Interpreting Deep Learning Models with Marginal Attribution by Conditioning on Quantiles (MACQ)

Mario V. Wüthrich
RiskLab, ETH Zurich

Joint work with:
Michael Merz, Ronald Richman & Andreas Tsanakas

May 20, 2021

ASTIN 2021 Online Colloquium

Overview of contents

- Refresher on neural network forecasting
- Explainability of network forecasts

This presentation is based on:

Interpreting Deep Learning Models with Marginal Attribution by Conditioning on Quantiles (2021).
`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3809674`

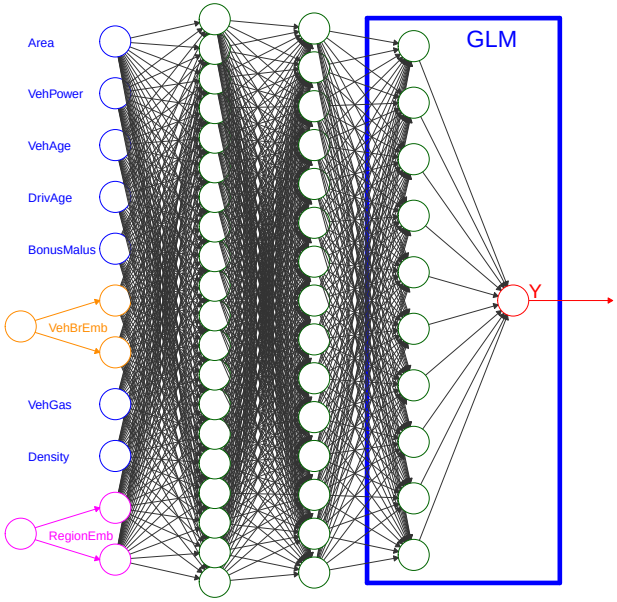
Statistical Foundations of Actuarial Learning and its Applications (2021).
`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3822407`

- **Refresher on neural network forecasting**

Refresher on neural network forecasting

- Model response Y with covariates $\boldsymbol{x}$ using a network regression of depth $d \in \mathbb{N}$

$$\boldsymbol{x} \quad \mapsto \quad \mathbb{E}[Y|\boldsymbol{x}] \;=\; \mu(\boldsymbol{x}) \;=\; g\left\langle \theta^{(d+1)}, \left(\boldsymbol{z}^{(d)} \circ \cdots \circ \boldsymbol{z}^{(1)}\right)(\boldsymbol{x}) \right\rangle.$$



- $\left(\boldsymbol{z}^{(d)} \circ \cdots \circ \boldsymbol{z}^{(1)}\right)$ is a composition of d layers $\boldsymbol{z}^{(k)}$ $\Rightarrow$ covariate $\boldsymbol{x}$ pre-processing;
- $g\langle \theta^{(d+1)}, \cdot \rangle$ is the GLM part with inverse link function g and parameter $\theta^{(d+1)}$.

Fitting a neural network

- The above network has network parameter $\boldsymbol{\theta}$ of dimension $r = 792$.
- Log-likelihood for given observations $\boldsymbol{Y}$: $\boldsymbol{\theta} \mapsto \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(Y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$.
- Maximum likelihood estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$ will over-fit to the data $\boldsymbol{Y}$.
- Use early stopping in gradient descent fitting to get "a" good estimate $\widehat{\boldsymbol{\theta}}$.
- "Good" network predictors are not unique, i.e., they have an element of subjectivity.

Short summary

+ *Typically*, the network outperforms the GLM approach in terms of out-of-sample prediction accuracy (covariate $\boldsymbol{x}$ pre-processing).

+ Use embedding layers in networks for categorical variables.

– Resulting (network) prices are not unique, but depend on seeds (subjective).

– The network does not build on improving the GLM.

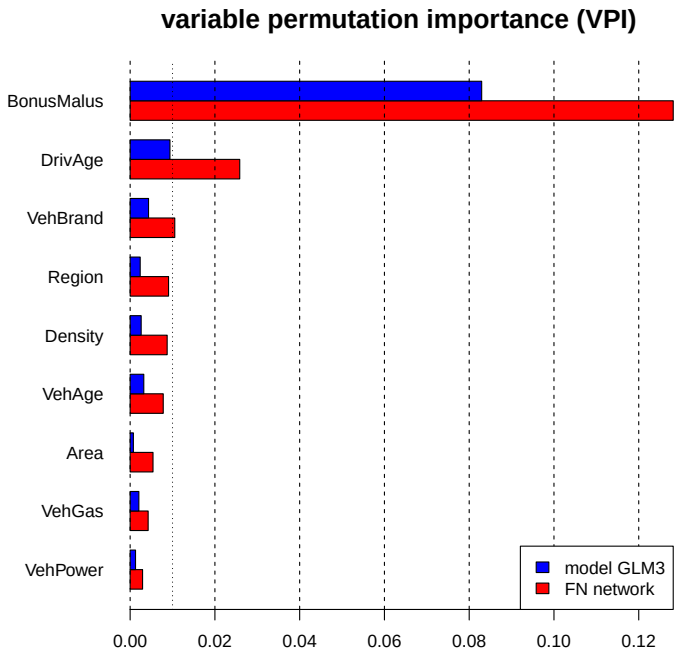
– The network fails to have the balance property $\sum_{i=1}^n Y_i \overset{??}{=} \sum_{i=1}^n \widehat{\mu}(\boldsymbol{x}_i)$.

– Network predictors are not transparent / explainable.

- **Explainability of network forecasts**

Variable permutation importance (VPI)

- Most popular (and most simple) measure is variable permutation importance.
- Permute one component of $\boldsymbol{x}$ in $\mu(\boldsymbol{x})$ at a time across the whole portfolio and analyze the increase/change in out-of-sample loss.



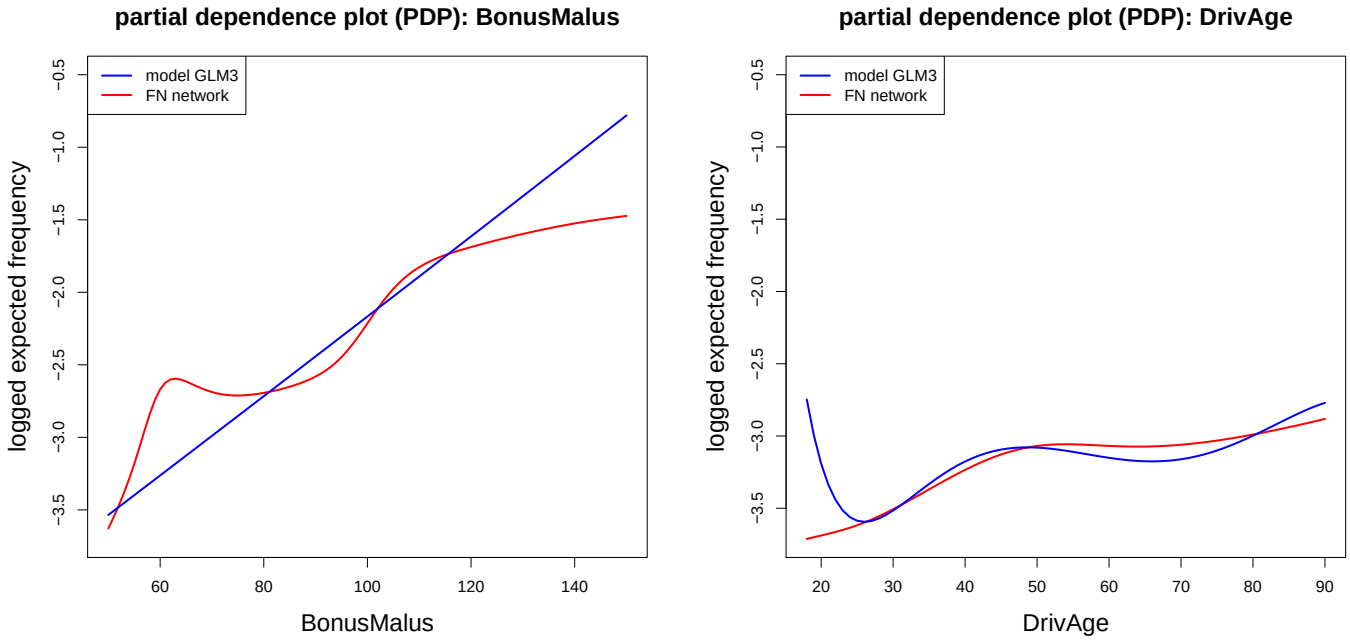
- Does not appropriately respect dependence between covariate components in $\boldsymbol{x}$.

Partial dependence plots (PDP)

- Partial dependence plots: marginal profile for fixed covariate component x_j

$$x_j \;\mapsto\; \bar{\mu}^{(j)}(x_j) = \int \mu(x_j \cup \boldsymbol{x}_{\setminus j})\; dp(\boldsymbol{x}_{\setminus j}),$$

$p(\cdot)$ describing the portfolio distribution.



Accumulated local effects (ALE)

- Consider partial derivative (sensitivity) w.r.t. x_j

$$\mu_j(\boldsymbol{x}) = \frac{\partial \mu(\boldsymbol{x})}{\partial x_j}.$$

- Average local effect of component j is received by

$$x_j \;\mapsto\; \Delta_j(x_j; \mu) = \int \mu_j(x_j, \boldsymbol{x}_{\setminus j})\; dp(\boldsymbol{x}_{\setminus j} | x_j).$$

- Accumulated local effects

$$x_j \;\mapsto\; \tilde{\mu}^{(j)}(x_j) = \int_{-\infty}^{x_j} \Delta_j(z_j; \mu) dz_j.$$

Sensitivities of distortion risk measures

- Tsanakas–Millossovich (2015) use directional derivatives to study sensitivities of (distortion) risk measures w.r.t. risk factors $\boldsymbol{X}$.

- Example of distortion risk measure: Value-at-Risk.

- Consider risk measure $\varrho(\mu(\boldsymbol{X}))$ of portfolio mean $\mu(\boldsymbol{X})$ w.r.t. risk factors $\boldsymbol{X}$.

- For the Value-at-Risk distortion risk measure ϱ, and under certain assumptions,

$$\begin{aligned} S_j &\overset{\text{def.}}{=} \lim_{\epsilon \to 0} \frac{\varrho(\mu(\boldsymbol{X} + \epsilon X_j \boldsymbol{e}_j)) - \varrho(\mu(\boldsymbol{X}))}{\epsilon} \\ &= \mathbb{E}\left[X_j \mu_j(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \right]. \end{aligned}$$

- Thus, sensitivities on given quantile levels $\{\mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\}$ are studied.

Marginal attribution by conditioning on quantiles

- Study how 1st and 2nd order terms contribute to $\mu(\boldsymbol{X})$ on given quantile levels $\{\mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\}$; $\boldsymbol{X}$ are covariates here.

- A Taylor approximation provides (initialize in $\boldsymbol{x} = \mathbf{0}$)

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

with first and second order marginal attributions

$$\begin{aligned} S_j(\alpha) &= \mathbb{E}\left[X_j \mu_j(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \right], \\ T_{j,k}(\alpha) &= \mathbb{E}\left[X_j X_k \mu_{j,k}(\boldsymbol{X}) \,\middle|\, \mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \right]. \end{aligned}$$

Marginal attribution by conditioning on quantiles

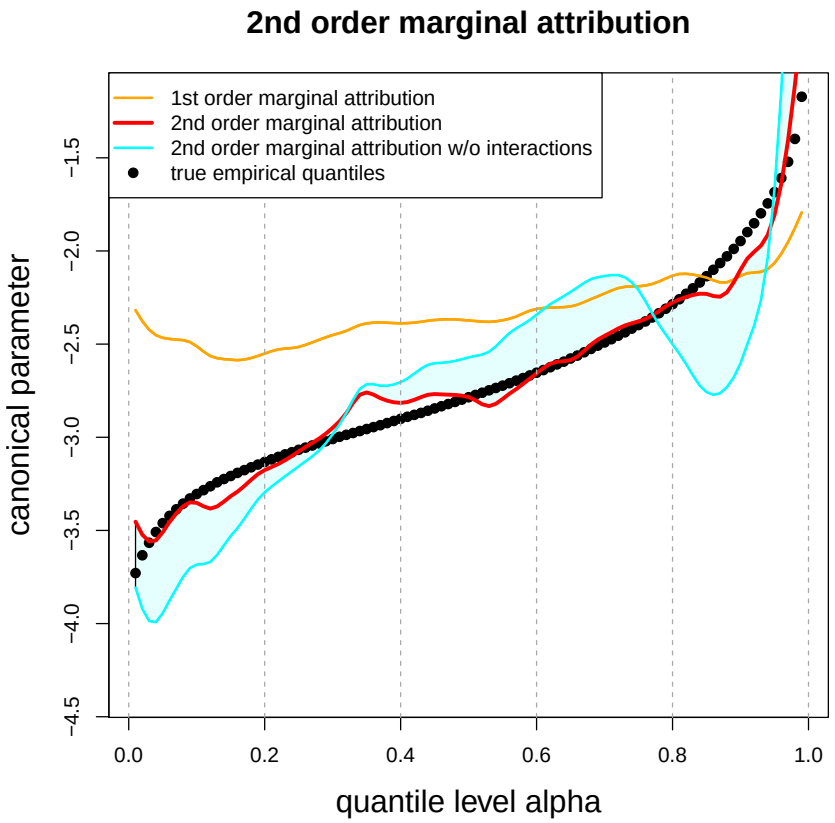
- A Taylor approximation provides

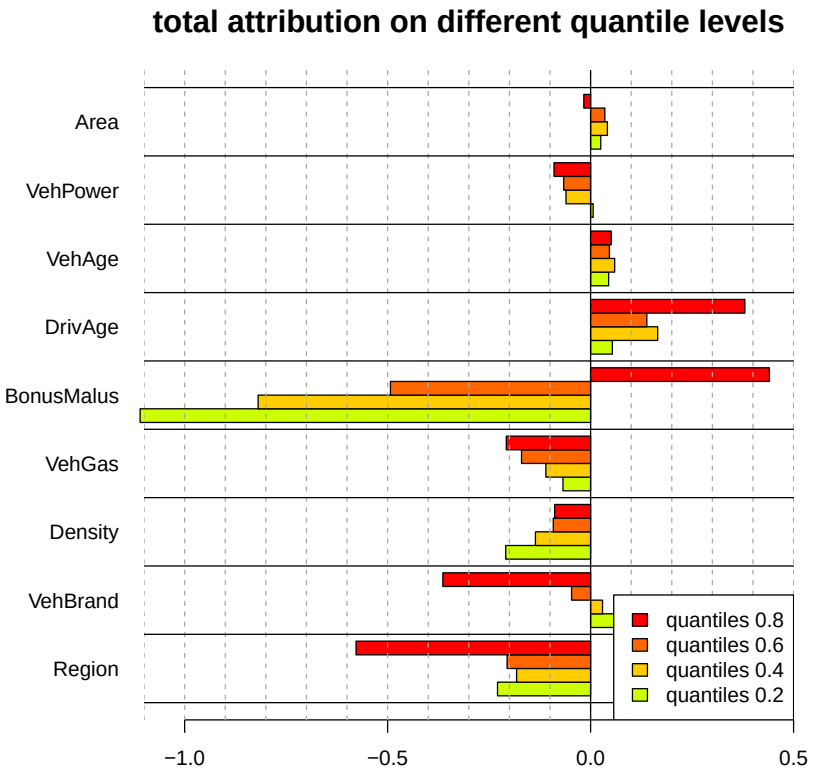
$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \;\approx\; \mu(\mathbf{0}) + \sum_{j=1}^{q} \left(S_j(\alpha) - \frac{1}{2} T_{j,j}(\alpha) \right) - \sum_{1 \le j < k \le q} T_{j,k}(\alpha).$$

- The green terms describe the contribution of components $1 \le j \le q$ of $\boldsymbol{X}$.
- The magenta terms describe the interaction terms.
- Respects dependence structure in covariates $\boldsymbol{X} = \boldsymbol{x}$.

This allows us to study variable importance on different quantile levels $\{\mu(\boldsymbol{X}) = F^{-1}_{\mu(\boldsymbol{X})}(\alpha)\}$ for $\alpha \in (0, 1)$.

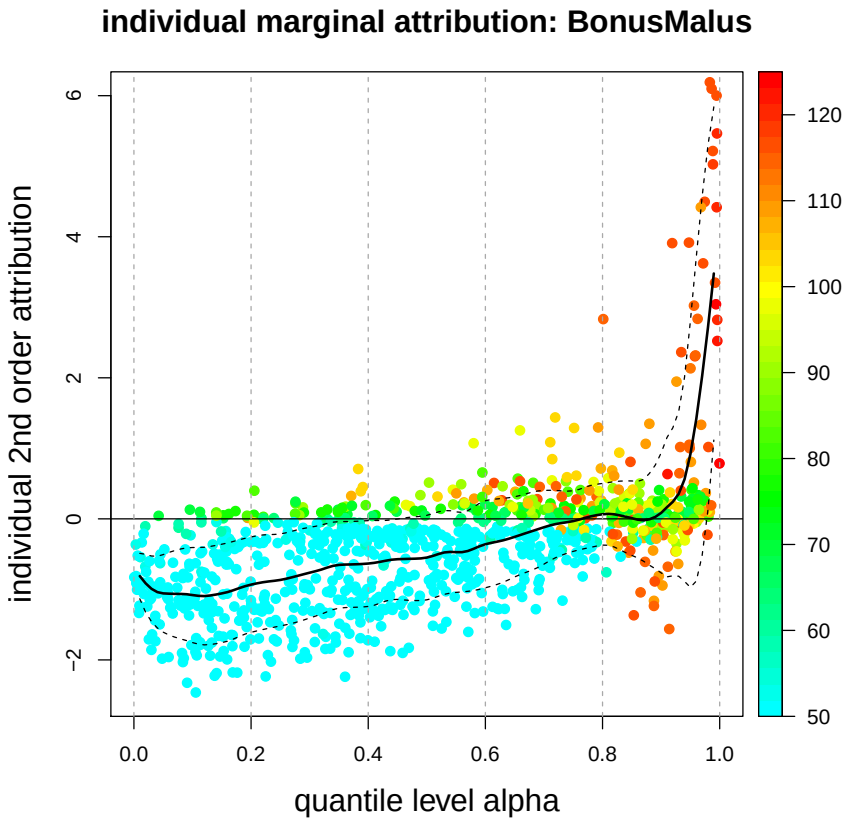
Results MACQ method (1/2)

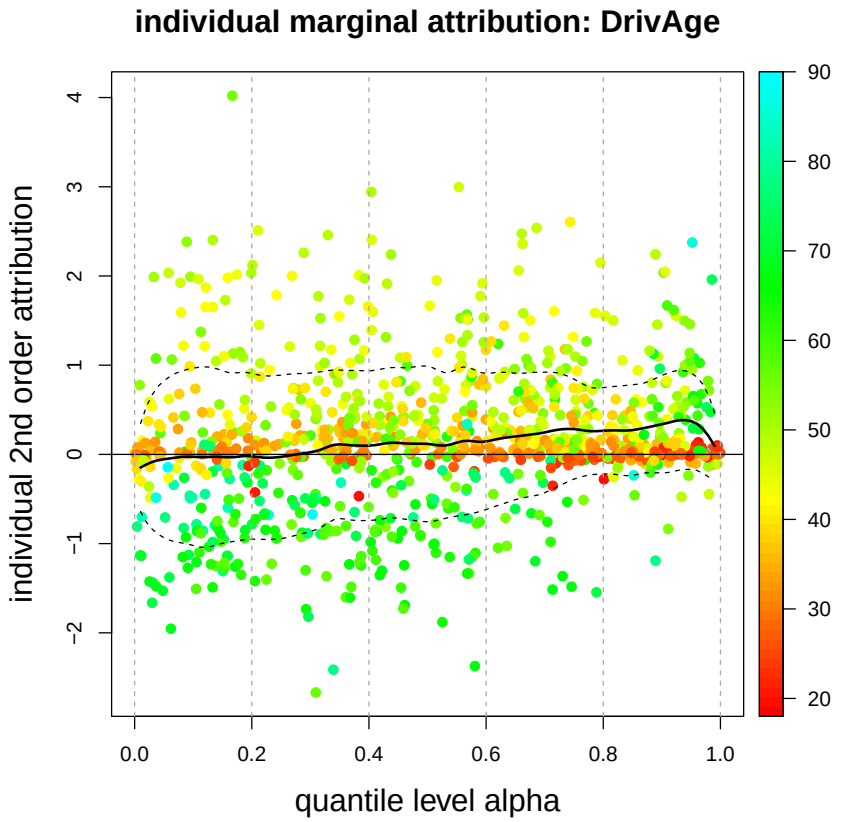




$$F^{-1}_{\mu(\boldsymbol{X})}(\alpha) \;\approx\; \mu(\mathbf{0}) + \sum_{j=1}^{q}\left(S_j(\alpha) - \frac{1}{2}T_{j,j}(\alpha)\right) - \sum_{1\le j<k\le q} T_{j,k}(\alpha).$$

Results MACQ method (2/2)





$S_j(\alpha) - \frac{1}{2}T_{j,j}(\alpha)$ for covariate components $j = \text{BonusMalus}, \text{DrivAge}$.

Take Aways

- A neural network is an extension of a GLM.
- Neural networks do covariate pre-processing themselves.
- 'Sufficiently good' network regression models are not unique.
- There is a vastly growing literature on explaining networks.
- Not all of these methods are suitable for explanation because they may not reflect the portfolio structure.
- MACQ studies regression functions on given quantile levels.

References

[1] Ancona, M., Ceolini, E., Öztireli, C., Gross, M. (2019). Gradient-based attribution methods. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller K.-R. (Eds.). Springer, Lecture Notes in Artificial Intelligence 11700, 168-191.

[2] Apley, D.W., Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society, Series B* **82/4**, 1059-1086.

[3] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29/5**, 1189-1232.

[4] Lorentzen, C., Mayer, M. (2020). Peeking into the black box: an actuarial case study for interpretable machine learning. *SSRN Manuscript* ID 3595944. Version May 7, 2020.

[5] Merz, M., Richman, R., Tsanakas, A., Wüthrich, M.V. (2021). Interpreting deep learning models with marginal attribution by conditioning on quantiles. *SSRN Manuscript* ID 3809674.

[6] Tsanakas, A., Millossovich, P. (2015). Sensitivity analysis using risk measures. *Risk Analysis* **36/1**, 30-48.

[7] Wüthrich, M.V., Merz, M. (2021). *Statistical Foundations of Actuarial Learning and its Applications*. SSRN Manuscript ID 3822407.

[8] Zhao, Q., Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics* **39/1**, 272-281.