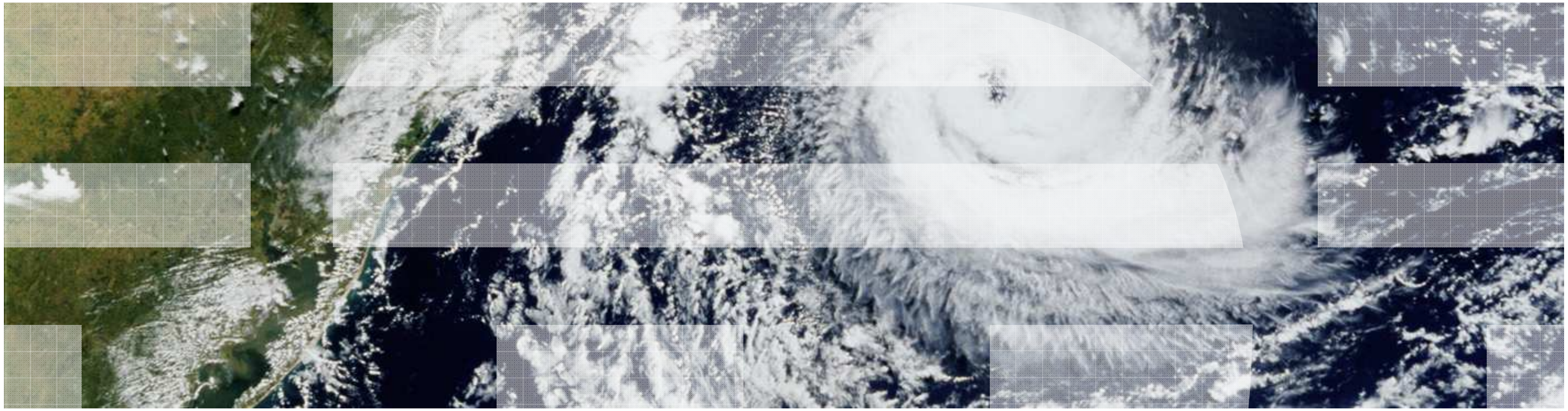
Eric Cope – IBM Research Zurich
27 March 2010

Penalized Likelihood Estimation for Truncated Data



Maximum likelihood estimation with truncated data

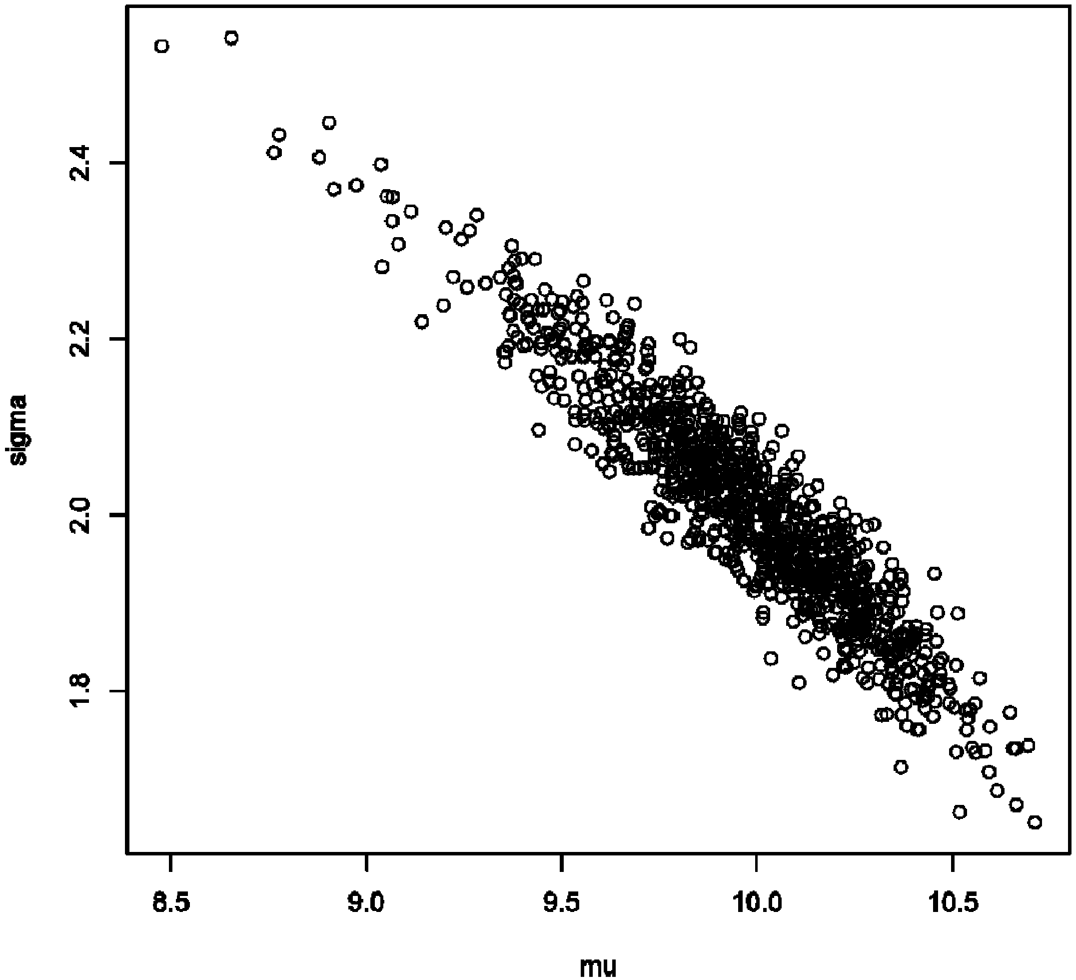
- In operational risk contexts, severity data is usually left-truncated
  – ORX has a loss reporting threshold of €20,000

- Data truncation has a distorting influence on the likelihood function
  – Causes the contours of the likelihood function to be oblong in shape (banana shape)
  – Result: high variability in estimating the maximum, high correlation among parameters
  – Especially true for location-scale families of distributions

- We have observed this
  – E.g., in fitting lognormals, possible to get very low values of μ and high values of σ
  – Although fits to the data can still be good, the values of the estimates can have a large effect when extrapolating to estimate high quantiles

- Can we reduce the distorting effects of truncation to get more stable & robust estimators?

Variance of sample solutions for estimating a Lognormal (10,2) truncated at 20,000



Maximum likelihood estimation with truncated data

- Suppose we have data $x_1, x_2, \ldots, x_n$ which is left-truncated a level t and we are fitting a distribution function F(x; θ) with density function f(x; θ)

- In maximum likelihood estimation with truncated data we find the value of θ minimizing

$$L_n(\theta; x) = -\frac{1}{n} \sum_{i=1}^{n} [\underbrace{\log f(x_i;\theta)}_{\text{(a)}} - \underbrace{\log(1 - F(t;\theta))}_{\text{(b)}}]$$

  – The term (a) is the usual objective function for maximum likelihood estimation

  – The presence of the second term (b) is needed to account for data truncation

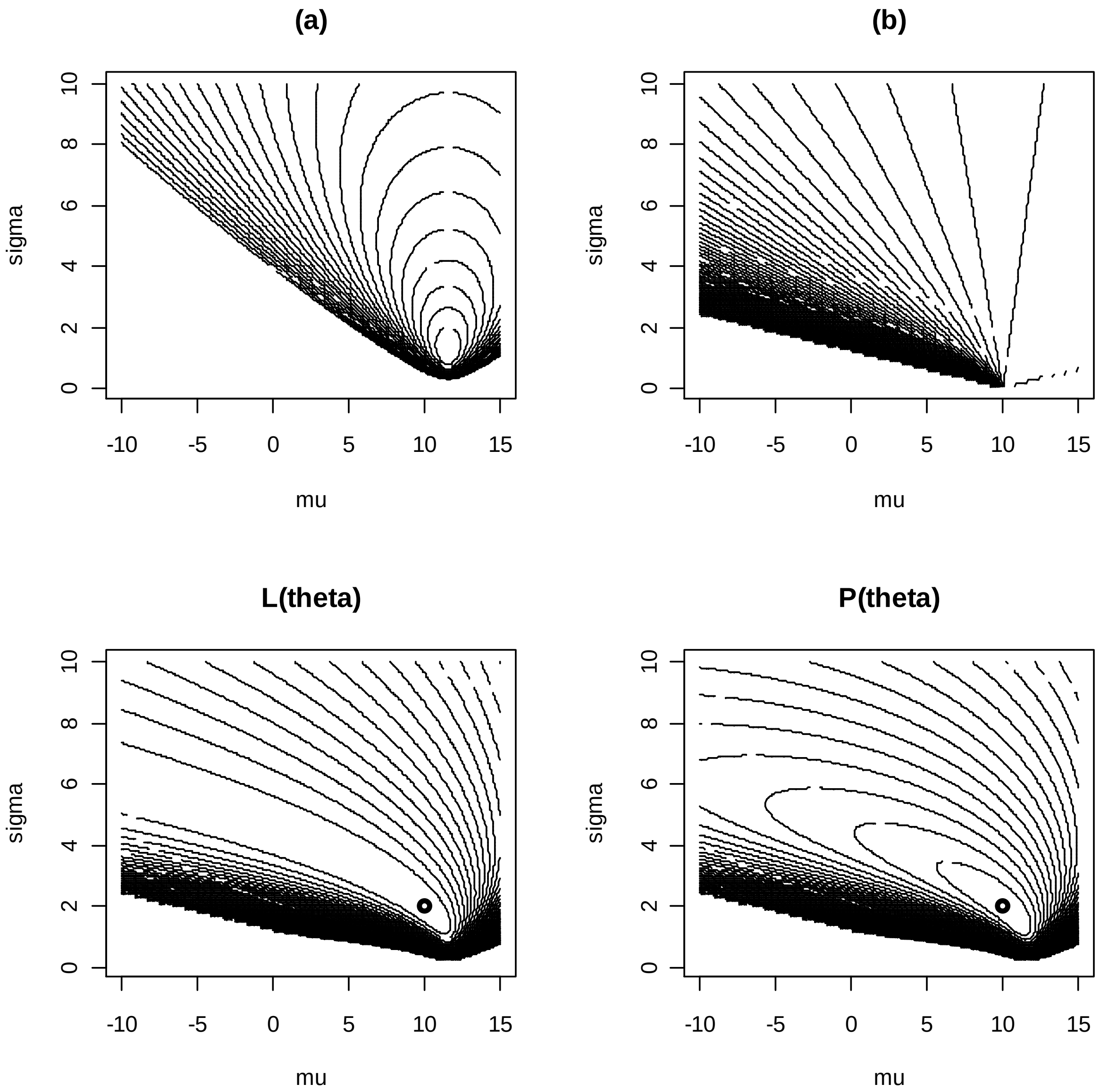
- We propose to add a *linear penalty term* that reduces the distorting effect of term (b)

$$P_n(\theta; x) = L_n(\theta; x) + \frac{\kappa}{n} \nu^T \theta$$

penalty parameter

penalty vector ||v|| = 1

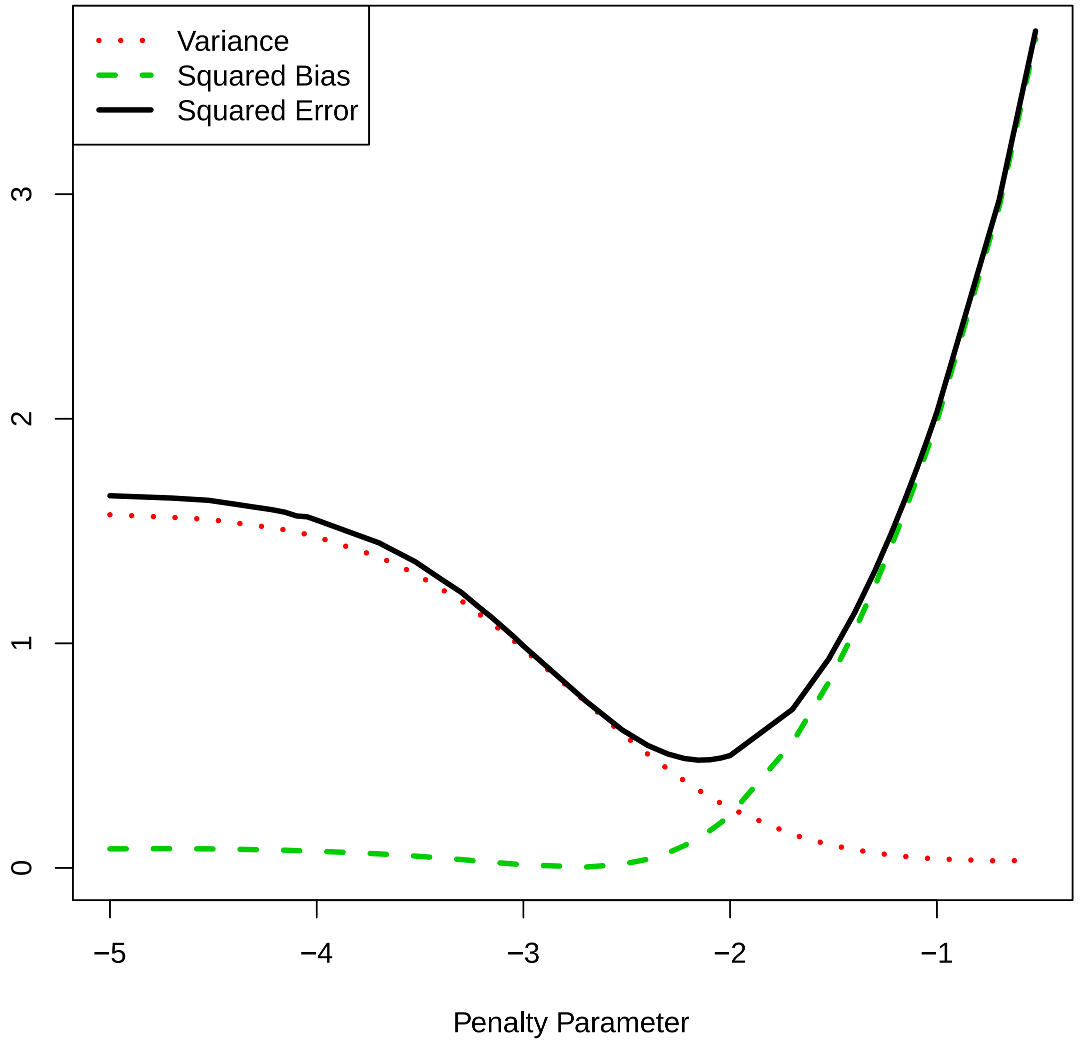
Likelihood function contours for Lognormal(10,2) truncated at 20,000



Properties of the penalized likelihood estimator

- The addition of the linear term "tilts" the likelihood function in a particular direction
  - Should tilt in a direction that achieves greatest reduction in variance of estimator
    - Direction of tilt should be aligned with the "valley floor" of the likelihood function
    - Rule of thumb: align penalty vector v with the first principal component of inverse of Hessian matrix of the likelihood function evaluated at the ML estimator
- Penalized likelihood estimator is an M-estimator
  - Distinct setting from robust estimation / Huber estimators
- As an M-estimator, it inherits many properties of maximum likelihood estimator
  - Consistent
  - Asymptotically Normal
  - Asymptotically efficient
- Trades off large amount of variance for a small increase in bias – overall reduction in error
  - In the spirit of ridge regression and shrinkage estimators (James-Stein)
  - Does not shrink estimates automatically to zero; rather, we choose penalty direction

Bias-variance tradeoff in selection of penalty parameter



Consistency and asymptotic normality: two views

- Assuming the penalty parameter κ/n is $o_P(1)$ as $n \to \infty$, asymptotically the distribution of the penalized likelihood estimator converges to the maximum likelihood estimator

- Suppose rather that $\kappa' \equiv \kappa/n$ is a fixed penalty parameter and define

$$P_n'(\theta; x) = L_n(\theta; x) + \kappa' \nu^T \theta$$

  – This estimator is asymptotically biased, as the penalty term does not reduce with n
  – We use this representation of the estimator for the purposes of deriving a suitable heuristic for selecting the penalty terms κ and ν for a given sample size n

- Let

$$\begin{aligned} l(\theta; x) &= \log f(x; \theta) - \log(1 - F(t, \theta)) \\ m_\theta(x) &= -l(\theta; x) + \kappa' \nu^T \theta \end{aligned}$$

  and let θ' minimize $Em(x)$

- Then under this view, the estimator converges to θ' and

$$\sqrt{n}(\hat{\theta}_n' - \theta') \Rightarrow N(0, V_{\theta'}^{-1} E[\dot{m}_{\theta'} \dot{m}_{\theta'}^T] V_{\theta'}^{-1})$$

  where $V_{\theta'}^{-1}$ is the Hessian of $El(\theta'; x)$, and $\dot{m}_{\theta'}$ is the derivative of $m_{\theta'}(x)$

How should we choose the penalty parameter κ ?

- Often can make a reasonable guess from the geometry of likelihood function
- Heuristic developed to automatically choose κ for a given choice of v
  - Based on first-order and asymptotic approximations of the change in bias and variance components of the estimator error

$$\Delta\mathrm{MSE} \approx \underbrace{-\frac{\kappa}{n^2}\nabla J(\theta)V_\theta^{-1}\nu}_{\text{variance}} + \underbrace{\frac{\kappa^2}{n^2}\nu^T V_\theta^{-1}V_\theta^{-1}\nu}_{\text{squared bias}}$$

V_θ^{-1} = inverse Hessian of LL function at θ

$J(\theta)$ = est. variance at MLE value θ = trace(V_θ^{-1})

  - Optimizing this expression for κ yields

$$\kappa^* = \frac{\nabla J(\theta)V_\theta^{-1}\nu}{2\nu^T V_\theta^{-1}V_\theta^{-1}\nu}$$

does not depend on *n*!

- Related expressions can be derived to minimize approximate estimation error for functionals of the distribution (such as high quantile values) rather than parameter error

IBM

Choosing the penalty parameters: derivation

- The MSE of the estimator is equal to its variance plus the square of its bias
  - The covariance matrix can be approximated using

$$\frac{1}{n}V_{\theta'}^{-1}E\left[\dot{m}_{\theta'}\dot{m}_{\theta'}^T\right]V_{\theta'}^{-1} = \frac{1}{n}V_{\theta'}^{-1} + \frac{\kappa^2}{n^3}V_{\theta'}^{-1}\nu\nu^T V_{\theta'}^{-1}$$

  - Using a second-order Taylor expansion of the likelihood, the bias can be approximated

$$-\frac{\kappa}{n}V_{\theta_0}^{-1}\nu$$

- The MSE may therefore be approximated as

$$\mathrm{MSE} \approx \frac{1}{n}\mathrm{trace}(V_{\theta'}^{-1}) + \frac{\kappa^2}{n^2}\nu^T V_{\theta_0}^{-1}V_{\theta_0}^{-1}\nu.$$

- The variance expression requires θ' rather than θ_o. Substituting a first-order approximation of $J(\theta) \equiv \mathrm{trace}(V_\theta^{-1})$,

$$J(\theta') \approx J(\theta_0) + \nabla J(\theta_0)(\theta' - \theta_0)$$

  we can derive the expression for ΔMSE on the previous slide

Choosing the penalty direction

- From the expression for ΔMSE we can deduce
  – If ν is chosen among values such that $||V_{\theta_o}{}^{-1}\nu||_2 = 1$, then MSE is minimized when $V_{\theta_o}{}^{-1}\nu$ is aligned with $\nabla J(\theta_o)$
  – Thus a good choice for ν is $V_{\theta_o}\nabla J(\theta_o)^T$

- Alternatively, from heuristic reasoning one could select ν to be aligned with the largest principal component of $V_{\theta_o}{}^{-1}$
  – Supposing the associated eigenvalue of the principal component is λ, we have (assuming $||\nu||_2 = 1$)
    - $\kappa^* = \nabla J(\theta_o)\nu \,/\, 2\lambda$
    - $\Delta\text{MSE} \approx -(\nabla J(\theta_o)\nu)^2 \,/\, 4n^2\lambda$

- Estimating $\nabla J(\theta_o)$ can be done based on the data sample
  – We find that using a finite-difference approximation of the variance trace in the principal component direction yields a stable estimator and a well-performing κ^*

Choosing the penalty parameter to minimize quantile error

- Instead of minimizing mean squared error in the parameters, we may rather be interested in minimizing a particular function $Q(\theta)$ of the resulting distribution
  - For example, the error of estimating high quantiles of the fitted distribution

- Using the delta method, the variance of the penalized estimator can be expressed

$$G(\theta') \equiv \frac{1}{n} \nabla Q(\theta') V_{\theta'}^{-1} (\nabla Q(\theta'))^T$$

- The bias can be written

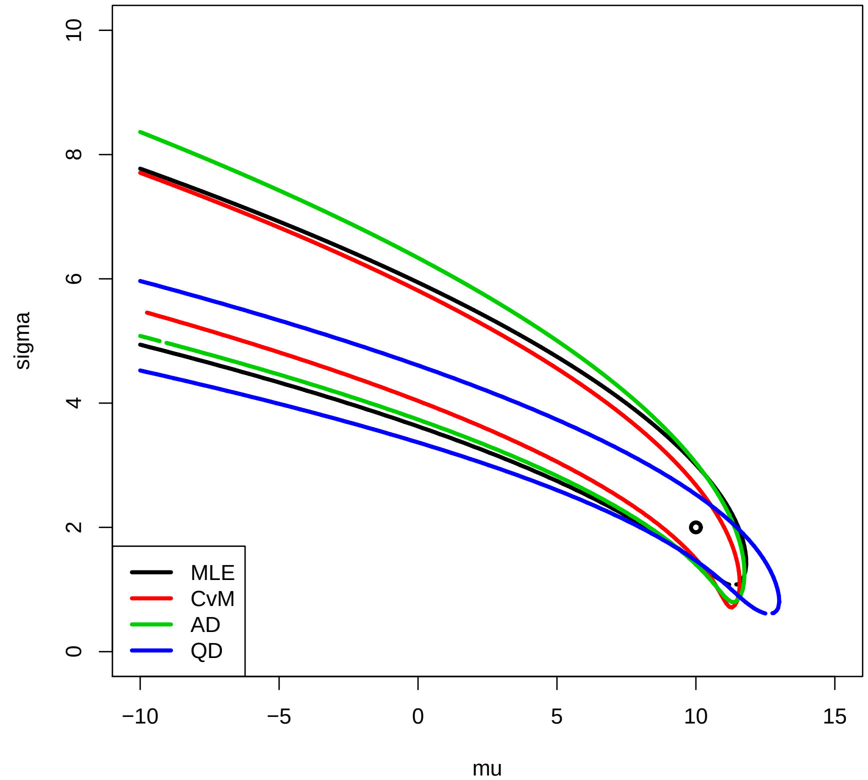
$$(\kappa/n)^2 \nabla Q(\theta_0) V_{\theta_0}^{-1} \nu \nu^T V_{\theta_0}^{-1} (\nabla Q(\theta_0))^T$$

- Again, using a Taylor expansion to approximate G(θ') in terms of θ_o an approximately optimal value of κ is

$$\kappa^* = \frac{\nabla G(\theta_0) V_{\theta_0}^{-1} \nu}{2 \nabla Q(\theta_0) V_{\theta_0}^{-1} \nu \nu^T V_{\theta_0}^{-1} (\nabla Q(\theta_0))^T}$$

Applicability to other non-ML estimation contexts ?

- The same distortions due to data truncation are apparent for the Cramer-von Mises and Anderson-Darling minimum-distance estimators, as well as the quantile distance estimator
- However, in these settings we are not dealing with M-estimators, so the theory presented does not apply



Simulation tests of estimator efficiency

- Ran several simulations for 1,000 iterations each
  - Tested left-truncated Lognormal, right-truncated log-logistic, and left-truncated Weibull
  - Sample sizes: 100 or 1,000
  - Truncation points: 25th, 50th, 75th, and 90th percentiles of distribution
- Compared four different estimators
  - Maximum likelihood estimator (No penalty)
  - Penalized estimator using κ* based on knowledge of true parameters (κ* - True Param)
  - Penalized estimator using κ based on bootstrap search for best value ($\hat{\kappa}$ - True Param)
  - Penalized estimator using κ* based on estimated parameters (κ* - Est Param)
- Results summary
  - Substantial improvements in MSE in nearly every case
  - Works best in low-data scenario; in high-data scenario performs closer to MLE
  - Best results when truncation occurs near the median of the underlying distribution

Simulation Results: left-truncated Lognormal(10,2), sample size 100

Trunc % (Value)[a]	Estimator	Variance	Bias	MSE[b]	SE Err[c]	% Redux[d]
25% (5,716)	No Penalty	0.39	0.002	0.39	0.029	
	κ^* - True Param	0.16	0.072	0.23	0.007	42 %
	$\hat{\kappa}$ - True Param	0.16	0.072	0.23	0.007	42 %
	κ^* - Est Param	0.22	0.061	0.28	0.014	28 %
50% (22,026)	No Penalty	2.0	0.063	2.1	0.22	
	κ^* - True Param	0.24	0.38	0.62	0.018	70 %
	$\hat{\kappa}$ - True Param	0.42	0.14	0.56	0.020	73 %
	κ^* - Est Param	0.35	0.33	0.68	0.021	67 %
75% (84,879)	No Penalty	4.9	0.14	5.1	0.39	
	κ^* - True Param	0.24	1.9	2.1	0.040	59 %
	$\hat{\kappa}$ - True Param	0.86	0.45	1.3	0.042	74 %
	κ^* - Est Param	0.39	2.0	2.4	0.052	54 %
90% (285,816)	No Penalty	6.7	0.015	7.3	0.45	
	κ^* - True Param	0.39	4.2	4.6	0.076	38 %
	$\hat{\kappa}$ - True Param	1.5	1.4	2.9	0.086	60 %
	κ^* - Est Param	0.42	6.0	6.4	0.10	12 %

Simulation Results: left-truncated Lognormal(10,2) sample size 1000

Trunc % (Value)	Estimator	Variance	Bias	MSE	SE Err	% Redux
25% (5,716)	No Penalty	0.0327	0.0000	0.0327	0.0015	
	κ^* - True Param	0.0296	0.0012	0.0308	0.0013	6 %
	$\hat{\kappa}$ - True Param	0.0271	0.0043	0.0314	0.0012	4 %
	κ^* - Est Param	0.0303	0.0012	0.0315	0.0013	4 %
50% (22,026)	No Penalty	0.112	0.0007	0.113	0.0065	
	κ^* - True Param	0.0892	0.0041	0.0932	0.0044	18 %
	$\hat{\kappa}$ - True Param	0.0892	0.0041	0.0932	0.0044	18 %
	κ^* - Est Param	0.0944	0.0050	0.0994	0.0047	12 %
75% (84,879)	No Penalty	0.383	0.0011	0.384	0.0280	
	κ^* - True Param	0.243	0.0345	0.277	0.0132	28 %
	$\hat{\kappa}$ - True Param	0.243	0.0345	0.277	0.0132	28 %
	κ^* - Est Param	0.238	0.0538	0.292	0.0130	24 %
90% (285,816)	No Penalty	1.34	0.0061	1.35	0.110	
	κ^* - True Param	0.310	0.671	0.981	0.0275	27 %
	$\hat{\kappa}$ - True Param	0.505	0.257	0.762	0.0261	43 %
	κ^* - Est Param	0.381	0.802	1.18	0.0432	12 %

Simulation Results: right-truncated Loglogistic(10,0.8), sample size 100

Trunc % (Value)	Estimator	Variance	Bias	MSE	SE Err	% Redux
25% (9,146)	No Penalty	4.2	0.54	4.8	0.31	
	κ^* - True Param	0.11	0.51	0.62	0.014	87 %
	$\hat{\kappa}$ - True Param	0.21	0.25	0.46	0.014	90 %
	κ^* - Est Param	0.23	0.65	0.88	0.028	81 %
50% (22,026)	No Penalty	0.83	0.032	0.87	0.11	
	κ^* - True Param	0.085	0.091	0.18	0.005	80 %
	$\hat{\kappa}$ - True Param	0.14	0.029	0.17	0.006	81 %
	κ^* - Est Param	0.11	0.083	0.19	0.006	78 %
75% (53,045)	No Penalty	0.088	0.000	0.088	0.005	
	κ^* - True Param	0.047	0.016	0.064	0.002	28 %
	$\hat{\kappa}$ - True Param	0.047	0.016	0.064	0.002	28 %
	κ^* - Est Param	0.054	0.014	0.068	0.002	23 %
90% (127,744)	No Penalty	0.038	0.000	0.038	0.002	
	κ^* - True Param	0.032	0.003	0.036	0.001	6 %
	$\hat{\kappa}$ - True Param	0.032	0.003	0.036	0.001	6 %
	κ^* - Est Param	0.031	0.003	0.034	0.001	9 %

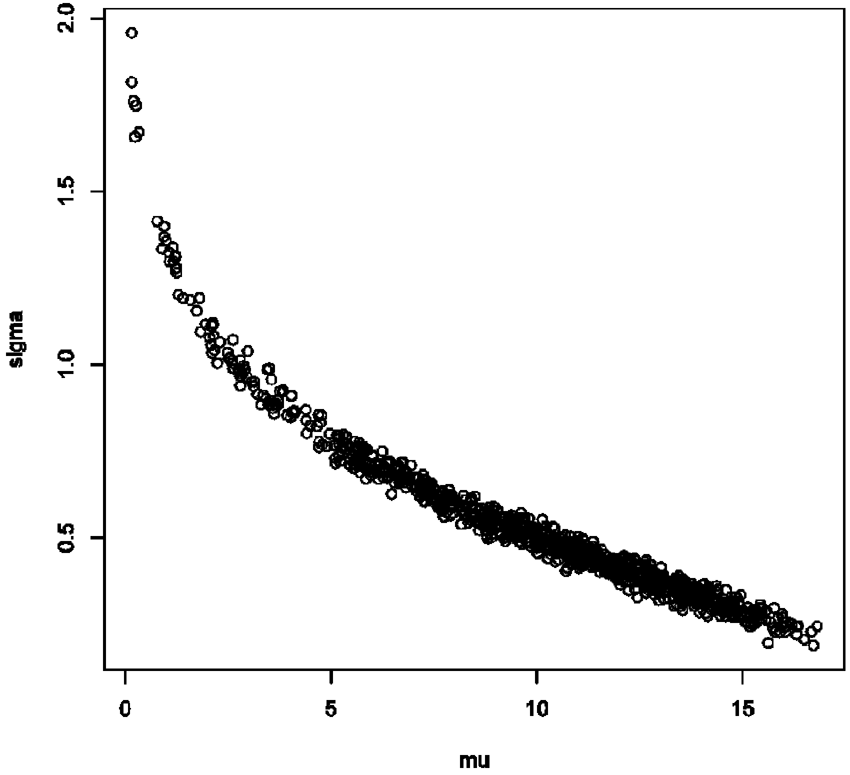
Simulation Results: right-truncated Loglogistic(10,0.8), sample size 1000

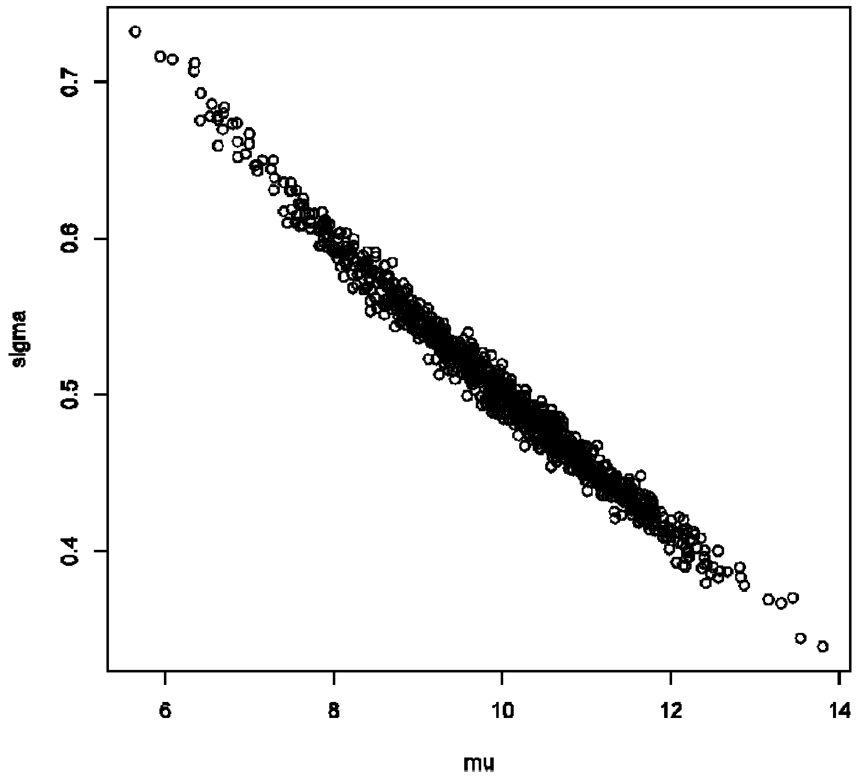
Trunc % (Value)	Estimator	Variance	Bias	MSE	SE Err	% Redux
25% (9,146)	No Penalty	0.278	0.0075	0.286	0.0363	
	κ^* - True Param	0.0874	0.0197	0.107	0.0042	63 %
	$\hat{\kappa}$ - True Param	0.0874	0.0197	0.107	0.0042	63 %
	κ^* - Est Param	0.0896	0.0248	0.114	0.0060	60 %
50% (22,026)	No Penalty	0.0349	0.0002	0.0350	0.0018	
	κ^* - True Param	0.0277	0.0015	0.0292	0.0013	17 %
	$\hat{\kappa}$ - True Param	0.0277	0.0015	0.0292	0.0013	17 %
	κ^* - Est Param	0.0291	0.0014	0.0305	0.0013	13 %
75% (53,045)	No Penalty	0.0080	0.0000	0.0080	0.0004	
	κ^* - True Param	0.0075	0.0002	0.0077	0.0003	4 %
	$\hat{\kappa}$ - True Param	0.0075	0.0002	0.0077	0.0003	4 %
	κ^* - Est Param	0.0076	0.0002	0.0078	0.0003	3 %
90% (127,744)	No Penalty	0.0038	0.0000	0.0038	0.0002	
	κ^* - True Param	0.0038	0.0000	0.0038	0.0001	1 %
	$\hat{\kappa}$ - True Param	0.0037	0.0001	0.0038	0.0001	1 %
	κ^* - Est Param	0.0038	0.0000	0.0038	0.0001	2 %

Simulation Results: left-truncated Weibull(10,0.5), sample size 100

Trunc % (Value)	Estimator	Variance	Bias	MSE	SE Err	% Redux
25% (5.36)	No Penalty	0.84	0.000	0.84	0.044	
	κ^* - True Param	0.68	0.057	0.74	0.032	12 %
	$\hat{\kappa}$ - True Param	0.68	0.057	0.74	0.032	12 %
	κ^* - Est Param	0.69	0.052	0.74	0.033	12 %
50% (8.33)	No Penalty	2.1	0.000	2.1	0.098	
	κ^* - True Param	1.3	0.38	1.7	0.064	20 %
	$\hat{\kappa}$ - True Param	1.6	0.12	1.8	0.070	18 %
	κ^* - Est Param	1.4	0.45	1.9	0.072	12 %
75% (11.7)	No Penalty	6.1	0.000	6.1	0.303	
	κ^* - True Param	2.5	2.4	4.8	0.155	21 %
	$\hat{\kappa}$ - True Param	3.7	0.86	4.5	0.179	26 %
	κ^* - Est Param	2.6	2.7	5.3	0.185	14 %
90% (15.2)	No Penalty	15.1	0.001	15.2	0.605	
	κ^* - True Param	3.8	10.4	14.2	0.369	7 %
	$\hat{\kappa}$ - True Param	9.3	1.7	11.1	0.389	27 %
	κ^* - Est Param	3.7	15.5	19.6	0.531	-29 %

Extreme distortions in the Weibull likelihood

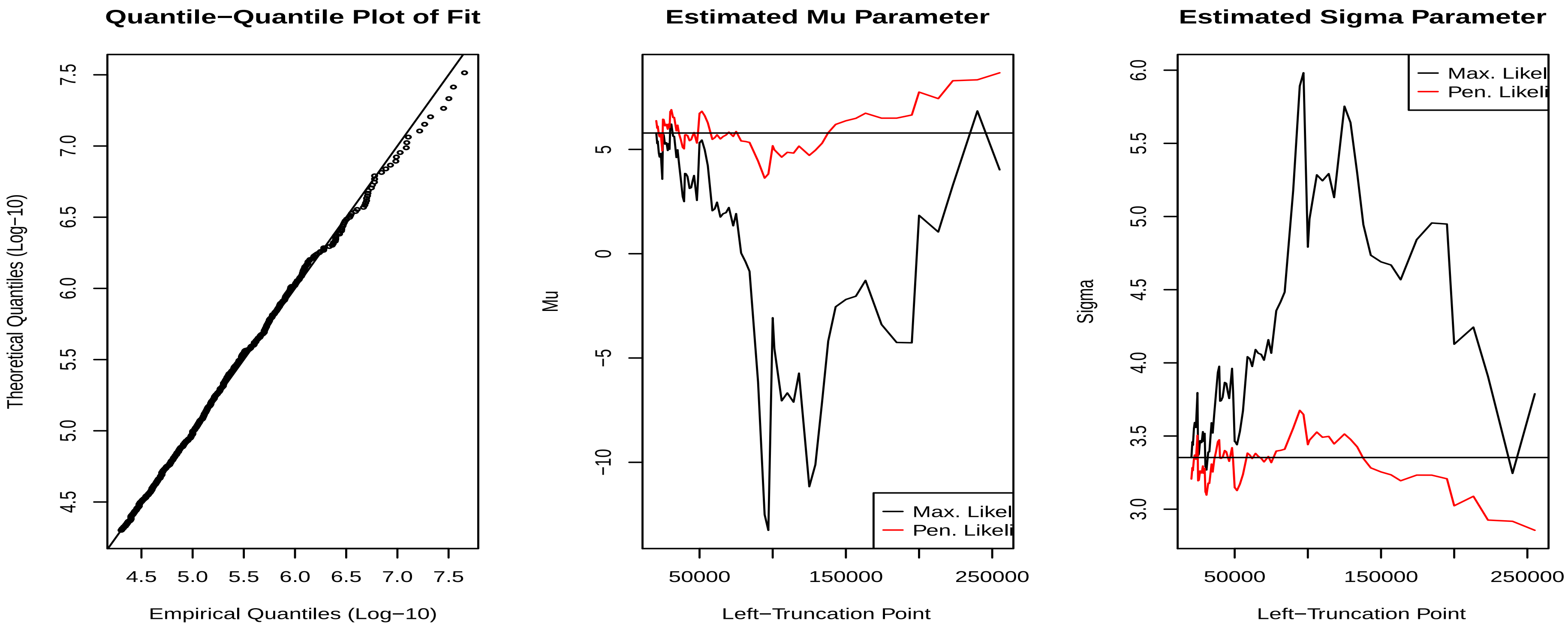




- Estimated parameters from 1,000 samples of a left-truncated Weibull(10,0.5) truncated at the 90th percentile (15.2)
  – Left graph shows estimates with sample size 100, right graph sample size 1,000

Improved stability of the penalized likelihood estimator

- Data: Private Banking / Clients Products and Business Practices 2002Q1 – 2008Q2 (1,309 data values); fit using Lognormal distribution over increasing truncation levels



Robustness to data contamination

Fit of Lognormal distributions to data generated from mixture of Lognormals (10,2) + (12,4) with increasing mixture weights

% Mixture[a]	Estimator	μ	σ	Variance	Bias[b]	MSE	SE Err	% Redux
	No Penalty	9.76	2.04	1.732	0.061	1.800	0.177	
0%	κ^* - True Param	10.55	1.75	0.239	0.363	0.602	0.017	67 %
	κ^* - Est Param	10.48	1.77	0.406	0.283	0.693	0.021	61 %
	No Penalty	8.37	2.66	4.873	3.105	8.934	0.593	
5%	κ^* - True Param	10.00	2.19	0.828	0.035	0.863	0.046	90 %
	κ^* - Est Param	10.26	2.05	0.668	0.072	0.913	0.117	90 %
	No Penalty	7.15	3.17	5.652	9.476	17.938	0.924	
10%	κ^* - True Param	9.60	2.53	1.060	0.448	1.509	0.071	92 %
	κ^* - Est Param	10.33	2.20	0.526	0.149	1.043	0.075	94 %
	No Penalty	6.47	3.56	4.855	14.903	23.232	1.096	
15%	κ^* - True Param	9.44	2.78	0.982	0.919	1.900	0.075	92 %
	κ^* - Est Param	10.34	2.36	0.431	0.245	1.155	0.076	95 %
	No Penalty	6.03	3.85	3.989	19.165	27.580	1.276	
20%	κ^* - True Param	9.29	3.01	1.062	1.523	2.585	0.090	91 %
	κ^* - Est Param	10.45	2.48	0.406	0.433	1.497	0.096	95 %

Conclusions

- Penalized likelihood estimator achieves improved error over maximum likelihood estimator for truncated data in most cases
  - Best applied to estimation on small sample sizes
  - Automatic selection of penalty parameter κ^* provides good results but can be improved
  - Also achieves substantially reduced error in high quantile estimation
  - Weibull distribution showed extreme distortions in likelihood surface, leading to poor estimation of good penalty parameters
- More robust and stable outcomes than ML estimates
- Can also be applied to Cramer-von Mises, Anderson-Darling, quantile distance estimators
- Future work: determine how covariates can best be incorporated into penalty scheme