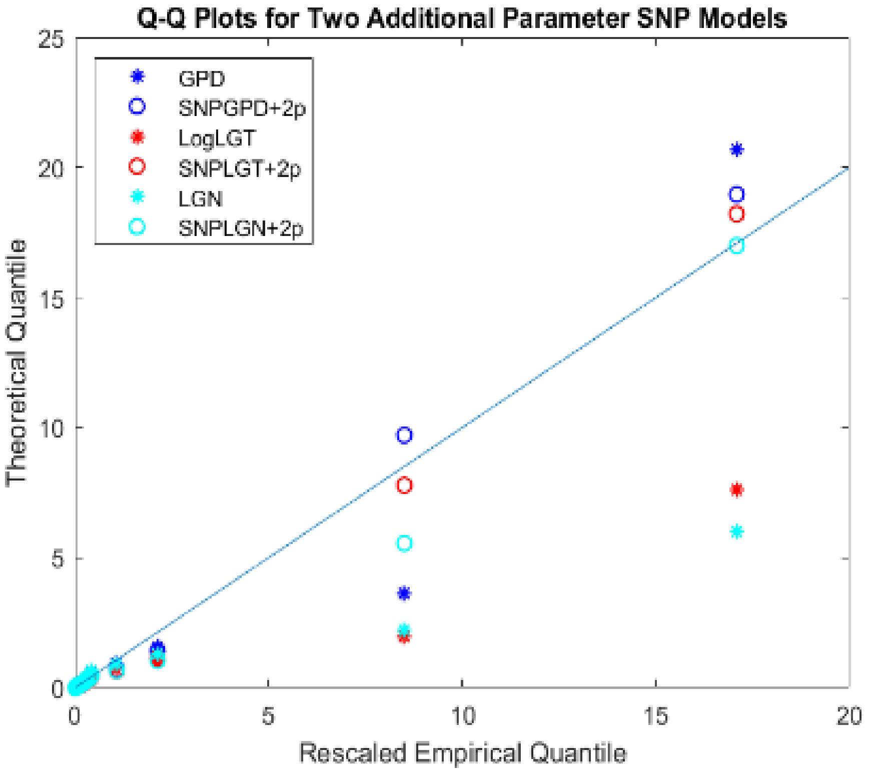
LogLGT and LGN Models			Parameters		t-Statistics	
Models	LogL	Quantile at 99.9%	μ	σ	μ	σ
LogLGT	41.70	143.2	-3.2692	1.1921	-10.375	7.836
LGN	41.33	35.7	-3.1454	2.1745	-9.485	9.11

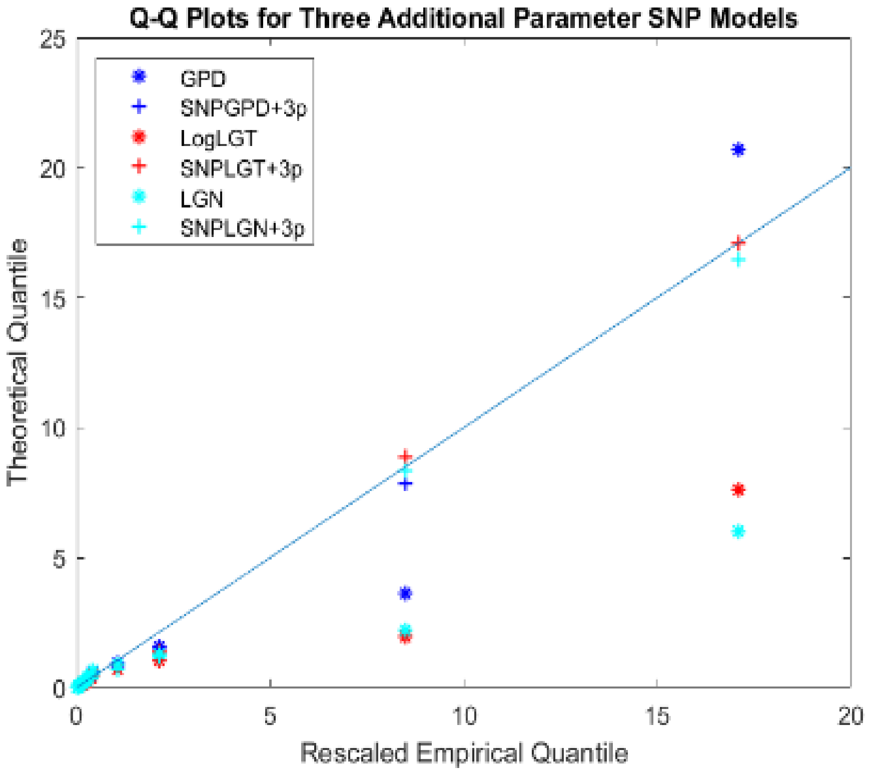
- The SNP models with two and three additional parameters exhibit a significant improvement over the selected parametric models by the LR test or Student t-test.

Model Performance		SNPGPD2p	SNPLGT2p	SNPLGN2p	SNPGPD3p	SNPLGT3p	SNPLGN3p
Log-likelihood Value		44.91	44.88	45.61	45.66	45.29	45.93
Quantile at 99.9%		27.4	25.0	20.8	23.2	21.0	18.9
Parameters	c	1.26109	0.91034	1.89910	1.51098	0.91615	1.87869
	θ_0	5.62724	5.23231	5.13742	5.88279	5.26112	5.18562
	θ_1	-1.85799	-1.71674	-1.64903	-0.05124	-2.43365	-2.39066
	θ_2	0.10966	0.10095	0.09576	-0.34192	0.39482	0.37245
	θ_3				0.02268	-0.01758	-0.01610
t-Statistics	c	3.5286	7.9220	9.2455	3.0861	7.7492	8.9210
	θ_0	5.9368	6.8134	6.9050	5.6268	6.8553	6.9400
	θ_1	-4.1974	-4.6080	-4.8425	-0.0180	-3.2068	-3.4846
	θ_2	3.3901	3.6545	3.8913	-0.4770	2.6941	2.9223
	θ_3				0.6196	-2.5005	-2.7456

The Model Performance Comparison: Q-Q Plots

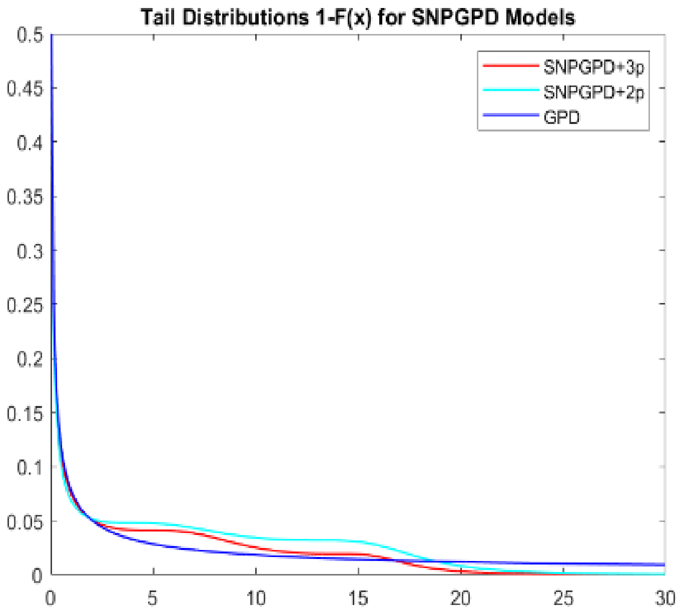
- The SNP model's performance can be improved by increasing the number of SNP model parameters.

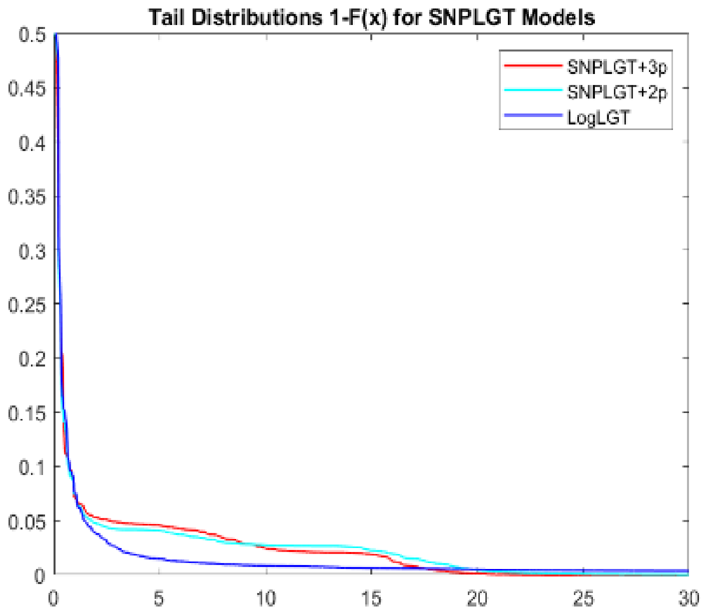


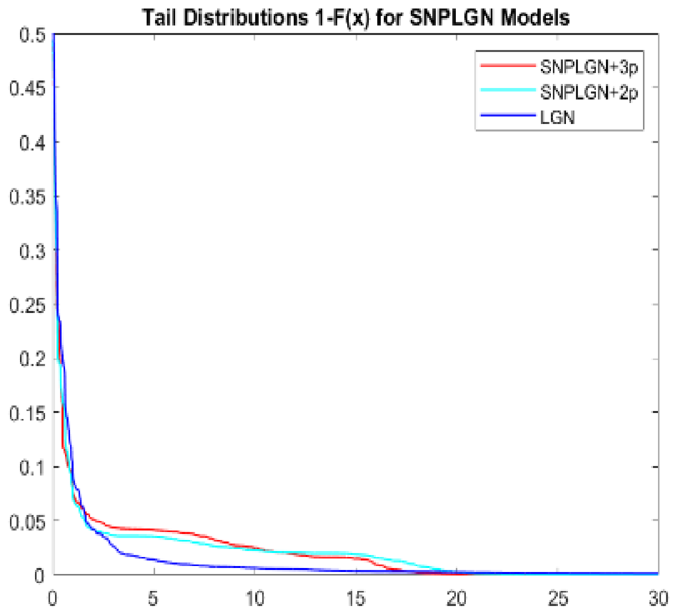


The Model Performance Comparison – Tail Distributions

- The SNP model tails converge to zero much faster than the parametric models. The additional SNP model parameters provide the flexibility to improve the model fit.
- The GPD model clearly has a heavier tail than that of the LGN and LogLGT models, which will result in a larger capital estimate.


Tail Distributions 1-F(x) for SNPGPD Models


Tail Distributions 1-F(x) for SNPLGT Models


Tail Distributions 1-F(x) for SNPLGN Models

The SNP Model’s Shape and Scale Parameters

- The SNP model’s scale is determined by the polynomial transformation, a power function with the estimated parameters $\hat{\theta}$s.
- For the GPD and LogLGT models in the Fréchet MDA, the estimated shape parameters are *c*=1.5858 and $1/\sigma$ = 0.8389, respectively, suggesting different tail behaviors.
- On the other hand, the estimated shape parameters of the SNPGPD3p and SNPLGT3p models are $\xi = c/m = 0.2157$ and $\xi = 1/(c * m) = 0.1559$, respectively. They are smaller than the corresponding parametric models.
- As a result, the SNP model tail behavior is more stable than the parametric model with a smaller tail $L(x)x^{-1/\xi}$ as $x \rightarrow \infty$ asymptotically, where $L(x)$ is a slow varying function.

The Economic Capital Comparison: VaR at 99.9%

- Since there is a minimum reporting threshold for bank operational risk losses, there are 281 loss events in the body between the minimum reporting threshold (10,000 USD) rt and the body-tail threshold bt that yields 43 tail events.
- Following industry practice, the distribution of the body loss events is estimated by truncated Lognormal model.

$$\tilde{f}(x|\mu,\sigma) = \frac{f(x|\mu,\sigma)}{F(bt|\mu,\sigma)-F(rt|\mu,\sigma)}$$

The Body LGN Model: Log-Likelihood = 1709.3					
Parameter	Estimate	Std Error	DF	t Value	Pr > |t|
μ	-10.0620	2.2565	281	-4.46	<.0001
σ	2.9345	1.1312	281	2.59	0.01

- The annual capital estimate VaR at 99.9% is simulated with 100,000 iterations in the following table. The SNPLGN3p model is selected due to its model performance.

Model Comparison	GPD	LogLGT	LGN	SNPGPD3p	SNPLGT3p	SNPLGN3p
Economic Capital VaR at 99.9%	10,410	847	92	41	40	36
Log-Likelihood Values	42.39	41.70	41.33	45.66	45.29	45.93

Conclusions

- This research extends the SNP estimation to model the operational risk capital, leading to a more stable and intuitive capital estimate than the parametric models.
- The SNP model enriches the family of distributions to estimate heavy tails with shape parameter as a function of the order of polynomial series "*m*" and the chosen kernel shape parameter "*c*". The SNP model scale parameter is also a function of the power series.
- On the model performance, since SNP models nest any chosen parametric model as a special case, the LR test and Student t-test can be assessed to ensure that the SNP model specification is justified.
- Q-Q plots can be evaluated to ensure that the incremental parameters are needed to accommodate the salient empirical regularities of heavy tails in the operational risk loss events. Tail distribution can be compared to visualize the stability of economic capital estimates.
- The SNP model specification is easy to implement, which yields the model parameter estimates in just one step. MATLAB enables this research to select a wide variety of parametric distributions and optimization algorithms to satisfy the precision requirement in the SNP model estimation and VaR capital simulation.