Bachelor's Thesis

Neural Structured Additive Distributional Regression

Author: Maximilian Schneider
Supervisor: Dr. David Rügamer

Department Of Statistics
Ludwig-Maximilians-Universität

July 12, 2022

Introduction: Exemplary underlying *data generating process*

Introduction: Model estimation

- Two types of model parameters[1]:
  - Model coefficients: $\boldsymbol{\beta}$
  - Smoothing parameters pertaining to splines, regulating flexibility: $\boldsymbol{\lambda}$
- Comparison of three methods
  - Two variants of the generalized Fellner-Schall method (GFS) (Wood and Fasiolo 2017)
  - Optimization using an artificial neural network (NN) and *Adam* (Kingma and Ba 2014)
- GFS outperforms the other method except for settings where the number of coefficients is close to the number of observations.

[1] not to be confused with *distributional* parameters

Contents

Notation

- Bold lower-case symbols (e. g. $\boldsymbol{\phi}_i$): vectors
- Bold upper-case symbols (e. g. $\boldsymbol{X}$): matrices
- Regular symbols (e. g. ϕ_{di}):
  - Scalars
  - or (accompanied by $(\cdot)$): functions
- In some contexts:
  - Regular upper-case symbols (e. g. Y_i): random variables
  - Corresponding y_i: realizations
- "Hats" (e. g. $\hat{\boldsymbol{\beta}}$): estimates

Model framework

A generalized additive model for location, scale and shape (GAMLSS) [2] is defined via

$$
\begin{aligned}
Y_i &\sim D(\boldsymbol{\phi}_i),\ i = 1, \ldots, n,\ d = 1, \ldots, \dim(\boldsymbol{\phi}_i) \\
g_d(\phi_{di}) &= \eta_{di} = \beta_{d0} + \sum_{k=1}^{K_d} s_{dk}(x_{dki}).
\end{aligned} \tag{1}
$$

$$
(\dim(\boldsymbol{\phi}_i) = \dim(\boldsymbol{\phi}_j)\, \forall i, j)
$$

Typically D is parameterized in a way that $\boldsymbol{\phi}_i \equiv (\mu_i, \sigma_i, \nu_i)$; μ_i being location, σ_i scale and ν_i shape.

[2] Building on the notation of Aeberhard et al. (2021)

Smoothing parameters

Prevention of overfitting via penalization of the functions wigglyness, measured as its second derivative, rewritten as a quadratic penalty

$$\lambda_k \int_{x_1}^{x_{J_k}} \hat{s}_k''(x)^2 \, dx = \lambda_k \hat{\boldsymbol{\beta}}_k^\top \boldsymbol{\mathcal{S}}_k \hat{\boldsymbol{\beta}}_k \tag{2}$$

Including more terms/splines the penalty can be expressed as

$$\hat{\boldsymbol{\beta}}^\top \mathbf{S}_{\boldsymbol{\lambda}} \hat{\boldsymbol{\beta}}. \tag{3}$$

Containing e. g.

$$\hat{\boldsymbol{\beta}}^\top \mathbf{S}_{\boldsymbol{\lambda}} \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}^\top \begin{pmatrix} 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda_1 \boldsymbol{\mathcal{S}}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \lambda_2 \boldsymbol{\mathcal{S}}_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}. \tag{4}$$

Objective function

Estimation approaches compared in this thesis are based on penalized maximum likelihood, with corresponding objective function

$$\begin{aligned} \ell_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_{\boldsymbol{\lambda}}(\boldsymbol{\beta})_i &= \sum_{i=1}^{n} \ell(\boldsymbol{\beta})_i - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta} \\ &= \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}. \end{aligned} \tag{5}$$

Newton's method

Optimization of (5) for $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\lambda}}$ via Newton's method

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \boldsymbol{\mathcal{H}}_{\hat{\boldsymbol{\lambda}}}^{-1} \boldsymbol{\mathcal{G}}_{\hat{\boldsymbol{\lambda}}}, \tag{6}$$

where $\boldsymbol{\mathcal{G}}_{\boldsymbol{\lambda}} = \partial \ell_{\boldsymbol{\lambda}} / \partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ is the gradient vector[3] and
$\boldsymbol{\mathcal{H}}_{\boldsymbol{\lambda}} = -(\partial^2 \ell_{\boldsymbol{\lambda}} / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}})$ is the negative Hessian matrix of $\ell_{\boldsymbol{\lambda}}(\cdot)$.

[3] Following the notation of Wood, Pya, and Säfken (2016)

Trust region algorithm

Optimization of (5) for $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\lambda}}$ via the trust region algorithm[4]

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \underset{\boldsymbol{e}:\|\boldsymbol{e}\|\leq\Delta}{\arg\max} \left\{ \ell_{\hat{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\beta}}) + \boldsymbol{e}^\top \boldsymbol{\mathcal{G}}_{\hat{\boldsymbol{\lambda}}} - \frac{1}{2} \boldsymbol{e}^\top \boldsymbol{\mathcal{H}}_{\hat{\boldsymbol{\lambda}}} \boldsymbol{e} \right\}, \tag{7}$$

where Δ denotes the current radius of a sphere with center $\hat{\boldsymbol{\beta}}$.

[4] Following Aeberhard et al. (2021)

Generalized Fellner-Schall method

Based on Laplace approximate marginal likelihood (LAML), which is optimized to obtain an estimate for $\boldsymbol{\lambda}$, one can derive the update formula

$$\lambda_k^* = \frac{\operatorname{tr}(\mathbf{S}_{\hat{\boldsymbol{\lambda}}}^- \frac{\partial \mathbf{S}_{\boldsymbol{\lambda}}}{\partial \lambda_k}\big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}) - \operatorname{tr}(\boldsymbol{\mathcal{H}}_{\hat{\boldsymbol{\lambda}}}^{-1} \frac{\partial \mathbf{S}_{\boldsymbol{\lambda}}}{\partial \lambda_k}\big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}})}{\hat{\boldsymbol{\beta}}^\top (\frac{\partial \mathbf{S}_{\boldsymbol{\lambda}}}{\partial \lambda_k}\big|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}) \hat{\boldsymbol{\beta}}} \hat{\lambda}_k, \tag{8}$$

which maintains the statistical consistency of reduced rank splines like the ones employed in this thesis given that $K = O(n^\alpha)$, where $\alpha < 1/3$ (Wood, Pya, and Säfken 2016).

Estimation using an artificial neural network

At step m all model parameters, written as one vector $\boldsymbol{\theta}$, are updated by[5]

$$\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}} - \alpha \, \boldsymbol{v}_m^{'*} / (\sqrt{\boldsymbol{w}_m^{'*}} + 10^{-7}), \tag{9}$$

where $\boldsymbol{\mathcal{G}} = \partial - \ell_{\boldsymbol{\lambda}} / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, $\boldsymbol{v}_m^* = 0.9 \, \hat{\boldsymbol{v}}_m - 0.1 \, \boldsymbol{\mathcal{G}}$, with a bias corrected version $\boldsymbol{v}_m^{'*} = \boldsymbol{v}_m^* / (1 - 0.9^m)$, $\boldsymbol{w}_m^* = 0.999 \, \hat{\boldsymbol{w}}_m - 0.001 \, \boldsymbol{\mathcal{G}}^2$, with $\boldsymbol{w}_m^{'*} = \boldsymbol{w}_m^* / (1 - 0.999^m)$ and α is the step size[6].

- $\ell_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is only evaluated at random subsets of $\boldsymbol{y}$.
- Implicit regularization
- Expectation $\hat{\boldsymbol{\lambda}} = \mathbf{0}$
- Early stopping using 10% of data $\Rightarrow$ Explicit regularization through $\boldsymbol{\lambda}$ if stopping occurs early enough

[5] Note that Adam minimizes functions, e. g. $-\ell_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$.

[6] Note that some tunable hyperparametes already have been replaced by the values used for model fitting in this thesis.

Data simulation

Effects on linear predictors

Exemplary data generating process

Gamma distribution, p = 5, $p_1 = 2$, $p_2 = 4$, n = 100

$$
\begin{aligned}
Y_i &\sim \mathsf{GA}(\exp(\eta_{1i}), \exp(\eta_{2i})),\ i = 1, \ldots, 100 \\
\eta_{1i} &= f_3(x_{2i}) + f_1(x_{5i}) \\
\eta_{2i} &= f_4(x_{5i}) + f_1(x_{1i}) + f_2(x_{3i}) + f_2(x_{2i}) + c,
\end{aligned} \tag{10}
$$

The packages in turn estimate this specification:

$$
\begin{aligned}
Y_i &\sim \mathsf{GA}(g_1^{-1}(\eta_{1i}), g_2^{-1}(\eta_{2i})),\ i = 1, \ldots, 100 \\
g_1(\mu_i) &= \eta_{1i} = \beta_{10} + s_{11}(x_{2i}) + s_{12}(x_{5i}) \\
g_2(\sigma_i) &= \eta_{2i} = \beta_{20} + s_{21}(x_{5i}) + s_{22}(x_{1i}) + s_{23}(x_{3i}) + s_{24}(x_{2i}),
\end{aligned} \tag{11}
$$

where $J_{dk} = 10\ \forall d, k$ and $g_d(\cdot)$ are the employed links[7].

[7] Note that linear effects are specified as smooth terms too.

Signal-to-noise ratio

The signal-to-noise ratio (SNR) is defined for this thesis as

$$\text{SNR} = \frac{\mathbb{V}(\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}))}{\mathbb{V}(\boldsymbol{y} - \mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}))}, \tag{12}$$

where $\boldsymbol{Y}$ is a vector of i.i.d. response random variables and $\boldsymbol{X}$ denotes the model matrix.
Inverse transform sampling (see e. g. Wikipedia 2021):

1. Draw observations u_i of $U_i \overset{i.i.d.}{\sim} \text{Unif}[0, 1]$
2. Transform $\boldsymbol{u}$ into $\boldsymbol{y}$ using respective quantile function $q(\cdot)$

$$\boldsymbol{y} = q(\boldsymbol{u}|g_1^{-1}(\boldsymbol{\eta}_1), g_2^{-1}(\boldsymbol{\eta}_2 + c)), \tag{13}$$

where the SNR is controlled through c.

Integrated mean squared error

Integrated mean squared error

Package	Minimum	1st Quartil	Median	Mean	3rd Quartil	Maximum
DR	0.001	0.007	0.024	0.062	0.099	0.725
GJRM	0.000	0.003	0.009	0.579	0.037	47.436
mgcv	0.000	0.003	0.007	8.997	0.030	5925.334
DR	0.003	0.073	0.123	0.135	0.182	0.752
GJRM	0.000	0.003	0.009	0.720	0.041	939.321
mgcv	0.000	0.003	0.009	0.186	0.033	90.577
DR	0.001	0.010	0.048	2.314	0.126	1569.680
GJRM	0.000	0.003	0.010	0.434	0.045	165.361
mgcv	0.000	0.000	0.010	52.100	0.070	82731.930

Table: Summary mean integrated mean squared error (IMSE). (i) Normal, (ii) Gamma, (iii) Gumbel

Integrated mean squared error

Integrated mean squared error

Log-likelihood

Log-likelihood

Package	Minimum	1st Quartil	Median	Mean	3rd Quartil	Maximum
DR	0.052	0.170	0.298	0.417	0.509	6.299
GJRM	0.000	0.000	0.000	15211.000	0.000	$2.333 \cdot 10^7$
mgcv	0.013	0.067	0.111	23.691	0.210	19587.182
DR	0.079	0.392	0.568	0.714	0.856	10.217
GJRM	0.009	0.071	0.127	5.762	0.251	2475.874
mgcv	0.009	0.072	0.122	0.193	0.220	8.151
DR	0.164	0.393	0.522	∞	0.797	∞
GJRM	0.000	0.000	0.000	$7.721 \cdot 10^5$	0.000	$1.444 \cdot 10^9$
mgcv	0.018	0.099	0.163	∞	0.336	∞

Table: Summary log-lik. (i) Normal, (ii) Gamma, (iii) Gumbel

Experiences with `deepregression`

- Good performance using the Adam optimizer and a step size of $\alpha = 0.01$
- Early stopping often at around 100 evolutions
- Some $\hat{\boldsymbol{\lambda}} \neq \mathbf{0}$, maybe due to early stopping
- Some degree of overfitting observable using plots like the figure for the IMSE
- Maybe test optimizing the improper joint density on which GFS is based, instead of $\ell_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$
- Batch-sizes: $\mathrm{round}(\sqrt{10\,\mathsf{n}})$

Outlook

- Training the NN with a constant batch size
- Change generation of the structure of the relationship between the response and its covariates
- More grid search-like approach for settings of special interest, e. g. ones with extreme errors
- Extension of possible smooth effects to include more difficult ones (e. g. $f_5(\cdot)$)
- Test other spline base types, e. g. P-splines are readily available
- Specify the packages to estimate effects for all available covariates and check if variables independent of $\boldsymbol{Y}$ get estimates $\hat{s}(\cdot) \equiv 0$
- Planned implementation of GFS in `deepregression` $\Rightarrow$ code suitable for beta-testing
- Estimation in an NN was revealed to be superior in cases where the assumptions of GFS are too restrictive.

Bibliography I

- Aeberhard, W. H. et al. (2021). "Robust fitting for generalized additive models for location, scale and shape". In: *Statistics and Computing* 31.1, pp. 1–16.
- Kingma, D. P. and J. Ba (2014). "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations*. URL: https://arxiv.org/abs/1412.6980.
- Wikipedia (July 10, 2021). *Inverse transform sampling — Wikipedia, The Free Encyclopedia*. URL: https://en.wikipedia.org/w/index.php?title=Inverse_transform_sampling&oldid=1030177737.
- Wood, S. N. and M. Fasiolo (2017). "A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models". In: *Biometrics* 73.4, pp. 1071–1081. DOI: 10.1111/biom.12666.

Bibliography II

Wood, S. N., N. Pya, and B. Säfken (2016). "Smoothing parameter and model selection for general smooth models". In: *Journal of the American Statistical Association* 111.516, pp. 1548–1563. DOI: 10.1080/01621459.2016.1180986.