Estimation of Probability Density and Distribution Functions using Variable-Knot Splines with Applications to Insurance Loss Data

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

Perspectives on Actuarial Risks in Talks of Young Researchers 2025

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

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.

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

- The **pdf, $f(x)$**, is approximated by a spline function, from the space of all n-th order spline functions, $S_{\boldsymbol{t}_{k,n}}$,

$$f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{N}_n(x) = \sum_{j=1}^{p} \theta_j N_{j,n}(x), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ is a vector of *non-negative coefficients* and $\boldsymbol{N}_n(x) = (N_{1,n}(x), \ldots, N_{p,n}(x))^T$, $p = n + k$, are normalized B-splines of order n, defined on the set of non-decreasing knots,

$$\boldsymbol{t}_{k,n} = \{t_1 = \ldots = t_n < t_{n+1} < \ldots < t_{n+k} < t_{n+k+1} = \ldots = t_{2n+k}\}. \tag{2}$$

In order for $f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta})$, to *integrate to one* and be a valid pdf, then

$$\sum_{j=1}^{p} \theta_j \frac{(t_{j+n} - t_j)}{n} = 1, \text{ must hold.} \tag{3}$$

- For the **cdf, $F(x)$,** it is natural to take

$$F(x;\boldsymbol{t}_{k,n+1},\boldsymbol{\theta}') = \int_{\underline{x}}^{x} f(u;\boldsymbol{t}_{k,n};\boldsymbol{\theta})du = \sum_{j=1}^{p} \theta'_j N_{j,n+1}(x), \quad \underline{x} \leq x \leq \overline{x}, \tag{4}$$

where

$$\boldsymbol{\theta}' = \left(\theta'_1, \ldots, \theta'_p\right)^T, \quad \theta'_j = \sum_{i=1}^{j} \theta_i \frac{t_{i+n} - t_i}{n}, \tag{5}$$

and $N_{j,n+1}(x), j = 1, \ldots, p$, are B-splines of order $n+1$, on the set of non-decreasing knots

$$\begin{aligned}\boldsymbol{t}_{k,n+1} = \{&t_0 = t_1 = \ldots = t_n < t_{n+1} < \ldots \\ &\ldots < t_{n+k} < t_{n+k+1} = \ldots = t_{2n+k} = t_{2n+k+1},\}\end{aligned} \tag{6}$$

with $N_{j,n+1}(x)$ supported on $[t_j, t_{j+n+1}]$, $j = 1, \ldots, p$.

Univariate pdf/cdf estimation

1. Given a set of knots $\{t_{n+1}, \ldots, t_{n+k}\} \subset \boldsymbol{t}_{k,n} \subset \boldsymbol{t}_{k,n+1}$, we find the maximum likelihood estimates (MLE) of the parameters, $\boldsymbol{\theta}$, by solving the optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log f(X_i; \boldsymbol{t}_{k,n}, \boldsymbol{\theta}), \tag{7}$$

subject to the constraints

$$\theta_j \geq 0, \quad j = 1, \ldots, p, \quad \text{and} \quad \sum_{j=1}^{p} \theta_j \frac{t_{j+n} - t_j}{n} = 1. \tag{8}$$

2. For a MLE estimate, $\hat{\boldsymbol{\theta}}$, obtained from (7), and estimates, $\hat{\boldsymbol{\theta}}'$, obtained from (5), find an updated set of knots, $\hat{\boldsymbol{t}}_{k,n+1}$, such that it minimizes the L_2-distance between $F_N(x)$ and $F(x; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}')$:

$$\hat{\boldsymbol{t}}_{k,n+1} = \arg \min_{\boldsymbol{t}_{k,n+1}} \sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2 \tag{9}$$

subject to the constraints $\{t_{n+1} < \ldots < t_{n+k}\} \subset \boldsymbol{t}_{k,n} \subset \boldsymbol{t}_{k,n+1}$.

- Since, k may be large, (9) is in general a *highly multivariate* optimization problem, *non-linear* in $t_{n+1}, \ldots, t_{n+k}$, that is virtually impossible to solve (see De Boor, 1978, Lindstrom, 1999).

- Therefore, instead of (9), we sequentially estimate the knots, $t_{n+1}, \ldots, t_{n+k}$, applying the procedure underpinning Stage A of the **Geometrically Designed Splines (GeDS)** (Kaishev et al., 2016 and Dimitrova et al., 2023):
  - At each iteration we place a knot, δ^*, within the cluster that maximizes the following-bias dominated measure:

$$w_j = \beta m'_j + (1-\beta)\eta'_j, j = 1, ..., u \tag{10}$$

where m'_j, η'_j denote the normalized mean and range values of the j-th cluster of residuals.

Iterative estimation process

Step 1. Given $\boldsymbol{t}_{k,n}$ find the MLE estimates, $\hat{\boldsymbol{\theta}}$, by solving (7).
Step 2. Use $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{t}_{k,n}$ to compute $\hat{\boldsymbol{\theta}}'$, following (5) and then apply (4) to 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 $t_{2n+k+1} = t_{2n+k}$, (cf. (2) and (6)).
Step 3. Compute the residuals

$$\rho_i = F_N(X_i) - F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}'), \quad i = 1, \ldots, N, \tag{11}$$

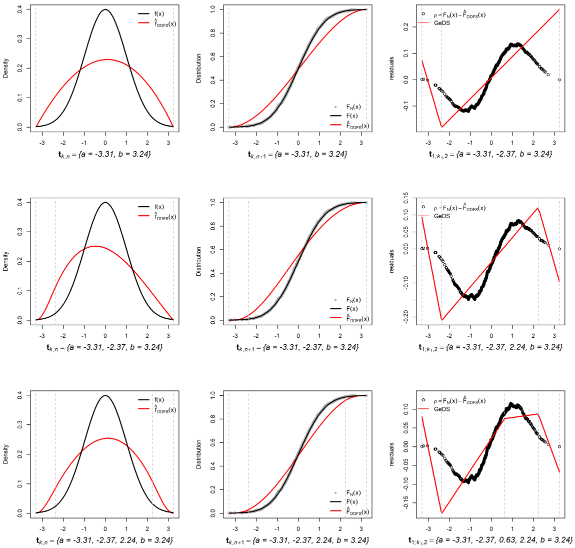
that provide information for the discrepancy between the ecdf, $F_N(x)$, and the estimate, $F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}')$, of the cdf F.

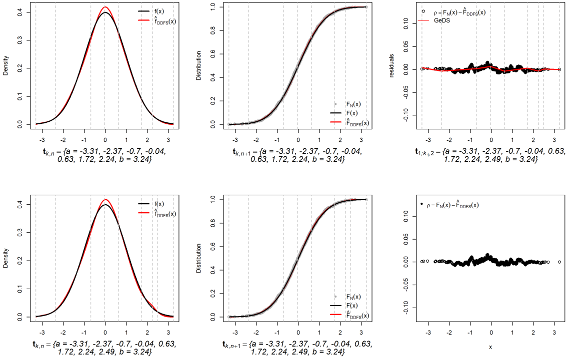
On the k-th iteration, if $k \leq q$, go to Step 4. Otherwise, use the residuals, ρ_i, $i = 1, \ldots, N$, to compute the ratio

$$\phi = \frac{\sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2}{\sum_{i=1}^{N} \left\{ F_N(X_i) - F(X_i; \boldsymbol{t}_{k-q,n+1}, \hat{\boldsymbol{\theta}}') \right\}^2}. \tag{12}$$

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 (11), 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)
$\hat{f}_{DDFS}(x)$
Density
Distribution
$F_N(x)$
F(x)
$\hat{F}_{DDFS}(x)$
residuals
$\rho = F_N(x) - \hat{F}_{DDFS}(x)$
GeDS
$\mathbf{t}_{k,n} = \{a = -3.31, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, b = 3.24\}$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, 2.24, b = 3.24\}$
$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, 2.24, b = 3.24\}$
$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, 0.63, 2.24, b = 3.24\}$



$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, b = 3.24\}$

$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, b = 3.24\}$

$\mathbf{t}_{1;k_1,2} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, 2.49, b = 3.24\}$

$\mathbf{t}_{k,n} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, 2.49, b = 3.24\}$

$\mathbf{t}_{k,n+1} = \{a = -3.31, -2.37, -0.7, -0.04, 0.63, 1.72, 2.24, 2.49, b = 3.24\}$

Multivariate extension

- Without loss of generality, for notational simplicity, we will layout the method for the case of two dimensions.

We now assume, we can approximate the pdf, $f(\boldsymbol{x})$, with a two-dimensional tensor product spline function, $f(\boldsymbol{x}; \mathbf{T}_{\boldsymbol{k},\boldsymbol{n}}, \boldsymbol{\theta})$, defined as

$$f(\boldsymbol{x}; \mathbf{T}_{\boldsymbol{k},\boldsymbol{n}}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \left(\boldsymbol{N}_{n_1}(x_1; \boldsymbol{t}_{1;k_1,n_1}) \otimes \boldsymbol{N}_{n_2}(x_2; \boldsymbol{t}_{2;k_2,n_2})\right) = \tag{13}$$

$$= \sum_{j_1=1}^{p_1} \sum_{j_2=1}^{p_2} \theta_{j_1 j_2} N_{j_1,n_1}(x_1; \boldsymbol{t}_{1;k_1,n_1}) N_{j_2,n_2}(x_2; \boldsymbol{t}_{2;k_2,n_2})$$

where $\boldsymbol{t}_{1;k_1,n_1}$, $\boldsymbol{t}_{2;k_2,n_2}$ are sets of knots with respect to x_1 and x_2, with k_1 and k_2 internal knots; $p_1 = n_1 + k_1$ and $p_2 = n_2 + k_2$; $\boldsymbol{N}_{n_1}(x_1; \boldsymbol{t}_{1;k_1,n_1})$ and $\boldsymbol{N}_{n_2}(x_2; \boldsymbol{t}_{2;k_2,n_2})$, are vectors of B-spline basis functions of order n_1 and n_2, on the sets of knots $\boldsymbol{t}_{1;k_1,n_1}$ and $\boldsymbol{t}_{2;k_2,n_2}$; $\boldsymbol{\theta}$ is a vector of spline coefficients.

As in the univariate case, for the spline model of $F(\boldsymbol{x})$, we take

$$F(\boldsymbol{x};\mathbf{T}_{\boldsymbol{k},\boldsymbol{n}+\mathbf{1}},\boldsymbol{\theta}^{'}) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(u,v;\mathbf{T}_{\boldsymbol{k},\boldsymbol{n}},\boldsymbol{\theta})dudv \tag{14}$$

$$= \sum_{j_1=1}^{p_1}\sum_{j_2=1}^{p_2} \theta^{'}_{j_1 j_2} N_{j_1,n_1+1}(x_1;\boldsymbol{t}_{1;k_1,n_1+1}) N_{j_2,n_2+1}(x_2;\boldsymbol{t}_{2;k_2,n_2+1}), \tag{15}$$

where,

$$\boldsymbol{\theta}^{'} = \left(\theta^{'}_{11},\theta^{'}_{12},...,\theta^{'}_{1p_2},\ldots,\theta^{'}_{p_1p_2}\right)^T,$$

$$\theta^{'}_{j_1j_2} = \sum_{i_1=1}^{j_1}\sum_{i_2=1}^{j_2} \theta_{i_1i_2} \frac{t_{i_1+n_1}-t_{i_1}}{n_1}\frac{t_{i_2+n_2}-t_{i_2}}{n_2} \tag{16}$$

Multivariate pdf/cdf estimation

1. For fixed numbers of knots, k_1, k_2 and knot locations, $\{\{t_{1,n_1+1},\ldots,t_{1,n_1+k_1}\} \times \{t_{2,n_2+1},\ldots,t_{2,n_2+k_2}\}\} \subset \mathbf{T}_{\boldsymbol{k},\boldsymbol{n}} \subset \mathbf{T}_{\boldsymbol{k},\boldsymbol{n}+\mathbf{1}}$, we find the maximum likelihood estimates (MLE) of the parameters, $\boldsymbol{\theta}$, by solving the optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log f(\boldsymbol{X_i}; \mathbf{T}_{\boldsymbol{k},\boldsymbol{n}}, \boldsymbol{\theta}), \tag{17}$$

subject to the constraints, $\theta_{j_1 j_2} \geq 0$, $j_1 = 1,\ldots,p_1$, $j_2 = 1,\ldots,p_2$, and:

$$\sum_{j_1=1}^{p_1} \sum_{j_2=1}^{p_2} \theta_{j_1,j_2} \frac{(t_{j_1+n_1} - t_{j_1})}{n_1} \frac{(t_{j_2+n_2} - t_{j_2})}{n_2} = 1. \tag{18}$$

2. For a MLE estimate, $\hat{\boldsymbol{\theta}}$, obtained from (17), and estimates, $\hat{\boldsymbol{\theta}}'$, obtained from (16), update the set of knots, $\hat{\mathbf{T}}_{\boldsymbol{k},\boldsymbol{n}+\mathbf{1}} = \boldsymbol{t}_{1;k_1,n_1+1} \times \boldsymbol{t}_{2;k_2,n_2+1}$.

The optimal location of an additional knot in either, $\boldsymbol{t}_{1;k_1,n_1+1}$ or $\boldsymbol{t}_{2;k_2,n_2+1}$, is found, so that a measure of the bias between $F_N(x)$ and $F(\boldsymbol{x}; \mathbf{T}_{\boldsymbol{k},\boldsymbol{n}+\mathbf{1}}, \boldsymbol{\theta}')$ is minimized following the two dimensional generalization of stage A of GeDS (see Dimitrova et al., 2023).

Asymptotic properties of DDFS

- **Lemma 1.** The spline density estimator, $f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta})$ can be expressed as the following mixture

$$f(x; \boldsymbol{t}_{k,n}, \boldsymbol{\theta}) = \alpha_1 f_{L_1}(x; \boldsymbol{\tau}_1) + \ldots + \alpha_p f_{L_p}(x; \boldsymbol{\tau}_p), \tag{19}$$

  where $\alpha_j = \theta_j \frac{t_{j+n} - t_j}{n} \geq 0$, $\sum_{j=1}^{p} \alpha_j = 1$, and where $f_{L_j}(x; \boldsymbol{\tau}_j)$ $j = 1, \ldots, p$, are densities, with respect to the Lebesgue measure, of the linear combinations, L_j, $j = 1, \ldots, p$ of Dirichlet random variables (Lemma 1 follows from (1), (3) and a result due to Ignatov and Kaishev, 1989).
- **Theorem.** For a fixed set of knots, $\boldsymbol{t}_{k,n}$, the MLE estimates $\hat{\boldsymbol{\theta}}$ are *strongly consistent* and *asymptotically normal* as $N \to \infty$ (the result of the theorem follows from Lemma 1 and a result due to Redner and Walker, 1984).

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

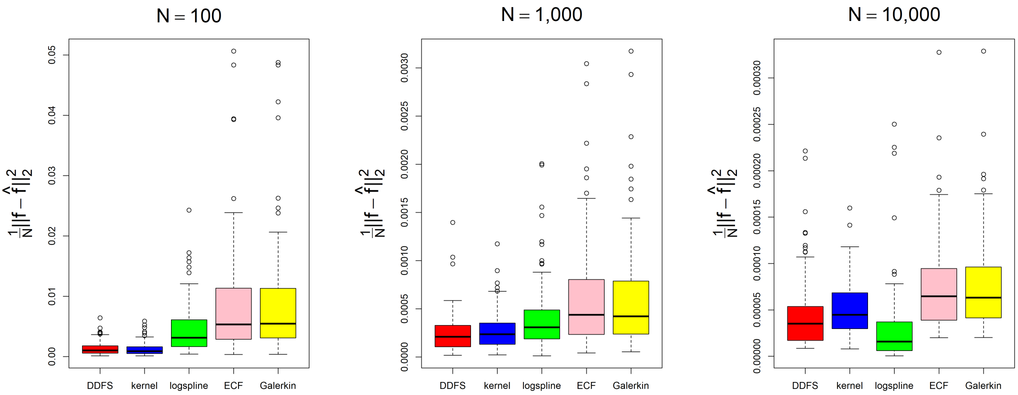
- **Mean Integrated Squared Error (MISE)**:

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

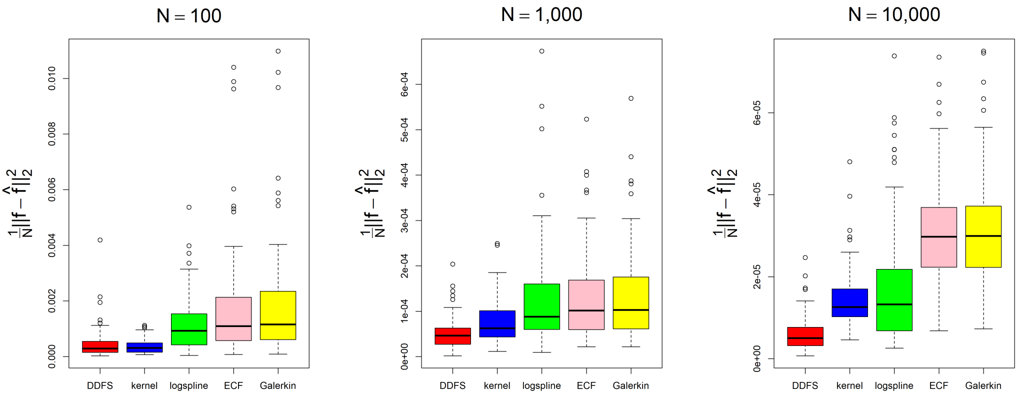
- **Roughness,** $R(f)$:

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

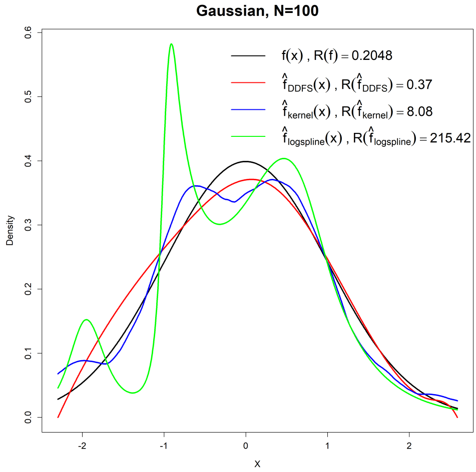
MISE boxplots - Gaussian

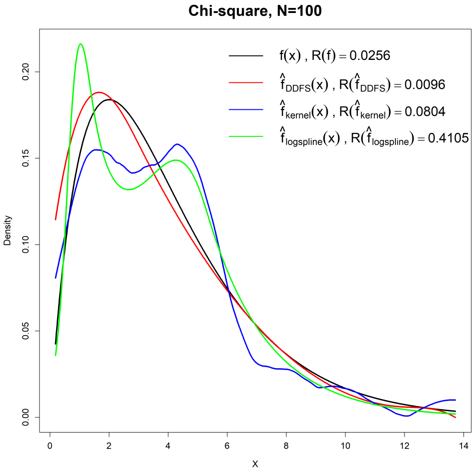


MISE boxplots - Chi-square

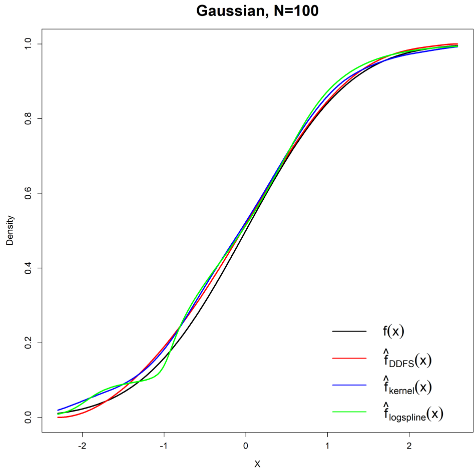


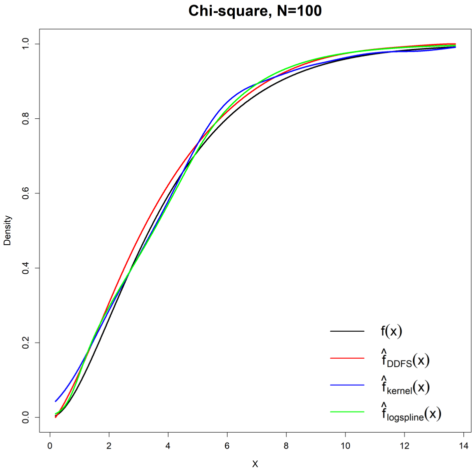
Roughness - Gaussian & Chi-square (pdf)



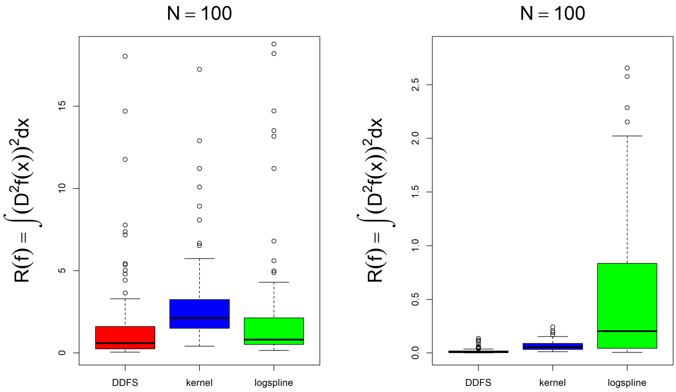


Roughness - Gaussian & Chi-square (cdf)

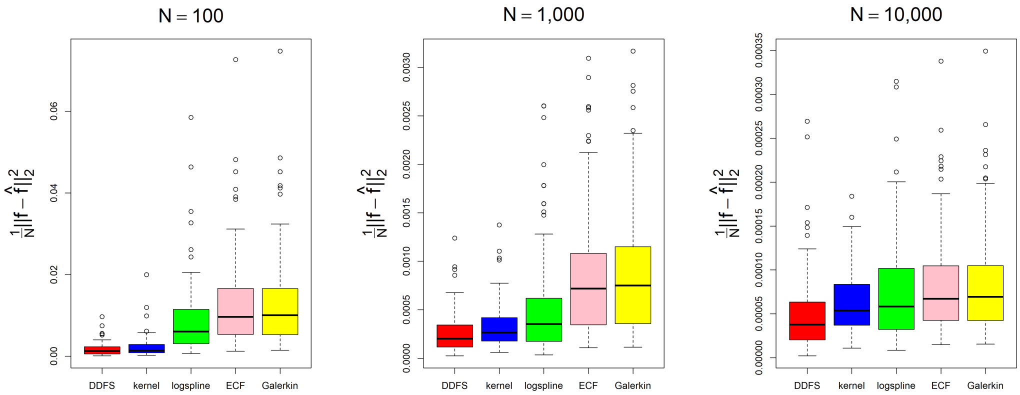




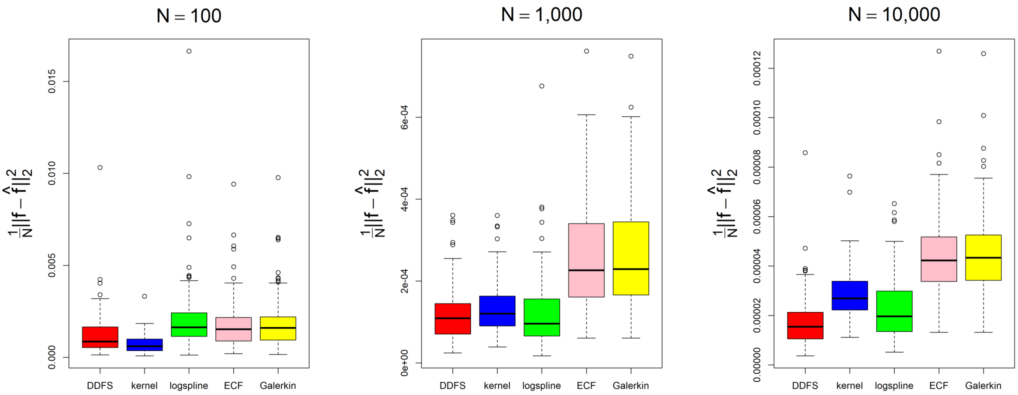
Roughness boxplots - Gaussian & Chi-square



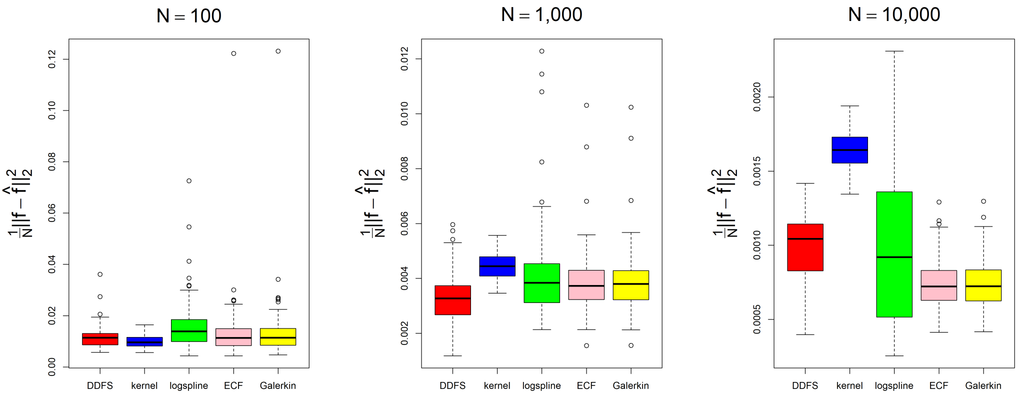
MISE boxplots - Generalized extreme value, Type III



MISE boxplots - Mix1d



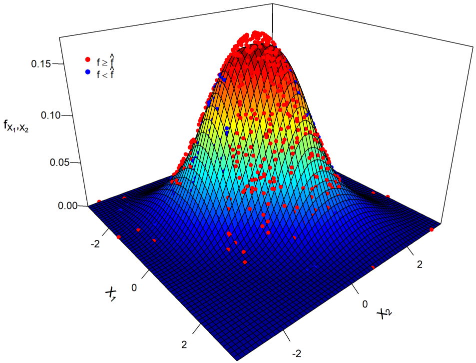
MISE boxplots - Smooth comb

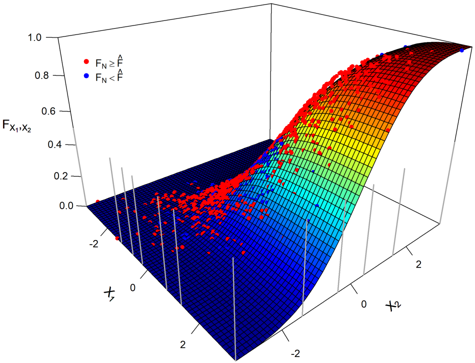


Bivariate Gaussian, $\rho = 0.5$

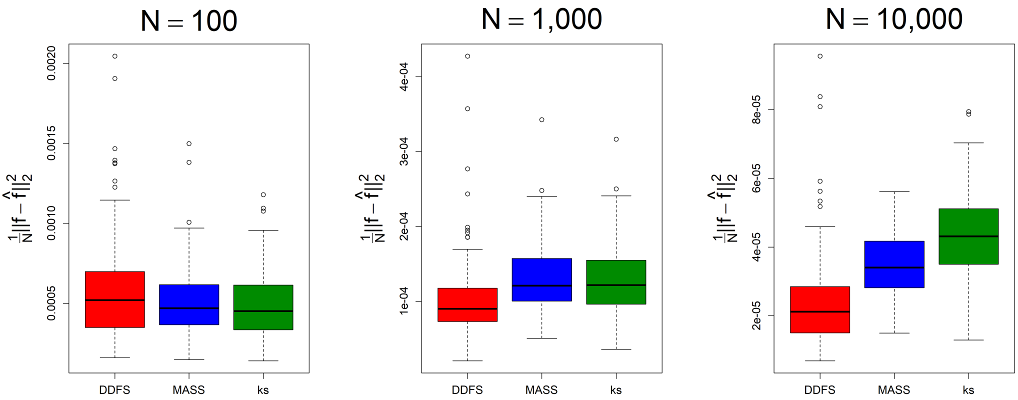
$\mathbf{t}_{1;k_1,n_1+1} = \{a = -3.27, -1.24, -0.68, -0.23, 0.8, 1.37, b = 3.12\}$

$\mathbf{t}_{2;k_2,n_2+1} = \{a = -3.12, -1.46, -0.41, 0.43, 1.48, b = 3.45\}$





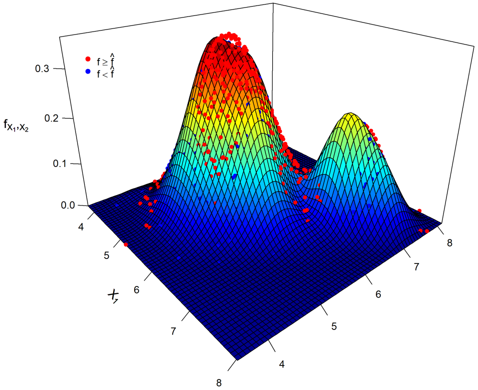
MISE boxplots - Bivariate Gaussian, $\rho = 0.5$

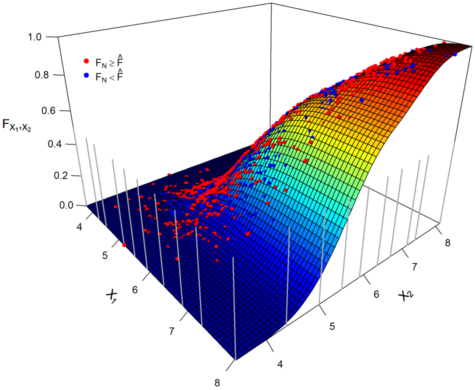


Bivariate Bimodal Mixed Gaussian

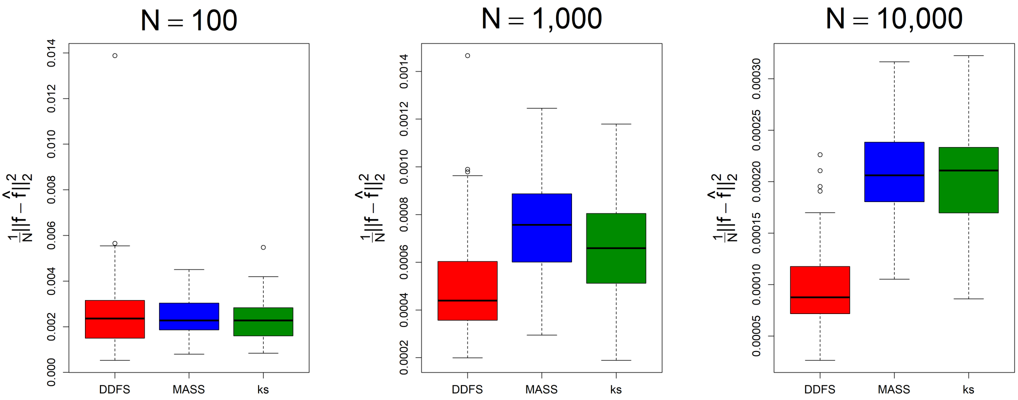
$\mathbf{t}_{1;k_1,n_1+1} = \{a = 3.74, 4.26, 4.55, 5.18, 5.54, 5.92, 6.39, 6.87, 7.33, b = 8.02\}$

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





Bivariate Bimodal Mixed Gaussian



Loss data modelling

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

➠ ML estimation often requires **large samples to converge** (data availability might be a problem) + involves a **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).

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 et al., 1995), *Conditional Coverage* (Christoffersen, 1998), *Dynamic Quantile* (Engle and and, 2004).

Danish Fire Insurance

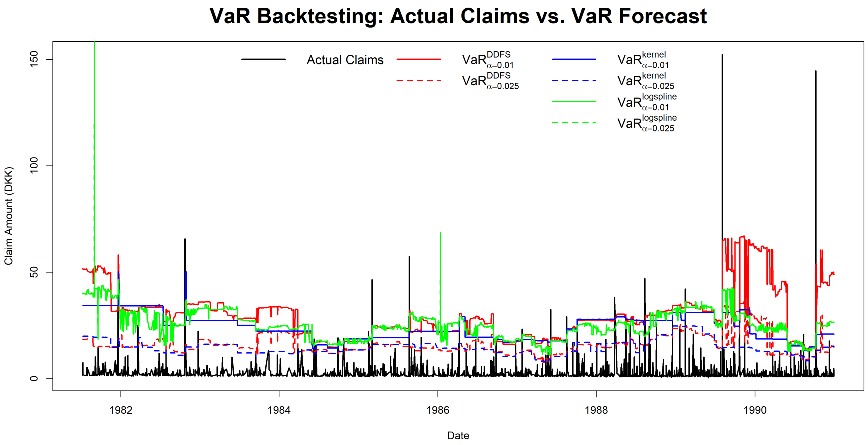


Table 1: Backtesting Statistics for VaR Models (data = danish)

	Viol. rate	TVaR MAE	UC_LRp	CC_LRp	DQ_LRp
$\alpha = 0.05$					
DDFS	0.0500	0.2929	0.9923	0.6001	0.0125
kernel	0.0513	0.5360	0.7796	0.2529	0.0013
logspline	0.0513	19.4176	0.7796	0.4360	0.8047
$\alpha = 0.025$					
DDFS	0.0250	1.8166	0.9946	0.1776	0.0280
kernel	0.0259	2.5763	0.7931	0.0623	0.0001
logspline	0.0245	28.6836	0.8867	0.4458	0.1013
$\alpha = 0.01$					
DDFS	0.0103	10.8978	0.9024	0.7819	0.0327
kernel	0.0116	5.7749	0.4586	0.5600	0.0006
logspline	0.0120	39.4870	0.3462	0.4617	0.0256

Norwegian Fire Insurance Data

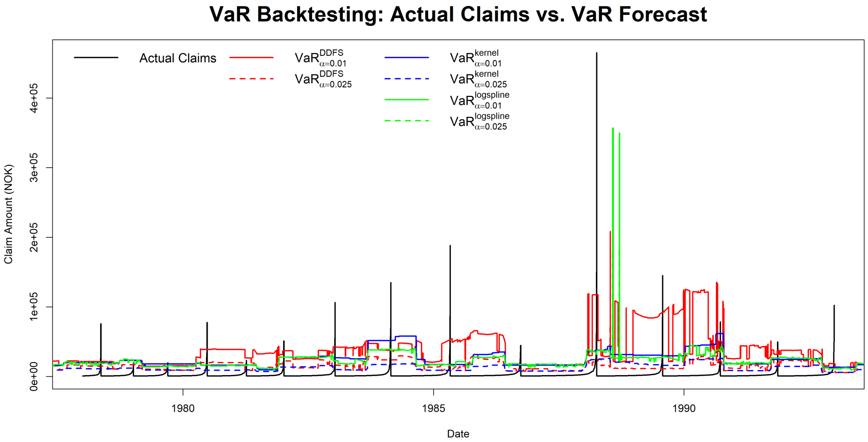


Table 2: Backtesting Statistics for VaR Models (data = norwegian)

	Viol. rate	TVaR MAE	UC_LRp	CC_LRp	DQ_LRp
$\alpha = 0.05$					
DDFS	0.0549	5728.4529	0.0459	0.0000	0.2577
kernel	0.0693	224.7367	0.0000	0.0000	0.2141
logspline	0.0687	14843.5071	0.0000	0.0000	0.2126
$\alpha = 0.025$					
DDFS	0.0251	4710.8357	0.9732	0.0000	0.0206
kernel	0.0351	72.1034	0.0000	0.0000	0.0078
logspline	0.0356	22755.3814	0.0000	0.0000	0.0074
$\alpha = 0.01$					
DDFS	0.0101	8140.1647	0.8951	0.0000	0.0001
kernel	0.0156	711.3990	0.0000	0.0000	0.0000
logspline	0.0147	39397.0583	0.0001	0.0000	0.0000

US Allocated Loss Adjustment Expenses

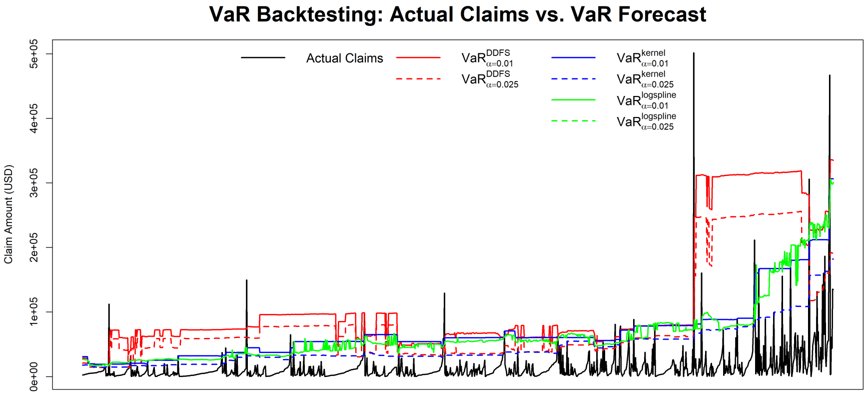


Table 3: Backtesting Statistics for VaR Models (data = lossalae)

	Viol. rate	TVaR MAE	UC_LRp	CC_LRp	DQ_LRp
$\alpha = 0.05$					
DDFS	0.0520	5924.9510	0.7471	0.0000	0.8879
kernel	0.0800	6318.5262	0.0000	0.0000	0.8910
logspline	0.0720	77978.2778	0.0008	0.0000	0.8230
$\alpha = 0.025$					
DDFS	0.0256	13695.9417	0.8923	0.0000	0.9663
kernel	0.0448	3869.2244	0.0001	0.0000	0.4172
logspline	0.0448	98390.4464	0.0001	0.0000	0.4292
$\alpha = 0.01$					
DDFS	0.0112	37127.2636	0.6757	0.0000	0.6161
kernel	0.0224	15161.8945	0.0002	0.0000	0.2674
logspline	0.0272	120016.2647	0.0000	0.0000	0.0133

- 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
- De Boor, C. (1978).A practical guide to splines. *Springer-Verlag google schola*, *2*, 4135–4195.
- 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
- Ignatov, Z., & Kaishev, V. (1989).A probabilistic interpretation of multivariate b-splines and some applications. *Serdica*, *15*.
- 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., et al. (1995). *Techniques for verifying the accuracy of risk measurement models* (Vol. 95). Division of Research; Statistics, Division of Monetary Affairs, Federal ...
- Lindstrom, M. J. (1999).Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics*, *8*(2), 333–352. https://doi.org/10.1080/10618600.1999.10474817
- Marambakuyana, W. A., & Shongwe, S. C. (2024).Composite and mixture distributions for heavy-tailed data—an application to insurance claims. *Mathematics*, *12*(2), 335.
- Redner, R. A., & Walker, H. F. (1984).Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, *26*(2), 195–239. https://doi.org/10.1137/1026034

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