Non-Parametric Insurance Loss Modelling using Variable-Knot Splines

Dimitrina D. Dimitrova[1]
Vladimir K. Kaishev[1]
Emilio L. Sáenz Guillén[1] (presenter)

Insurance
Data
Science

[1]Faculty of Actuarial Science and Insurance, Bayes Business School.

BAYES
BUSINESS SCHOOL
CITY ST GEORGE'S
UNIVERSITY OF LONDON

DDFS (Density & Distribution Func. variable-knot Spline estimation)

Novel non-parametric method for *simultaneous* variable knot spline estimation of both the pdf and the cdf of a random variable.

DDFS (Density & Distribution Func. variable-knot Spline estimation)

Novel non-parametric method for *simultaneous* variable knot spline estimation of both the pdf and the cdf of a random variable.

1. In the literature, non-parametric density estimation methods address the estimation of *solely the pdf*. However, pdf and cdf are closely connected.

DDFS (Density & Distribution Func. variable-knot Spline estimation)

Novel non-parametric method for *simultaneous* variable knot spline estimation of both the pdf and the cdf of a random variable.

1. In the literature, non-parametric density estimation methods address the estimation of *solely the pdf*. However, pdf and cdf are closely connected.
2. Model structure under which, both the pdf and cdf spline models share the *same set of knots* + their *coefficients are connected*.

DDFS (Density & Distribution Func. variable-knot Spline estimation)

Novel non-parametric method for *simultaneous* variable knot spline estimation of both the pdf and the cdf of a random variable.

1. In the literature, non-parametric density estimation methods address the estimation of *solely the pdf*. However, pdf and cdf are closely connected.
2. Model structure under which, both the pdf and cdf spline models share the *same set of knots* + their *coefficients are connected*.
3. Iterative approach combining:
   - **Constrained maximum likelihood estimation** of the spline coefficients.
   - **Sequential minimum bias driven estimation** of the underlying spline knots, following the Geometrically Designed Spline (GeDS) methodology (Kaishev et al., 2016, Dimitrova et al., 2023, Dimitrova et al., 2025).

DDFS (Density & Distribution Func. variable-knot Spline estimation)

Novel non-parametric method for *simultaneous* variable knot spline estimation of both the pdf and the cdf of a random variable.

1. In the literature, non-parametric density estimation methods address the estimation of *solely the pdf*. However, pdf and cdf are closely connected.
2. Model structure under which, both the pdf and cdf spline models share the *same set of knots* + their *coefficients are connected*.
3. Iterative approach combining:
   - **Constrained maximum likelihood estimation** of the spline coefficients.
   - **Sequential minimum bias driven estimation** of the underlying spline knots, following the Geometrically Designed Spline (GeDS) methodology (Kaishev et al., 2016, Dimitrova et al., 2023, Dimitrova et al., 2025).
4. Competitive alternative to state of the art methods, e.g., kernel, logsplines (Kooperberg and Stone, 1991), and recent spline-based methods (Cui et al., 2020, Kirkby et al., 2021) + *Large sample properties*.

Iterative estimation process

Iterative estimation process

Step 1. Given $\boldsymbol{t}_{k,n}$, find the MLE estimates, $\hat{\boldsymbol{\theta}}$, of $f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta})$.

Iterative estimation process

Step 1. Given $\boldsymbol{t}_{k,n}$, find the MLE estimates, $\hat{\boldsymbol{\theta}}$, of $f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta})$.

Step 2. Use $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{t}_{k,n}$ to compute $\hat{\boldsymbol{\theta}}'$, and then compute $F(x; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}')$, recalling that, $\boldsymbol{t}_{k,n+1}$ is obtained from $\boldsymbol{t}_{k,n}$, by adding the additional end knots, $t_0 = t_1$, and $t_{2n+k+1} = t_{2n+k}$.

Iterative estimation process

Step 1. Given $\boldsymbol{t}_{k,n}$, find the MLE estimates, $\hat{\boldsymbol{\theta}}$, of $f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta})$.
Step 2. Use $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{t}_{k,n}$ to compute $\hat{\boldsymbol{\theta}}'$, and then compute $F(x; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}')$, recalling that, $\boldsymbol{t}_{k,n+1}$ is obtained from $\boldsymbol{t}_{k,n}$, by adding the additional end knots, $t_0 = t_1$, and $t_{2n+k+1} = t_{2n+k}$.
Step 3. Compute the residuals

$$\rho_i = F_N(X_i) - F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}'), \quad i = 1, ..., N, \tag{1}$$

that provide information for the discrepancy between the ecdf, $F_N(x)$, and the estimate, $F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}')$, of the cdf F.

On the k-th iteration, if $k \leq q$, go to Step 4. Otherwise, use the residuals, ρ_i, $i = 1, \ldots, N$, to compute the ratio

$$\phi = \frac{\sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2}{\sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k-q,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2}. \tag{2}$$

If $\phi \geq \phi_{exit}$, then exit the iterations with final estimates, $f(x; \boldsymbol{t}_{k-q,n}, \hat{\boldsymbol{\theta}})$ and $F(x; \boldsymbol{t}_{k-q,n+1}, \hat{\boldsymbol{\theta}}')$. If $\phi < \phi_{exit}$, then go to Step 4.

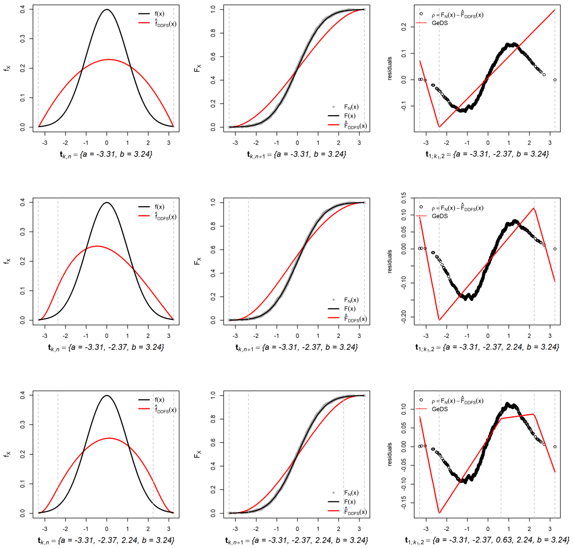
On the k-th iteration, if $k \leq q$, go to Step 4. Otherwise, use the residuals, ρ_i, $i = 1, \ldots, N$, to compute the ratio

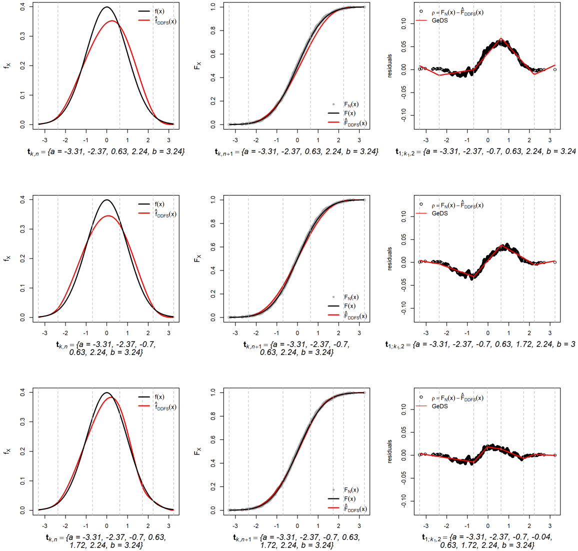
$$\phi = \frac{\sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2}{\sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k-q,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2}. \tag{2}$$

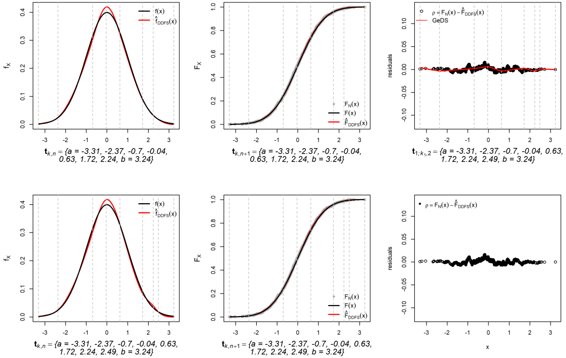
If $\phi \geq \phi_{exit}$, then exit the iterations with final estimates, $f(x; \boldsymbol{t}_{k-q,n}, \hat{\boldsymbol{\theta}})$ and $F(x; \boldsymbol{t}_{k-q,n+1}, \hat{\boldsymbol{\theta}}')$. If $\phi < \phi_{exit}$, then go to Step 4.

Step 4. Find a new knot δ^*, applying the *locally adaptive bias minimizing knot insertion scheme* of Stage A of GeDS, viewing ρ_i, from (1), as the observations y_i, $i = 1, \ldots, N$.

Update the current set of knots as, $\boldsymbol{t}_{k+1,n+1} = \boldsymbol{t}_{k,n+1} \cup \delta^*$. Set, $\boldsymbol{t}_{k,n} \leftarrow \boldsymbol{t}_{k+1,n+1}$ and go back to Step 1.


f(x)
$\hat{f}_{DDFS}(x)$
$F_N(x)$
F(x)
$\hat{F}_{DDFS}(x)$
$\rho = F_N(x) - \hat{F}_{DDFS}(x)$
GeDS
residuals
$\mathbf{t}_{k,n} = \{a = -3.31, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, b = 3.24\}$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, 2.24, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, 0.63, 2.24, b = 3.24\}$


f(x)
$\hat{f}_{DDFS}(x)$
$F_N(x)$
F(x)
$\hat{F}_{DDFS}(x)$
$\rho = F_N(x) - \hat{F}_{DDFS}(x)$
GeDS
f_X
F_X
residuals
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, 0.63, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, 0.63, 2.24, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, -0.7, 0.63, 2.24, b = 3.24$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, -0.7, 0.63, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, -0.7, 0.63, 2.24, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, -0.7, 0.63, 1.72, 2.24, b = 3$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, -0.7, 0.63, 1.72, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, -0.7, 0.63, 1.72, 2.24, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, b = 3.24\}$


$f(x)$
$\hat{f}_{DDFS}(x)$
f_X
$F_N(x)$
$F(x)$
$\hat{F}_{DDFS}(x)$
F_X
$\rho = F_N(x) - \hat{F}_{DDFS}(x)$
GeDS
residuals
x
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, 2.49, b = 3.24\}$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, 2.49, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, 2.49, b = 3.24\}$

Test case	Density	Test case	Density
Gaussian	$N(0,1)$	Merton's jump diffusion	$\sigma = 0.08$, $\lambda = 3$ $\mu_J = -0.01$, $\sigma_J = 0.4$, $\Delta_t = 1/4$
Student-t	t_v, $v = 6$	Kou's double exponential	$\sigma = 0.04$, $\lambda = 2$ $p_{up} = 0.4$, $\eta_1 = 3$, $\eta_2 = 5$, $\Delta_t = 1/4$
Exponential	Exp(λ), $\lambda = 1$, $x \geq 0$	Generalized extreme value	$\frac{1}{\sigma}\exp\left(-\left(1+k\frac{x-\mu}{\sigma}\right)^{-\frac{1}{k}}\right)(1+k\frac{x-\mu}{\sigma})^{-1-\frac{1}{k}}$ $k = -0.5, 0$ and 0.5, $\mu = 1$, $\sigma = 0$, $x > 0$
Chi-square	χ^2_k, $k = 4$, $x > 0$	MixGauss	$0.15N(-0.25, 1/3) + 0.85N(3.25, 1)$
Gamma	$x \geq 0$, $k = 9$, $\theta = 0.5$	Mix1d	$0.8\chi^2(3) + 0.2N(7,1)$
Weibull	$x \geq 0$, $\lambda = 1$, $k = 5$	MixGauss2	$\frac{5}{6}N(3,1) + \frac{5}{36}N(8,(1/3)^2) + \frac{1}{36}N(10,(1/9)^2)$
Log-normal	$x > 0$, $\mu = 0$, $\sigma = 1$	Bimodal	$\frac{1}{2}N\left(0, \left(\frac{1}{10}\right)^2\right) + \frac{1}{2}N(5,1)$
Nakagami	$\frac{2\mu^\mu x^{2\mu-1}}{\Gamma(\mu)\omega^\mu}\exp\left(-\frac{\mu}{\omega}x^2\right)$, $x > 0$, $\mu = \omega = 2$	Separated bimodal	$\frac{1}{2}N(-2, \left(\frac{1}{2}\right)^2) + \frac{1}{2}N(2, \left(\frac{1}{2}\right)^2)$
Kurtotic unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N\left(0, \left(\frac{1}{10}\right)^2\right)$	Skewed bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2}, \left(\frac{1}{3}\right)^2)$
Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N\left(0, \left(\frac{1}{10}\right)^2\right)$	Trimodal	$\frac{1}{3}\sum_{k=0}^{2} N\left(80k, (k+1)^4\right)$
Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2}, \left(\frac{2}{3}\right)^2) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$	Smooth comb	$\sum_{k=0}^{5}\frac{2^{5-k}}{63}N\left(\frac{65-96/2^k}{21}, \left(\frac{32/63}{2^k}\right)^2\right)$
Strongly skewed	$\sum_{k=0}^{7}\frac{1}{8}N\left(3\left(\left(\frac{2}{3}\right)^k - 1\right), \left(\frac{2}{3}\right)^{2k}\right)$	Claw	$\frac{1}{2}N(0,1) + \sum_{k=0}^{4}\frac{1}{10}N\left(\frac{k}{2}-1, \left(\frac{1}{10}\right)^2\right)$

We assess the goodness-of-fit over a regular grid of K evaluation points $x_1, \ldots, x_K$ with uniform spacing $\Delta x = x_{i+1} - x_i$, based on:

- **Mean Integrated Squared Error (MISE)**

$$\mathbb{E}\left[\int \left(\hat{f}(x;N) - f(x)\right)^2 dx\right] \approx \Delta x \sum_{i=1}^{K} \left(\hat{f}(x_i;N) - f(x_i)\right)^2$$

We assess the goodness-of-fit over a regular grid of K evaluation points $x_1, \ldots, x_K$ with uniform spacing $\Delta x = x_{i+1} - x_i$, based on:

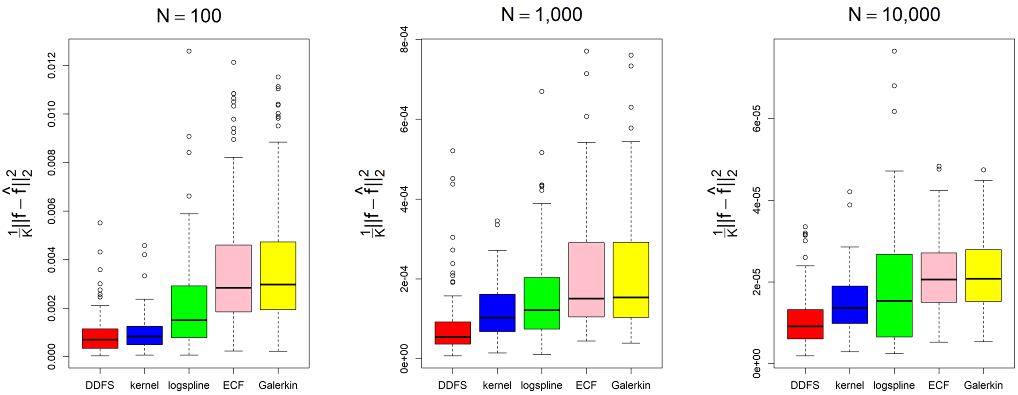
- **Mean Integrated Squared Error (MISE)**

$$\mathbb{E}\left[\int \left(\hat{f}(x;N) - f(x)\right)^2 dx\right] \approx \Delta x \sum_{i=1}^{K} \left(\hat{f}(x_i;N) - f(x_i)\right)^2$$

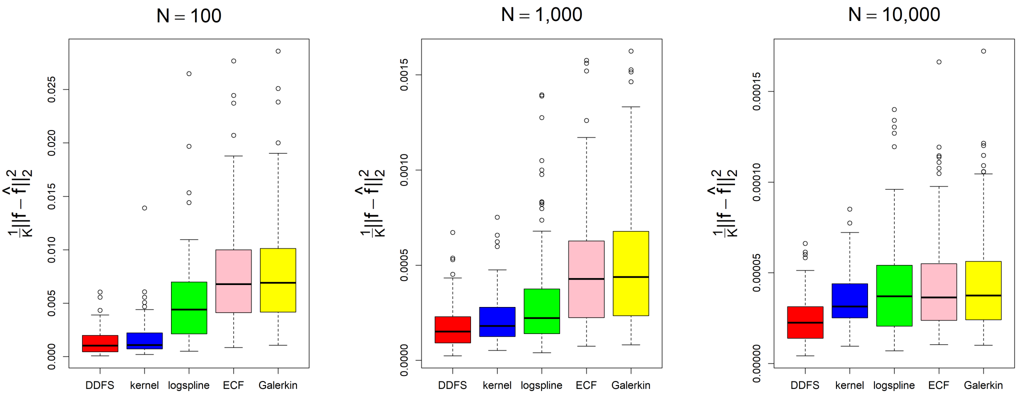
- **Roughness, $R(f)$**

$$R(f) = \int \left(f''(x)\right)^2 dx \approx \Delta x \sum_{i=2}^{K-1} \left(\frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{(\Delta x)^2}\right)^2$$

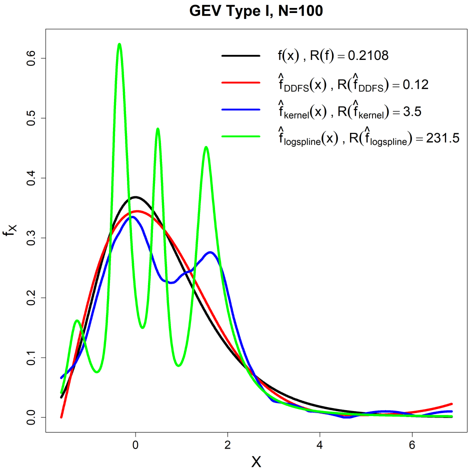
MISE boxplots - Generalized Extreme Value, Type I (Gumbel)

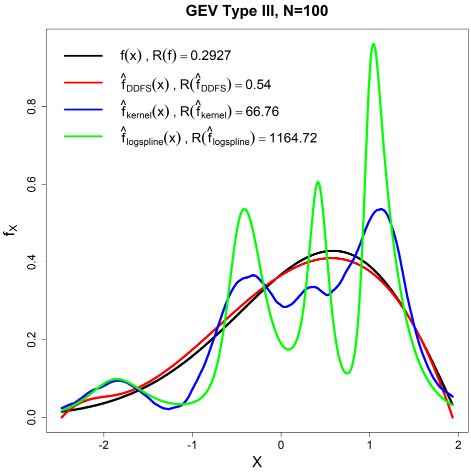


MISE boxplots - Generalized Extreme Value, Type III (Weibull)

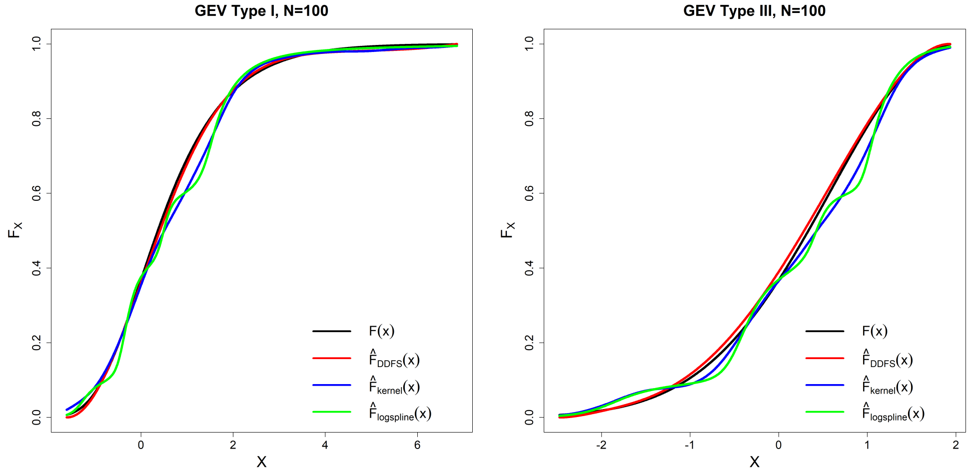


Roughness - Generalized Extreme Value, Type I and III (pdf)

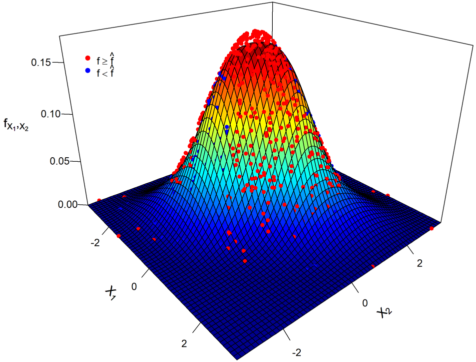


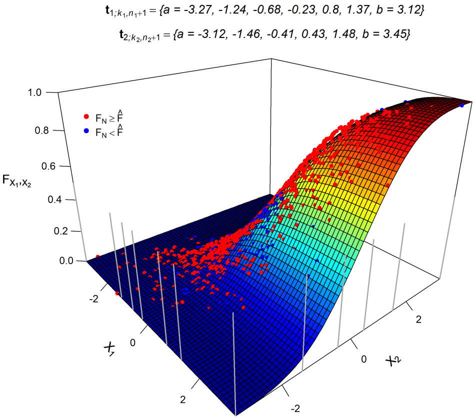


Roughness - Generalized Extreme Value, Type I and III (cdf)

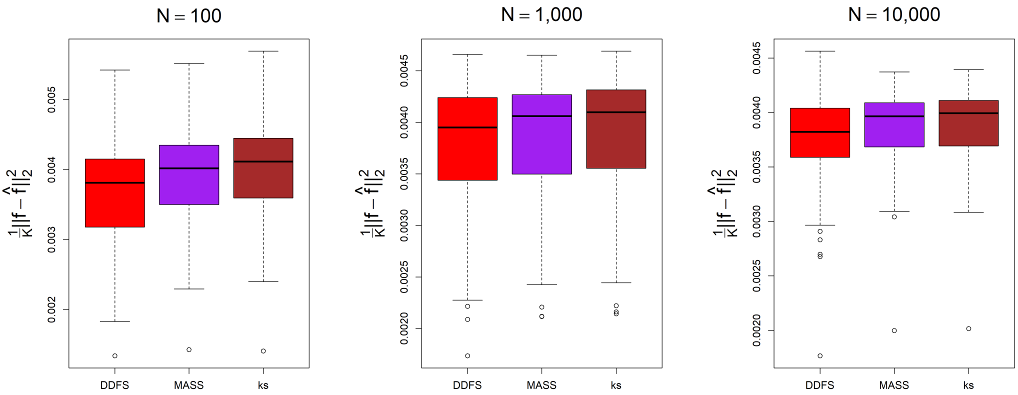


Bivariate Gaussian, $\rho = 0.5$





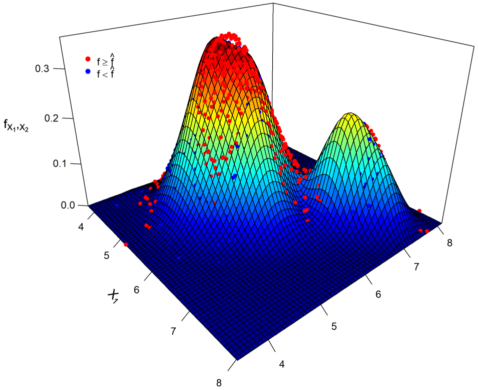
MISE boxplots - Bivariate Gaussian, $\rho = 0.5$

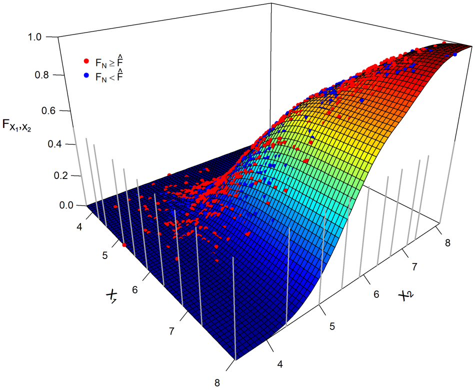


Bivariate Bimodal Mixed Gaussian

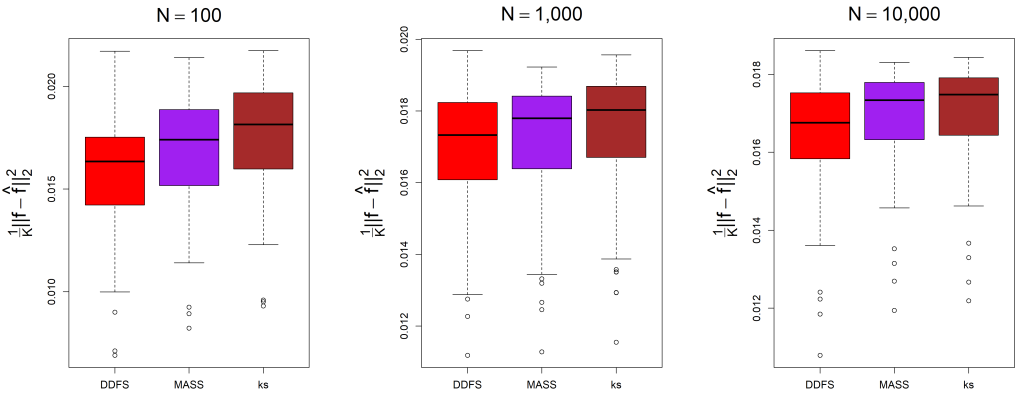
$\mathbf{t}_{1;k_1,n_1+1} = \{a = 3.74, 4.26, 4.55, 5.18, 5.54, 5.92, 6.39, 6.87, 7.33, b = 8.02\}$

$\mathbf{t}_{2;k_2,n_2+1} = \{a = 3.48, 4.36, 5, 5.49, 5.9, 6.34, 6.68, 7.14, 7.38, b = 8.14\}$





Bivariate Bimodal Mixed Gaussian



Loss data modelling

Proliferation of composite and mixture (parametric) models proposed for the modelling of insurance loss data (Marambakuyana and Shongwe, 2024).

Loss data modelling

Proliferation of composite and mixture (parametric) models proposed for the modelling of insurance loss data (Marambakuyana and Shongwe, 2024).

- Estimation often requires **large samples to converge** (data availability might be a problem) + **high computational burden**: $\longrightarrow$ many actuarial studies only fit the model once (on the whole dataset) and perform backtesting on a static forecast (see, e.g., Abu Bakar et al., 2015);

✗ *rolling-window historical simulation backtesting*.

Some frequently considered loss-datasets are:

1. *Danish Fire Insurance*.
2. *Norwegian Fire Insurance Data*.
3. *US Allocated Loss Adjustment Expenses*.

➤ Assess model reliability via **rolling window back-testing** (Basel III, EIOPA): *Proportion of Failures* (Kupiec, 1995), *Conditional Coverage* (Christoffersen, 1998), *Dynamic Quantile* (Engle and and, 2004).

Danish Fire Insurance

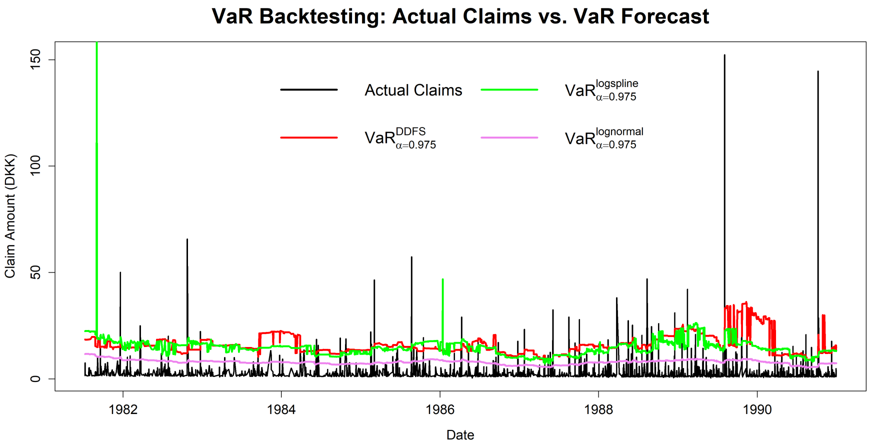


Table 1: Backtesting statistics for VaR models (data = danish).

	$\alpha = 0.975$				$\alpha = 0.99$			
	Viol. rate	UC	CC	DQ	Viol. rate	UC	CC	DQ
ddfs	0.0250	0.9946	0.1776	0.0244	0.0103	0.9024	0.7819	0.0349
kernel	0.0259	0.7931	0.0623	0.0001	0.0116	0.4586	0.5600	0.0006
logspline	0.0245	0.8867	0.4458	0.1013	0.0120	0.3462	0.4617	0.0256
lognormal	0.0531	0.0000	0.0000	0.0000	0.0397	0.0000	0.0000	0.0000
gamma	0.0437	0.0000	0.0000	0.0000	0.0361	0.0000	0.0000	0.0000

Norwegian Fire Insurance Data

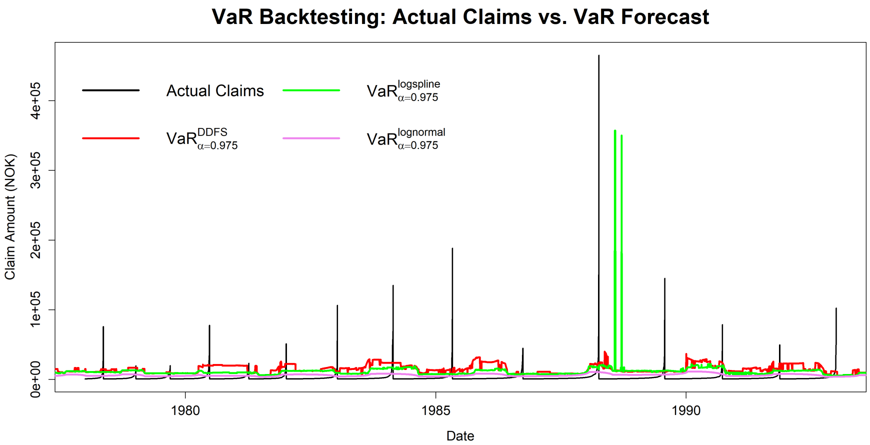


Table 2: Backtesting statistics for VaR models (data = norwegian).

	$\alpha = 0.975$				$\alpha = 0.99$			
	Viol. rate	UC	CC	DQ	Viol. rate	UC	CC	DQ
ddfs	0.0265	0.3816	0.0000	0.0206	0.0100	0.9832	0.0000	0.0001
kernel	0.0351	0.0000	0.0000	0.0078	0.0156	0.0000	0.0000	0.0000
logspline	0.0356	0.0000	0.0000	0.0074	0.0147	0.0001	0.0000	0.0000
lognormal	0.0583	0.0000	0.0000	0.0032	0.0390	0.0000	0.0000	0.0000
gamma	0.0466	0.0000	0.0000	0.0048	0.0336	0.0000	0.0000	0.0000

US Allocated Loss Adjustment Expenses

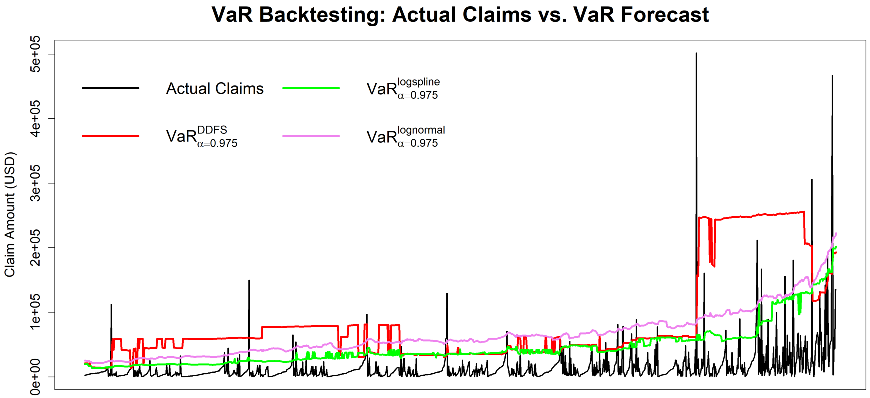


Table 3: Backtesting statistics for VaR models (data = lossalae).

	$\alpha = 0.975$				$\alpha = 0.99$			
	Viol. rate	UC	CC	DQ	Viol. rate	UC	CC	DQ
ddfs	0.0256	0.8923	0.0000	0.9663	0.0112	0.6757	0.0000	0.6161
kernel	0.0448	0.0001	0.0000	0.4172	0.0224	0.0002	0.0000	0.2674
logspline	0.0448	0.0001	0.0000	0.4292	0.0272	0.0000	0.0000	0.0133
lognormal	0.0200	0.2410	0.0218	0.9555	0.0072	0.2950	0.5413	0.6525
gamma	0.0488	0.0000	0.0000	0.4018	0.0360	0.0000	0.0000	0.0040

Abu Bakar, S., Hamzah, N., Maghsoudi, M., & Nadarajah, S. (2015).Modeling loss data using composite models. *Insurance: Mathematics and Economics*, *61*, 146–154. https://doi.org/https://doi.org/10.1016/j.insmatheco.2014.08.008

Christoffersen, P. F. (1998).Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862. Retrieved May 14, 2025, from http://www.jstor.org/stable/2527341

Cui, Z., Kirkby, J. L., & Nguyen, D. (2020).Nonparametric density estimation by b-spline duality. *Econometric Theory*, *36*(2), 250–291. https://doi.org/10.1017/S0266466619000112

Dimitrova, D. S., Kaishev, V. K., Lattuada, A., & Verrall, R. J. (2023).Geometrically designed variable knot splines in generalized (non-)linear models [© 2022. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/]. *Applied Mathematics and Computation*, *436*. https://doi.org/10.1016/j.amc.2022.127493

Dimitrova, D. S., Kaishev, V. K., & Saenz Guillen, E. (2025). *Geds: An r package for regression, generalized additive models and functional gradient boosting, based on geometrically designed (ged) splines* [Manuscript submitted for publication].

Engle, R. F., & and, S. M. (2004).Caviar. *Journal of Business & Economic Statistics*, *22*(4), 367–381. https://doi.org/10.1198/073500104000000370

Kaishev, V. K., Dimitrova, D. S., Haberman, S., & Verrall, R. J. (2016).Geometrically designed, variable knot regression splines. *Computational Statistics*, *31*(3), 1079–1105. https://doi.org/10.1007/s00180-015-0621-7

Kirkby, J. L., Leitao, Á., & Nguyen, D. (2021).Nonparametric density estimation and bandwidth selection with b-spline bases: A novel galerkin method. *Computational Statistics & Data Analysis*, *159*, 107202. https://doi.org/https://doi.org/10.1016/j.csda.2021.107202

Kooperberg, C., & Stone, C. J. (1991).A study of logspline density estimation. *Computational Statistics & Data Analysis*, *12*(3), 327–347. https://doi.org/https://doi.org/10.1016/0167-9473(91)90115-I

Kupiec, P. H. (1995).Techniques for verifying the accuracy of risk measurement models [Available at SSRN: https://ssrn.com/abstract=7065]. *The Journal of Derivatives*, *3*(2), 73–84.

Marambakuyana, W. A., & Shongwe, S. C. (2024).Composite and mixture distributions for heavy-tailed data—an application to insurance claims. *Mathematics*, *12*(2), 335.

Bayes Business School
106 Bunhill Row
London EC1Y 8TZ
Tel +44 (0)20 7040 8600
bayes.city.ac.uk