Interpreting Deep Learning Models with Marginal Attribution by Conditioning on Quantiles

Anna Kozak

29.03.2021

Interpreting Deep Learning Models with Marginal Attribution by Conditioning on Quantiles

Michael Merz* Ronald Richman†‡ Andreas Tsanakas§ Mario V. Wüthrich¶

Version of March 22, 2021

- PDP
- ICE
- ALE
- LIME
- Shapley values
- **Marginal Attribution by Conditioning on Quantiles (MACQ)**

Mario Wüthrich is Professor in the Department of Mathematics at ETH Zurich, Honorary Visiting Professor at City, University of London (2011-2022), Honorary Professor at University College London (2013-2019), and Adjunct Professor at University of Bologna (2014-2016). He holds a Ph.D. in Mathematics from ETH Zurich (1999). From 2000 to 2005, he held an actuarial position at Winterthur Insurance, Switzerland. He is Actuary SAA (2004), served on the board of the Swiss Association of Actuaries (2006-2018), and is Editor-in-Chief of ASTIN Bulletin (since 2018).

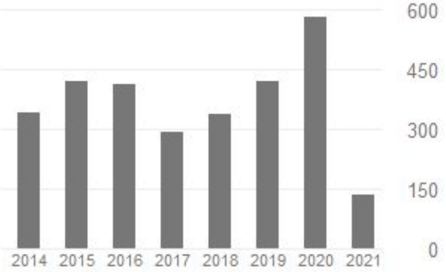
Mario Wüthrich

ETH Zurich
Keine bestätigte E-Mail-Adresse

TITEL	ZITIERT VON	JAHR
Stochastic claims reserving methods in insurance MV Wüthrich, M Merz John Wiley & Sons	462	2008
Stochastic mortality in life insurance: market reserves and mortality-linked insurance contracts M Dahl Insurance: mathematics and economics 35 (1), 113-136	452	2004
Copula convergence theorems for tail events A Juri, MV Wüthrich Insurance: Mathematics and Economics 30 (3), 405-420	178	2002
Modelling the claims development result for solvency purposes MV Wüthrich, M Merz, H Bühlmann, M De Felice, A Gisler, F Moriconi Casualty Actuarial Society E-Forum, 542-568	153	2008

Zitiert von — ALLE ANZEIGEN

	Alle	Seit 2016
Zitate	5153	2228
h-index	37	26
i10-index	104	70



$$\mu : \mathbb{R}^q \to \mathbb{R}, \qquad \boldsymbol{x} \mapsto \mu(\boldsymbol{x}), \qquad\qquad \mathbb{E}[Y|\boldsymbol{x}] = \mu(\boldsymbol{x})$$

Select a quantile level $\alpha \in (0,1)$, the α-quantile of $\mu(\boldsymbol{X})$ is given by

$$F^{-1}_{\mu(\mathbf{X})}(\alpha) = \inf\left\{y \in \mathbb{R};\ F_{\mu(\mathbf{X})}(y) \geq \alpha\right\},$$

where $F_{\mu(\mathbf{X})}(y) = P[\mu(\boldsymbol{X}) \leq y]$ describes the distribution function of $\mu(\boldsymbol{X})$.

$$\mu : \mathbb{R}^q \to \mathbb{R}, \qquad \boldsymbol{x} \mapsto \mu(\boldsymbol{x}), \qquad\qquad \mathbb{E}[Y|\boldsymbol{x}] = \mu(\boldsymbol{x})$$

Select a quantile level $\alpha \in (0,1)$, the α-quantile of $\mu(\boldsymbol{X})$ is given by

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) = \inf\left\{y \in \mathbb{R};\ F_{\mu(\boldsymbol{X})}(y) \geq \alpha\right\},$$

where $F_{\mu(\boldsymbol{X})}(y) = P[\mu(\boldsymbol{X}) \leq y]$ describes the distribution function of $\mu(\boldsymbol{X})$.

The *1st order attributions* to components $1 \leq j \leq q$ on quantile level α are defined by

$$S_j(\mu;\alpha) \ = \ \mathbb{E}_P\left[X_j \mu_j(\boldsymbol{X}) \middle| \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right].$$

These are the marginal attributions by conditioning on quantiles (MACQ).

Taylor expansion

$$\mu(\mathbf{0}) \approx \mu(\boldsymbol{x}) - (\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x}))^\top \boldsymbol{x}.$$

$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) = \mathbb{E}_P\left[\mu(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) \models F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\right] \approx \mu(\mathbf{0}) + \sum_{j=1}^{q} S_j(\mu;\alpha).$$

2nd order Taylor expansion

$$\mu(\boldsymbol{x}+\boldsymbol{\epsilon}) = \mu(\boldsymbol{x}) + (\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x}))^\top \boldsymbol{\epsilon} + \frac{1}{2}\boldsymbol{\epsilon}^\top (\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{x}))\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|^2),$$

$$F^{-1}_{\mu(\mathbf{X})}(\alpha) \;\approx\; \mu\left(\mathbf{0}\right) + \sum_{j=1}^{q} S_j(\mu;\alpha) - \frac{1}{2}\sum_{j,k=1}^{q} T_{j,k}(\mu;\alpha),$$

2nd order Taylor expansion

$$\mu(\boldsymbol{x}+\boldsymbol{\epsilon}) = \mu(\boldsymbol{x}) + (\nabla_{\boldsymbol{x}}\mu(\boldsymbol{x}))^\top \boldsymbol{\epsilon} + \frac{1}{2}\boldsymbol{\epsilon}^\top (\nabla^2_{\boldsymbol{x}}\mu(\boldsymbol{x}))\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|^2),$$

$$F^{-1}_{\mu(\mathbf{X})}(\alpha) \;\approx\; \mu\left(\mathbf{0}\right) + \sum_{j=1}^{q} S_j(\mu;\alpha) - \frac{1}{2}\sum_{j,k=1}^{q} T_{j,k}(\mu;\alpha),$$

$$1 \le j,k \le q, \qquad T_{j,k}(\mu;\alpha) = \mathbb{E}_P\left[X_j X_k \mu_{j,k}(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\mathbf{X})}(\alpha)\right]$$

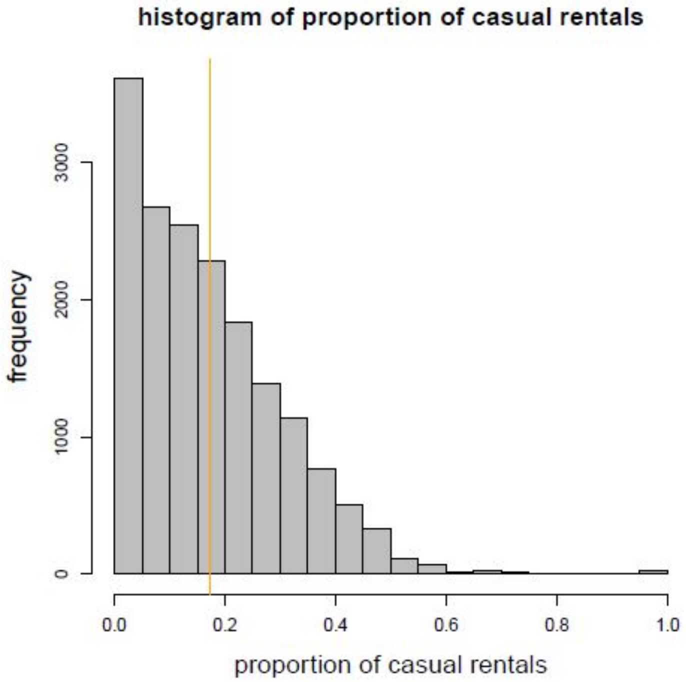
$$F^{-1}_{\mu(\mathbf{X})}(\alpha) \;\approx\; \mu\left(\mathbf{0}\right) + \sum_{j=1}^{q}\left(S_j(\mu;\alpha) - \frac{1}{2}T_{j,j}(\mu;\alpha)\right) - \sum_{1\le j<k\le q} T_{j,k}(\mu;\alpha)$$

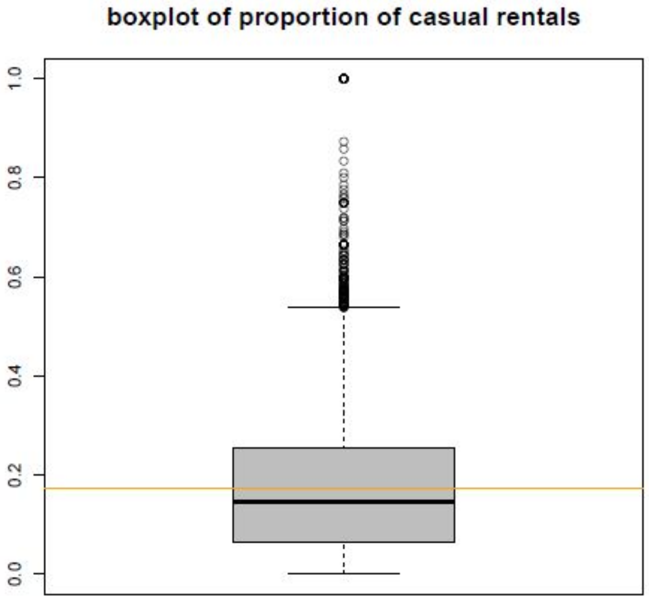
Example - bike rental data

Listing 1: Excerpt of bike rental data.

```
'data.frame':   17379 obs. of  13 variables:
 $ date       : Date, format: "2011-01-01" "2011-01-01" "2011-01-01" ...
 $ year       : num  2011 2011 2011 2011 2011 ...
 $ month      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ hour       : int  0 1 2 3 4 5 6 7 8 9 ...
 $ weekday    : int  6 6 6 6 6 6 6 6 6 6 ...
 $ holiday    : Factor w/ 2 levels "holiday","no-holiday": 2 2 2 2 2 2 2 2 2 2 ...
 $ workingday: Factor w/ 2 levels "no-working","workingday": 1 1 1 1 1 1 1 1 1 1 ...
 $ weather    : num  1 1 1 1 1 2 1 1 1 1 ...
 $ temp       : num  0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...
 $ temp_feel : num  0.288 0.273 0.273 0.288 0.288 ...
 $ humidity  : num  0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...
 $ windspeed : num  0 0 0 0 0 0.0896 0 0 0 0 ...
 $ casual     : int  3 8 5 3 0 0 2 1 1 8 ...
 $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
 $ count      : int  16 40 32 13 1 1 2 3 8 14 ...
```

Example - bike rental data

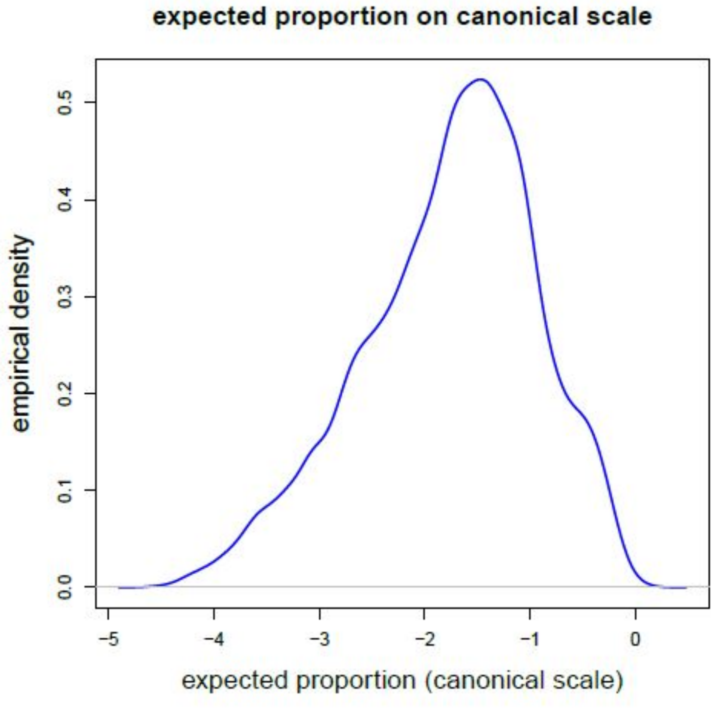


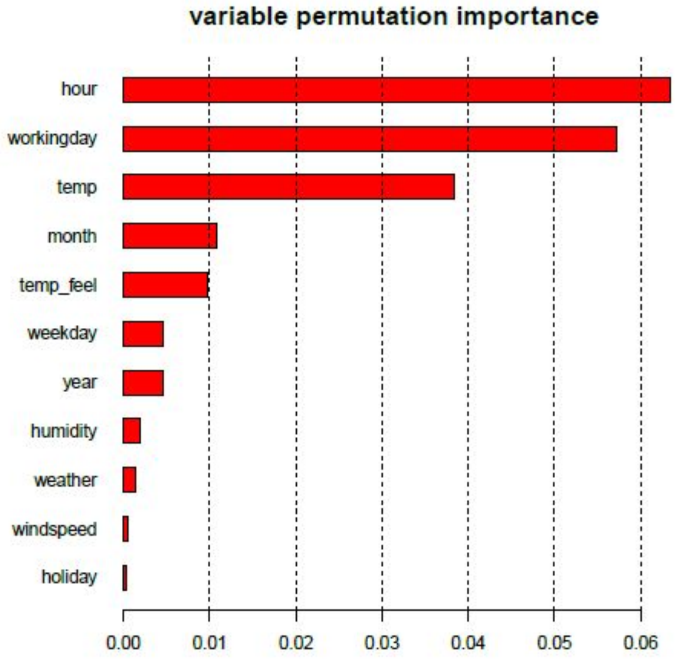


Model

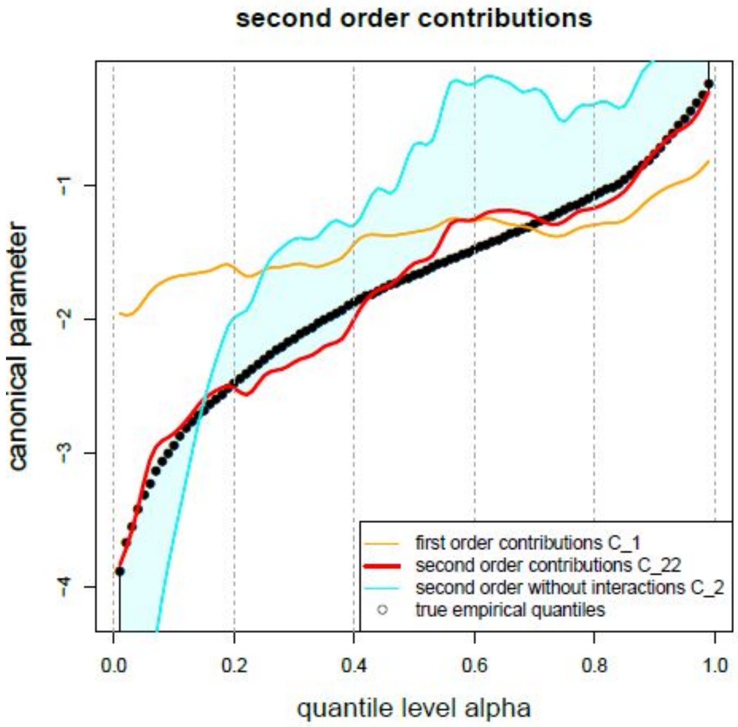
- fully-connected feed-forward neural network
- three hidden layers (20, 15, 10)
- sigmoid output activation
- hyperbolic tangent as activation function in 3 hidden layers

Example - bike rental data

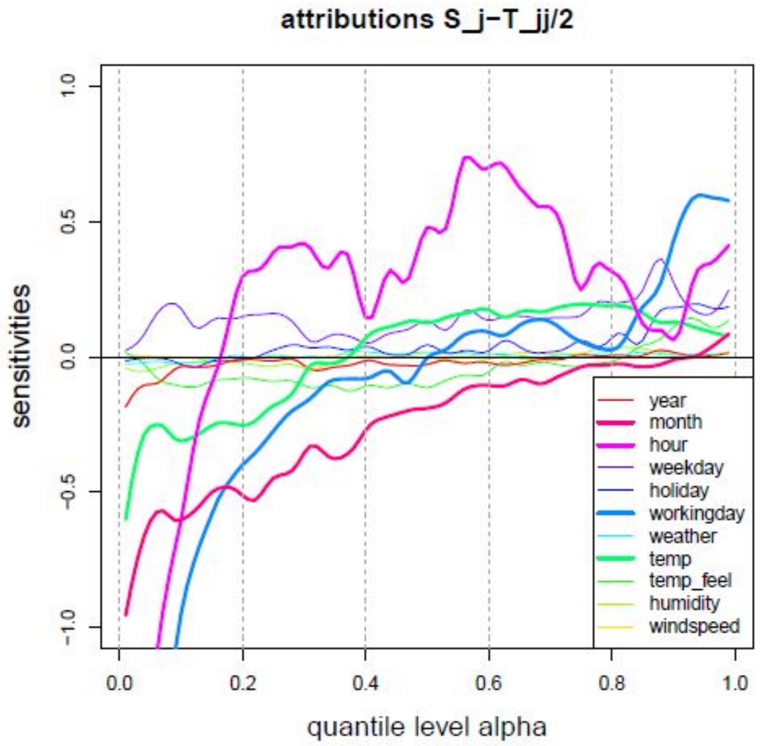


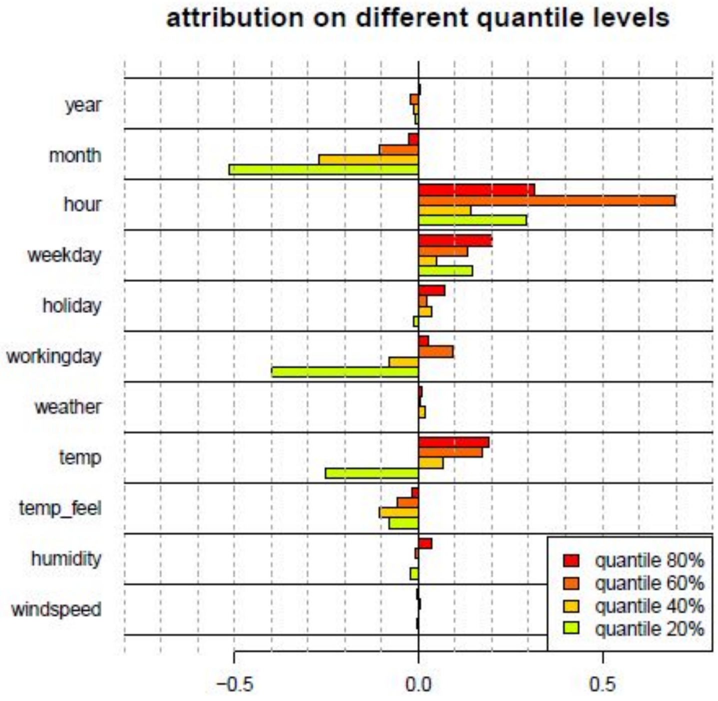


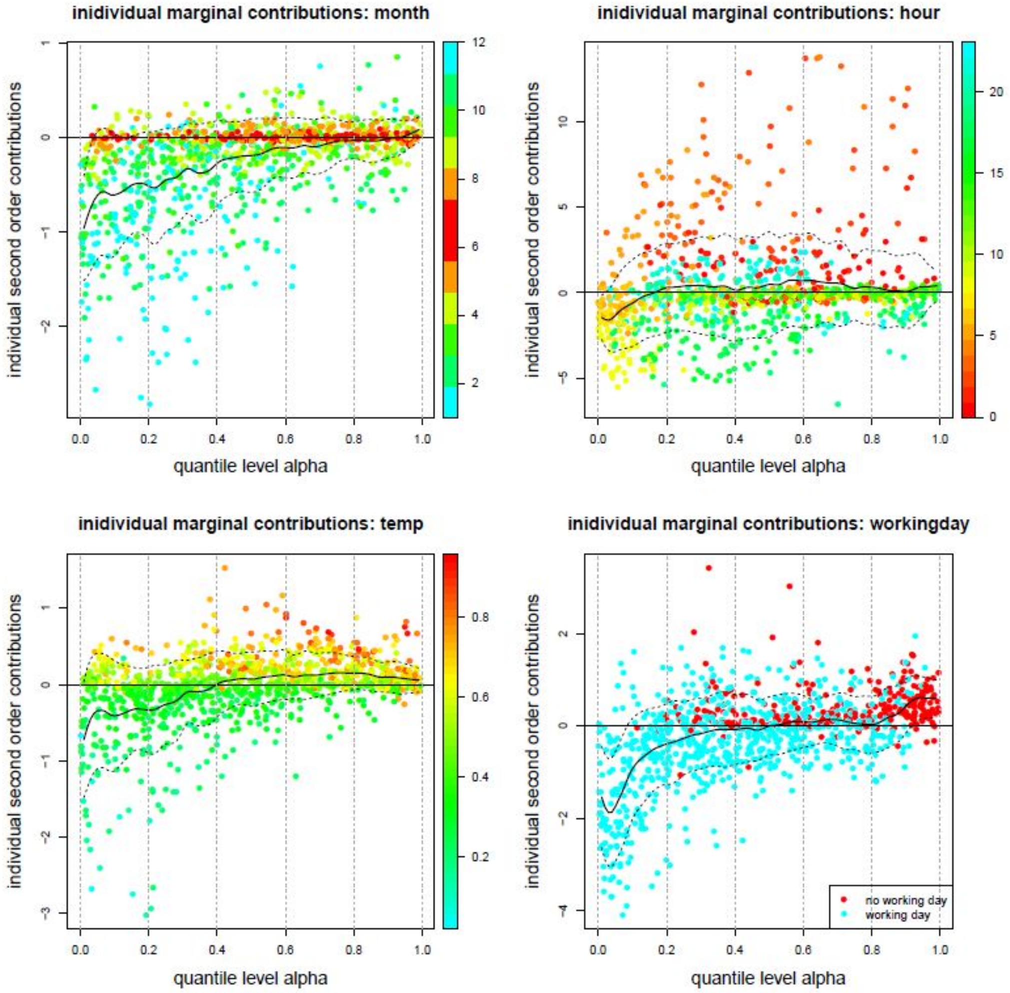
Example - bike rental data



Example - bike rental data






inidividual marginal contributions: month
inidividual marginal contributions: hour
inidividual marginal contributions: temp
inidividual marginal contributions: workingday
individual second order contributions
quantile level alpha
no working day
working day