Statistics and Probability Letters 97 (2015) 155–164

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

L_2 differentiability of generalized linear models

Daria Pupashenko [a,b], Peter Ruckdeschel [c,*], Matthias Kohl [a]

[a] Furtwangen University, Department of Medical and Life Sciences, Jakob-Kienzle-Str. 17, 78054 VS-Schwenningen, Germany
[b] TU Kaiserslautern, AG Statistik, FB. Mathematik, P.O.Box 3049, 67653 Kaiserslautern, Germany
[c] Fraunhofer ITWM, Abt. Finanzmathematik, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 24 November 2014
Accepted 25 November 2014
Available online 3 December 2014

MSC:
62F12
62F35

Keywords:
Generalized linear models
L_2-differentiability
Shape scale model
Time series model for shape

ABSTRACT

We derive conditions for L_2 differentiability of generalized linear models with error distributions not necessarily belonging to exponential families, covering both cases of stochastic and deterministic regressors. These conditions induce smoothness and integrability conditions for corresponding GLM-based time series models.

1. Motivation

Introduced by Nelder and Wedderburn (1972), generalized linear models (GLMs) have become one of the most frequently used statistical models with a vast amount of published results. Hence, trying to give a full account on relevant literature would be pretentious. We instead refer to the monographs McCullagh and Nelder (1989) and Fahrmeir and Tutz (2001). When it comes to regularity assumptions, though, this literature focuses on GLMs which are exponential families, compare Haberman (1974, 1977); Fahrmeir (1990); Fahrmeir and Kaufmann (1985), or uses quasi-likelihood or pseudo-likelihood techniques to account for over/under-dispersion effects, see, a.o., Gouriéroux et al. (1984); Nelder and Pregibon (1987) and McCullagh and Nelder (1989). In some situations, exponential families are a too narrow class, though: e.g., recently log-linear models for generalized Pareto distributions have found applications in operational risk (compare Dahen and Georges, 2010), but distributions of extreme value type with unknown shape parameter do not fall into the range of exponential families and so far are not yet covered.

Heading for asymptotic results and robustness, we are not only interested in consistency results for specific estimators like maximum likelihood estimators (MLEs), but rather in *local asymptotic normality (LAN)* in the sense of Le Cam (1970) and Hájek (1972). With the LAN property at hand a very powerful asymptotic framework as pioneered by Le Cam is available: It gives a precise setup in which to obtain strong optimality results for (estimators behaving asymptotically like) the MLE, i.e., the Asymptotic Convolution Theorem and the Asymptotic Minimax Theorem, see, e.g. Rieder (1994, Theorems 3.2.3

* Corresponding author.
E-mail address: peter.ruckdeschel@itwm.fraunhofer.de (P. Ruckdeschel).

http://dx.doi.org/10.1016/j.spl.2014.11.020

and 3.3.8) or van der Vaart (1998, Theorems 8.8 and 8.11). The LAN property entails necessary expansions for asymptotic maximin tests with explicit terms for the asymptotic maximin power under local alternatives (Le Cam, 1986, Section 11.9); it is the starting point for efficient and adaptive estimation in semiparametric models (compare Bickel et al., 1993) and for a comprehensive theory of optimally-robust procedures (see Rieder (1994, Chapters 5 and 7)).

Now, a sufficient condition for the LAN property is given by L_2-differentiability (see, e.g. Rieder (1994, Theorem 2.3.5)), and – at least in the i.i.d. setting – this is a necessary condition, too, compare Le Cam and Yang (2000, Chapter 7, Proposition 3). Hence in this light, deriving smoothness of the model in terms of L_2-differentiability would be a desirable goal; i.e., to consider GLMs as particular parametric models and to derive their L_2-differentiability. For GLMs which are exponential families, this has already been achieved in Schlather (1994). Typically, however, scale-shape families as e.g. the generalized Pareto distributions are non-exponential. In this article, we hence generalize results of Rieder (1994, Section 2.4) on L_2-differentiability for linear regression models to also cover error distributions with a k-dimensional parameter and with regressors of possibly different length for each parameter. More specifically, we separately treat the case of stochastic regressors, which is of particular interest for incorporating (space-)time dependence, and of deterministic regressors as occurring in planned experiments.

While in principle L_2-differentiability of these models could be settled by general auxiliary results from Hájek (1972, Lemmas A.1–A.3), or be placed in the framework of Rieder and Ruckdeschel (2001), our goal are sufficient conditions directly exploiting the regression structure. More specifically, these conditions refer to (i) smoothness of the error distribution model, (ii) (uniform) integrability of the scores (L_2-derivative) and (iii) suitably integrated continuity of the Fisher information of again the error distribution model.

At first glance, this might look like a technical exercise but setting up time series models where time-dependence is captured by a GLM-type link with (functions of) the own past observations as regressors, conditions (ii) and (iii) reveal to which extent the current error distribution may depend upon the past without making it "over-informative" for the present. More precisely, letting aside dimensionalities of the parameter of the error distribution and the regressors, the scores function of a GLM $\mathscr{P}$ with errors from a distribution model $\mathscr{Q}$, link function ℓ and regressor x is of form $\Lambda_\beta^{\mathscr{P}}(x, y) = \Lambda_{\ell(x\beta)}^{\mathscr{Q}}(y)\dot{\ell}(x\beta)x$, where $\Lambda_\vartheta^{\mathscr{Q}}$ are the parametric scores from model $\mathscr{Q}$. Now even if $\mathscr{Q}$ has fat tails and non-existing moments, in many cases $\Lambda_\vartheta^{\mathscr{Q}}$ still is square integrable, see e.g. the case of α-stable distributions as in DuMouchel (1973) or the generalized extreme value and Pareto distributions GEVD and GPD explicated later on in this paper. If however, as in a autoregressive (AR) time series context with identity link $\ell(\theta) = \theta$, x comes again from a distribution within $\mathscr{Q}$, the LAN property may fail due to a lack of integrability. This is the case in Andrews et al. (2009, Theorem 3.3), where in addition the authors obtain slower convergence rates for β in an AR-model with α-stable errors. One way to preserve the LAN property could consist in using a suitable link function ℓ such that the product $\dot{\ell}x$ becomes square integrable—see later in this paper for corresponding GPD and GEVD time series. This technique can be seen as an alternative/an extension to the approach using regression ranks as in Hallin et al. (2011), which in the respective case of a regression model with α-stable errors and deterministic regressors achieves the same goal, i.e., extending the availability of the LAN property.

In this paper, we explicate the respective conditions (i)–(iii) for the cases of stochastic and deterministic regressors, respectively, in examples including – for reference and comparison – linear regression, Poisson, and Binomial regression, as well as scale-shape regression for the GPD and GEVD.

In particular for the latter distributions we give conditions which render a corresponding time series model accessible to the LAN type framework and thus contribute a new sort of GLM for extreme value type distributions where the tail weight respectively, the shape parameter depends on past observations in an autoregressive way. Thus, large extreme observations may foster or dampen the occurrence of future large extreme observations and controlling the extremal index (see Embrechts et al. (1997, pp. 413–423)) this way.

The rest of the paper is organized as follows: Section 2 provides the mathematical setup and the main results with Theorem 2.3 (for random carriers) and Theorem 2.6 (for deterministic carriers). The examples are worked out in Section 3. The proofs of our assertions are given in the Appendix.

2. Main results

Let $(\Omega, \mathscr{A})$ be a measurable space and $\mathscr{M}_1(\mathscr{A})$ the set of all probability measures on $\mathscr{A}$. Consider $\mathscr{Q} = \{Q_\vartheta | \vartheta \in \Theta\} \subset \mathscr{M}_1(\mathscr{A})$ a parametric model with open parameter domain $\Theta \subset \mathbb{R}^k$. Following Le Cam and Rieder, we write dQ_ϑ for the densities w.r.t. some dominating measure ν on $\mathscr{A}$ and denote the norm in the respective $L_2(\nu)$ space by $\|\cdot\|_{\mathscr{L}_2}$; as usual, ν is suppressed from notation as the choice of ν has no effect on respective convergence assertions. In this context, L_2 differentiability in the case of i.i.d. observations is defined as follows.

Definition 2.1. Model $\mathscr{Q}$ is called L_2 differentiable at $\vartheta \in \Theta$ if there exists a function $\Lambda_\vartheta^{\mathscr{Q}} \in L_2^k(P_\vartheta)$ such that, as $h \to 0 \in \mathbb{R}^k$

$$\left\| \sqrt{dQ_{\vartheta+h}} - \sqrt{dQ_\vartheta}\left(1 + \tfrac{1}{2}(\Lambda_\vartheta^{\mathscr{Q}})^{\mathsf{T}} h\right) \right\|_{\mathscr{L}_2} = \mathrm{o}(|h|). \tag{2.1}$$

Then, $\Lambda_\vartheta^{\mathscr{Q}}$ is the L_2 derivative and the $k \times k$ matrix $\mathscr{I}_\vartheta^{\mathscr{Q}} = \mathrm{E}_\vartheta \Lambda_\vartheta^{\mathscr{Q}} (\Lambda_\vartheta^{\mathscr{Q}})^{\mathsf{T}}$ is the Fisher information of $\mathscr{Q}$ at ϑ.

We say that $\mathcal{Q}$ is continuously L_2 differentiable at ϑ if, for any $h \to 0 \in \mathbb{R}^k$,

$$\sup_{t\in\mathbb{R}^k:\,|t|\le 1} \left\| \sqrt{dQ_{\vartheta+h}}(\Lambda^{\mathcal{Q}}_{\vartheta+h})^{\mathsf T} t - \sqrt{dQ_{\vartheta}}(\Lambda^{\mathcal{Q}}_{\vartheta})^{\mathsf T} t \right\|_{\mathcal{L}_2} = \mathrm{o}(1). \tag{2.2}$$

Introducing regressors to explain parameter ϑ, we turn model $\mathcal{Q}$ into a regression model $\mathcal{P}$ with parameter β. To this end, for $p \in \mathbb{N}$, let $\pi \in \mathbb{N}^k$, $\pi = (p_h)_{h=1,\ldots,k}$ be a partition of the p coordinates into blocks of dimension p_h, i.e., $\sum_h p_h = p$. Obviously, then each $x \in \mathbb{R}^p$ can unambiguously be indexed by the double index $(x_{h,j})_{\substack{h=1,\ldots,k\\ j=1,\ldots,p_h}}$. For these blocks we define the following operators:

$$T_\pi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^k, \qquad (a,b) \mapsto T_\pi(a,b) =: a^{\mathsf T_\pi} b = \left(\sum_{j=1}^{p_h} a_{h,j} b_{h,j} \right)_{h=1,\ldots,k} \tag{2.3}$$

$$\rho_\pi : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p, \qquad (c,a) \mapsto \rho_\pi(c,a) =: c \;\cdot_\pi\; a = (c_h a_{h,j})_{\substack{h=1,\ldots,k\\ j=1,\ldots,p_h}} \tag{2.4}$$

$$M_\pi : \mathbb{R}^{k\times k} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{p\times p}, \qquad (C,a,b) \mapsto M_\pi(C,a,b) = (C_{h_1,h_2} a_{h_1,j_1} b_{h_2,j_2})_{\substack{h_1,h_2=1,\ldots,k\\ j_1,j_2=1,\ldots,p_h}}. \tag{2.5}$$

We also write $C \;\cdot_\pi\; a$ for a $k \times m$ matrix C, meaning that we apply ρ_π to C column by column as first argument, so that the result will be the respective $p \times m$ matrix $(c_{h,l} a_{h,j})_{\substack{h=1,\ldots,k\\ j=1,\ldots,p_h}; l=1,\ldots,m}$.

Then, the case of a k-dimensional parameter ϑ in Model $\mathcal{Q}$ and non-identically dimensional regressors for each of the k coordinates can be captured using a continuously differentiable link function $\ell : \mathbb{R}^k \to \Theta$ with derivative $\dot\ell$, so that for a p-dimensional regressor X and p-dimensional regression parameter β we obtain a regression as $\vartheta = \ell(\theta)$ for $\theta = X^{\mathsf T_\pi}\beta$. Applying the chain rule, the candidate L_2 derivative in this regression model is

$$\Lambda^{\mathcal{P}}_\beta(x,y) = \dot\ell(\theta)^{\mathsf T} \Lambda^{\mathcal{Q}}_\vartheta(y) \;\cdot_\pi\; x. \tag{2.6}$$

The case of the linear regression model treated in Rieder (1994, Section 2.4) is obtained as a special case for $\mathcal{Q}$ an L_2-differentiable $k = 1$-dimensional location model and ℓ the identity. As in Rieder (1994, Section 2.4), we distinguish the cases of stochastic and deterministic regressors.

To apply conditions as in Hájek (1972), we need the notion of *absolute continuity* in k dimensions: Let $f : \mathbb{R}^k \to \mathbb{R}$; we call f *absolutely continuous*, if for all $a, b \in \mathbb{R}^k$ the function $G : [0,1] \to \mathbb{R}$, $s \mapsto G(s) = f(a + s(b-a))$ is *absolutely continuous* (as usual, see Rudin (1986, Chapter 6)).

For later reference we recall the results of Hájek (1972, Lemmas A.1–A.3):

Proposition 2.2 (*Hájek*). *Assume that in some $\vartheta_0 \in \Theta$ surrounded by some open neighborhood U, model $\mathcal{Q}$ satisfies*

(H.1) *The densities $dQ_\vartheta(y)$ are absolutely continuous in each $\vartheta \in U$ for Q_{ϑ_0}-a.e. y.*
(H.2) *The derivative $\frac{\partial}{\partial\vartheta} dQ_\vartheta(y) = \Lambda_\vartheta(y)\, dQ_\vartheta(y)$ exists in each $\vartheta \in U$ for Q_{ϑ_0}-a.e. y.*
(H.3) *The Fisher information $\mathcal{I}_\vartheta = \int \Lambda_\vartheta(y)\Lambda_\vartheta(y)^{\mathsf T} Q_\vartheta(dy)$ exists, (i.e., the integral is finite) and is continuous in ϑ on U.*

Then, $\mathcal{Q}$ is continuously L_2 differentiable in ϑ_0 with derivative Λ_{ϑ_0} and Fisher information $\mathcal{I}_{\vartheta_0}$.

2.1. Random carriers

In this context the regressors x are stochastic with distribution K, but the observations $(x,y)_i$ are then modeled as i.i.d. observations. To this end, let model $\mathcal{Q}$ be a k-dimensional L_2-differentiable model with parameter $\vartheta \in \Theta$ and derivative $\Lambda^{\mathcal{Q}}_\vartheta$ and Fisher information $\mathcal{I}^{\mathcal{Q}}_\vartheta$. The corresponding GLM induced by the link function $\ell : \mathbb{R}^k \to \Theta$ (with derivative $\dot\ell$) and partition π is given as

$$\mathcal{P} = \left\{ P_\beta(dx,dy) = Q_{\ell(x^{\mathsf T_\pi}\beta)}(dy|x)\, K(dx) \mid \beta \in \mathbb{R}^p;\; Q_\vartheta \in \mathcal{Q} \right\}. \tag{2.7}$$

We state the following result.

Theorem 2.3. *Let $\beta_0 \in \mathbb{R}^p$ and $\vartheta_t = \ell(\theta_t)$ for $\theta_t = x^{\mathsf T_\pi}(\beta_0 + t)$ as well as $\dot\ell_t = \dot\ell(\theta_t)$; further define $\mathcal{I}^{\mathcal{P}}_{\vartheta_t}(x) := M_\pi\left(\dot\ell_t^{\mathsf T} \mathcal{I}^{\mathcal{Q}}_{\vartheta_t} \dot\ell_t, x, x\right)$.*

Then model $\mathcal{P}$ from (2.7) is L_2 differentiable in β_0 if subsequent conditions (i)–(iii) *hold.*

(i) *Model $\mathcal{Q}$ fulfills* (H.1)–(H.3) *with "Q_{ϑ_0}-a.e. y" replaced by "P_{β_0}-a.e. (x,y)" in* (H.1) *and* (H.2).
(ii)

$$\int |\mathcal{I}^{\mathcal{P}}_{\vartheta_0}(x)|\, K(dx) < \infty, \tag{2.8}$$

(iii) *for every* $b \in (0, \infty)$,

$$\lim_{s\to 0} \sup_{|t|\le b} \int \left| |\mathscr{I}^{\mathscr{P}}_{\vartheta_{st}}(x)| - |\mathscr{I}^{\mathscr{P}}_{\vartheta_0}(x)| \right| K(dx) = 0, \tag{2.9}$$

where $|\mathscr{I}|$ *is the Frobenius matrix norm, i.e.,* $|\mathscr{I}|^2 = \mathrm{tr}\mathscr{I}^2$.

Then model $\mathscr{P}$ *is continuously* L_2 *differentiable in* β_0 *with derivative* $\Lambda^{\mathscr{P}}_{\beta_0}(x,y) = \dot{\ell}_0^{\mathrm{T}} \Lambda^{\mathscr{Q}}_{\vartheta_0}(y) \cdot_{\pi} x$ *and Fisher information*

$$\mathscr{I}^{\mathscr{P}}_{\beta_0} = \mathrm{E}_{\beta_0} \Lambda^{\mathscr{P}}_{\beta_0} (\Lambda^{\mathscr{P}}_{\beta_0})^{\mathrm{T}} = \int \mathscr{I}^{\mathscr{P}}_{\vartheta_0}(x) K(dx).$$

Remark 2.4. Sufficient conditions for (2.8) and (2.9) are $\int |\mathscr{I}^{\mathscr{P}}_{\vartheta_0}|\, |\dot{l}_0|^2\, |x|^2\, K(dx) < \infty$, and for every $b \in (0, \infty)$, it is required that $\lim_{s\to 0} \sup_{|t|\le b} \int \left| |\mathscr{I}^{\mathscr{Q}}_{\vartheta_{st}}|\, |\dot{\ell}_{st}|^2 - |\mathscr{I}^{\mathscr{Q}}_{\vartheta_0}|\, |\dot{\ell}_0|^2 \right| |x|^2\, K(dx) = 0$.

As just seen, the general GLM case comes with additional conditions for the link function ℓ and its derivative. For the linear regression case, they boil down to (i) L_2 differentiability of the one dimensional location case and (ii) finite second moment of x w.r.t. K. (iii) becomes void, as $\dot{\ell} \equiv 1$ and $\mathscr{I}^{\mathscr{Q}}$ does not depend on the parameter—compare Rieder (1994, Theorem 2.4.7).

2.2. Deterministic carriers

The case of deterministic carriers canonically leads to triangular schemes of independent, but no longer identically distributed observations. To this end, we take up Rieder (1994, Definition 2.3.8) and define a corresponding notion of L_2-differentiability:

For $n \in \mathbb{N}$ and $i = 1, \ldots, i_n$, let $(\Omega_{n,i}, \mathscr{A}_{n,i})$ be general sample spaces and $\mathscr{M}_1(\mathscr{A}_{n,i})$ the set of all probability measures on $\mathscr{A}_{n,i}$. Consider the array of parametric families of probability measures $\mathscr{P}_{n,i} = \{P_{n,i,\beta} | \beta \in \mathbb{R}^p\} \subset \mathscr{M}_1(\mathscr{A}_{n,i})$.

Definition 2.5. The parametric array $\mathscr{P} = (\bigotimes_{i=1}^{i_n} \mathscr{P}_{n,i})$ is called L_2 differentiable at $\beta_0 \in \mathbb{R}^p$ if there exists an array of functions $\Lambda^{\mathscr{P}}_{n,i,\beta_0} \in L_2^k(P_{n,i,\beta_0})$ such that for all $i = 1, \ldots, i_n$ and $n \ge 1$ the following conditions (2.10)–(2.12) are fulfilled.

$$\mathrm{E}_{n,i,\beta_0} \Lambda^{\mathscr{P}}_{n,i,\beta_0} = 0. \tag{2.10}$$

Let $\mathscr{I}^{\mathscr{P}}_{n,i,\beta_0} = \mathrm{E}_{n,i,\beta_0} \Lambda^{\mathscr{P}}_{n,i,\beta_0} (\Lambda^{\mathscr{P}}_{n,i,\beta_0})^{\mathrm{T}}$ and $\mathscr{I}^{\mathscr{P}}_{n,\beta_0} = \sum_{i=1}^{i_n} \mathscr{I}^{\mathscr{P}}_{n,i,\beta_0}$ and for $t \in \mathbb{R}^k$, we define $t_n = (\mathscr{I}^{\mathscr{P}}_{n,\beta_0})^{-\frac{1}{2}} t$ and $U_{n,i} = U_{n,i,\beta_0}(t) = t_n^{\mathrm{T}} \Lambda^{\mathscr{P}}_{n,i,\beta_0}$. Then, for all $\varepsilon \in (0, \infty)$ and all $t \in \mathbb{R}^k$ we require

$$\lim_{n\to\infty} \sum_{i=1,\ldots,i_n} \int_{\{|U_{n,i}|>\varepsilon\}} U_{n,i}^2\; dP_{n,i,\beta_0} = 0. \tag{2.11}$$

Finally, for all $b \in (0, \infty)$ we need

$$\lim_{n\to\infty} \sup_{|t|\le b} \sum_{i=1}^{i_n} \left\| \sqrt{dP_{n,i,\beta_0+t_n}} - \sqrt{dP_{n,i,\beta_0}} \left(1 + \frac{1}{2} U_{n,i,\beta_0}(t)\right) \right\|^2_{\mathscr{L}_2} = 0. \tag{2.12}$$

Then, in β_0 and at time n, $\mathscr{P}$ has L_2 derivative $(\Lambda^{\mathscr{P}}_{n,i,\beta_0})$ and Fisher information $\mathscr{I}^{\mathscr{P}}_{n,\beta_0}$.

$\mathscr{P}$ is continuously differentiable in β_0, if for each sequence $h_n \to 0 \in \mathbb{R}^p$,

$$\lim_{n\to\infty} \sup_{|t|\le b} \sum_{i=1}^{i_n} \left\| \sqrt{dP_{n,i,\beta_0+h_n}} U_{n,i,\beta_0+h_n}(t) - \sqrt{dP_{n,i,\beta_0}} U_{n,i,\beta_0}(t) \right\|^2_{\mathscr{L}_2} = 0. \tag{2.13}$$

Our GLM with deterministic regressors $x_{n,i} \in \mathbb{R}^p$ correspondingly is defined as $\mathscr{P} = \bigotimes_{i=1}^{i_n} \mathscr{P}_{n,i}$ with

$$\mathscr{P}_{n,i} = \left\{ P_{n,i,\beta_0}(dy) = Q_{\vartheta_{n,i}}(dy) \mid \beta_0 \in \mathbb{R}^p;\; \vartheta_{n,i} = \ell(x_{n,i}^{\mathrm{T}_\pi} \beta_0),\; Q_{\vartheta_{n,i}} \in \mathscr{Q} \right\}. \tag{2.14}$$

Rieder (1994, Theorem 2.4.2) shows that in the linear regression case, conditions (2.11) and (2.12) follow from the (uniform) smallness of the hat matrix $H_n = H_{n;i,j} = x_{n,i}^{\mathrm{T}} (\sum_{g=1}^{i_n} x_{n,g} x_{n,g}^{\mathrm{T}})^{-1} x_{n,j}$, which, as H_n is a projector, reduces to the Feller type condition

$$\lim_{n} \max_{i=1,\ldots,i_n} H_{n;i,i} = 0. \tag{2.15}$$

In our more general framework, one may still define a corresponding projector H_n locally (i.e., in β_0) as

$$H_n = H_{n;i,j;\beta_0} = L^{\mathrm{T}}_{n,i;\beta_0}(\mathscr{I}^{\mathscr{P}}_{n,\beta_0})^{-1}L_{n,j;\beta_0}, \qquad L_{n,i;\beta_0} = \dot{\ell}(\theta_{n,i})^{\mathrm{T}}(\mathscr{I}^{\mathscr{P}}_{n,i,\beta_0})^{1/2} \cdot_\pi x_{n,i} \tag{2.16}$$

and, locally, the (changes in the) fitted parameters $\vartheta_{n,i}$ (in a corresponding Fisher scoring procedure) then can be written as

$$\vartheta^{(\mathrm{new})}_{n,i} = \vartheta_{n,i} + \sum_{j=1}^{i_n}(\mathscr{I}^{\mathscr{P}}_{n,i,\beta_0})^{-1/2}H_{n;i,j}(\mathscr{I}^{\mathscr{P}}_{n,j,\beta_0})^{-1/2}\Lambda^{\mathscr{Q}}_{\vartheta_{n,j}}(y_{n,j}).$$

However, contrary to the linear regression case, in the general GLM case, the distribution of the standardized scores $(\mathscr{I}^{\mathscr{P}}_{n,j,\beta_0})^{-1/2}\Lambda^{\mathscr{Q}}_{\vartheta_{n,j}}(y_{n,j})$ is not invariant in β_0. Therefore, the proof for the linear regression fails at this point and condition (2.15) is not sufficient—compare for instance the one-dimensional GLM $\mathscr{P}$ at $\beta_0 = 1$ induced by the one-dimensional Poisson model $\mathscr{Q}$ with parameter $\lambda > 0$, $i_n = n$, the identity as link function and regressors $x_{n,i} = 1/n$. In fact, this is the standard example for a scheme satisfying the Feller condition but violating the Lindeberg condition. Also, not surprisingly, it is easy to see that Lindeberg condition (2.11) entails condition (2.15).

Theorem 2.6. *Model $\mathscr{P}$ from (2.14) is continuously L_2 differentiable in $\beta_0 \in \mathbb{R}^p$ with L_2 derivative $\Lambda^{\mathscr{P}}_{n,i,\beta_0} = \Lambda^{\mathscr{P}}_{\beta_0}(x_{n,i}, y)$ with $\Lambda^{\mathscr{P}}_{\beta_0}$ from (2.6) and Fisher information $\mathscr{I}^{\mathscr{P}}_{n,\beta_0}$ as given in Definition 2.5 if the following conditions* (i)–(iii) *are fulfilled.*

(i) *Model $\mathscr{Q}$ fulfills* (H.1)–(H.3).
(ii) *The Lindeberg condition (2.11) holds for $U_{n,i}$ defined as in Definition 2.5.*
(iii) *Let $\vartheta_{n,i,t} = \ell(\theta_{n,i,t})$ for $\theta_{n,i,t} = x^{\mathrm{T}_\pi}_{n,i}\left(\beta_0 + (\mathscr{I}^{\mathscr{P}}_{n,\beta_0})^{-1/2}t\right)$ and introduce the abbreviations $\mathscr{I}^{\mathscr{Q}}_{n,i,t} = \mathscr{I}^{\mathscr{Q}}_{\theta_{n,i,t}}$, $\dot{\ell}_{n,i,t} = \dot{\ell}(\theta_{n,i,t})$, and $\mathscr{I}^{\mathscr{P}}_{n,i,t} = M_\pi\left(\dot{\ell}^{\mathrm{T}}_{n,i,t}\mathscr{I}^{\mathscr{Q}}_{n,i,t}\dot{\ell}_{n,i,t}, x_{n,i}, x_{n,i}\right)$. Then, for every $b \in (0,\infty)$ it must hold*

$$\lim_{n\to\infty}\sup_{|t|\le b}\sum_{i=1}^{i_n} t_n^{\mathrm{T}}(\mathscr{I}^{\mathscr{P}}_{n,i,t} - \mathscr{I}^{\mathscr{P}}_{n,i,0})t_n = 0. \tag{2.17}$$

3. Examples

Example 3.1 (*Linear Regression*)**.** It is obvious that Theorem 2.3 can be applied to the linear regression model

$$\mathscr{P} = \{P_\beta(dx, dy) = F(dy - x^{\mathrm{T}}\beta)K(dx)\} \tag{3.1}$$

about the one dimensional location model

$$\mathscr{Q} = \{Q_\vartheta(dy) = F(dy - \vartheta)\} \tag{3.2}$$

for some probability F on $(\mathbb{R}, \mathbb{B})$ with finite Fisher information of location, if the latter is defined as $\sup_\varphi(\int\varphi'(x)\,dF)^2/(\int\varphi^2\,dF)$ where φ varies in the set $\mathscr{C}^1_0(\mathbb{R}\to\mathbb{R})$ of all continuously differentiable functions with compact support, see Huber (1981, Definition 4.1/Theorem 4.2)—finite Fisher information of location settles condition (i) of Theorem 2.3, condition (ii) as already noted boils down to $\int|x|^2\,K(dx) < \infty$ and condition (iii) is void.

Example 3.2 (*Binomial GLM with Logit Link and Poisson GLM with Log Link*)**.** The Binomial model $\mathrm{Binom}(m, p)$ for known size $m \in \mathbb{N}$, usually $m = 1$, and unknown success probability $p \in (0, 1)$ has error distribution with counting density $q_p(y) = \binom{m}{y}p^y(1-p)^{m-y}$ (on $y \in \{0, \dots, m\}$), hence condition (i) of Theorem 2.3 is obviously fulfilled with Fisher information $\mathscr{I}^{\mathscr{Q}}_p = m(p(1-p))^{-1}$. Choosing a logit link, i.e., $\ell(\theta) = e^\theta/(1+e^\theta)$, $\mathscr{I}^{\mathscr{Q}}_p\dot{\ell}(\theta)^2 = mp(1-p)$, conditions (ii) and (iii) become

$$\text{(ii)} \int e^{x^{\mathrm{T}}\beta}(1+e^{x^{\mathrm{T}}\beta})^{-2}|x|^2\,K(dx) < \infty,$$

$$\text{(iii)} \int e^{x^{\mathrm{T}}\beta}\frac{(e^{x^{\mathrm{T}}s}-1)(1-e^{x^{\mathrm{T}}(2\beta+s)})}{(1+e^{x^{\mathrm{T}}(\beta+s)})^2(1+e^{x^{\mathrm{T}}\beta})^2}|x|^2\,K(dx) \to 0, \quad s \to 0.$$

As in these expressions both integrands are bounded pointwise in x, if $|x|^2$ is integrable w.r.t. K, the Binomial GLM with logit-link is continuously L_2 differentiable.

The Poisson model $\mathrm{Pois}(\lambda)$ ($\lambda \in (0,\infty)$) has error distribution with counting density $q_\lambda(y) = e^{-\lambda}\lambda^y/y!$ (on $y \in \mathbb{N}$), hence condition (i) of Theorem 2.3 is obviously fulfilled with Fisher information $\mathscr{I}^{\mathscr{Q}}_\lambda = \lambda^{-1}$. Choosing log link, i.e., $\ell(\theta) = e^\theta$, $\mathscr{I}^{\mathscr{Q}}_\lambda\dot{\ell}(\theta)^2 = \lambda$, conditions (ii) and (iii) become

$$\text{(ii)} \int e^{x^{\mathrm{T}}\beta}|x|^2K(dx) < \infty, \qquad \text{(iii)} \int e^{x^{\mathrm{T}}\beta}(e^{x^{\mathrm{T}}s}-1)\,|x|^2\,K(dx) \to 0, \quad s \to 0.$$

Hence integrability of $e^{|x|(|\beta|+\delta)}|x|^2$ w.r.t. K implies continuous L_2 differentiability of the Poisson GLM with log-link.

These two conditions, i.e., $|x| \in L_2(K)$ for Binomial logit and $e^{|x|(|\beta|+\delta)}|x|^2 \in L_1(K)$ for the Poisson GLM with log-link recover the conditions mentioned in Fahrmeir and Tutz (2001, p. 47).

Example 3.3 (*GEVD and GPD Joint Shape-Scale Models with Componentwise Log Link*)**.** Both, the generalized extreme value distribution (GEVD) and the generalized Pareto distribution (GPD) come with a three-dimensional parameter (μ, σ, ξ) for a location or threshold parameter $\mu \in \mathbb{R}$, a scale parameter $\sigma \in (0, \infty)$ and a shape parameter $\xi \in \mathbb{R}$. While for the GEVD, in principle the three dimensional model is L_2-differentiable for $\xi \in (-1/2, 0)$ and $\xi \in (0, \infty)$, respectively, in the GPD model, the model including the threshold parameter is not covered by our theory for L_2-differentiable error models. The reason is basically, that observations close to the endpoint of the support in the GPD model carry overwhelmingly much information on the threshold. To deal with GEVD and GPD in parallel let us hence assume μ known in both models, and, for simplicity, $\mu = 0$. Then, parameter ϑ consists of scale σ and shape ξ. In both models, the scores $\Lambda_\vartheta^{\mathcal{Q}}$ on the quantile scale, i.e., $\Lambda_\vartheta(F_\vartheta^{-1}(u))$ for $F_\vartheta^{-1}(u)$ the respective quantile function, include terms of order $(1-u)^\xi$. Hence for condition (i), we need to assume that at least $\xi > -1/2$. Depending on the context, it can be reasonable to add further restrictions. e.g., in case of the GPD, we only obtain an unbounded support if $\xi \geq 0$; similarly, if we restrict attention to the special case of Fréchet distributions for GEV distributions, $\xi > 0$ is a natural restriction.

For parameter ϑ, we consider a continuously differentiable componentwise link function $\ell: \mathbb{R}^2 \to \Theta$, i.e., the link function is of the form $\ell(\theta) = (\ell_\sigma(x_\sigma^{\mathrm{T}}\beta_\sigma), \ell_\xi(x_\xi^{\mathrm{T}}\beta_\xi))$ where we partition the p-dimensional regressor x and parameter β accordingly to $x = (x_\sigma, x_\xi)$ and $\beta = (\beta_\sigma, \beta_\xi)$ so that $\theta = x^{\mathrm{T}\pi}\beta = (x_\sigma^{\mathrm{T}}\beta_\sigma, x_\xi^{\mathrm{T}}\beta_\xi)$. Then, based on the 2×2 Fisher information matrix $\mathscr{I}_{\sigma,\xi}^{\mathcal{Q}}$ for joint scale and shape with entries $I_{\sigma\sigma}$, $I_{\sigma\xi}$ and $I_{\xi\xi}$, we obtain

$$\dot{\ell}^{\mathrm{T}} \mathscr{I}_{\sigma,\xi}^{\mathcal{Q}} \dot{\ell} = \begin{pmatrix} \dot{\ell}_\sigma^2 I_{\sigma\sigma} & \dot{\ell}_\sigma \dot{\ell}_\xi I_{\sigma\xi} \\ \dot{\ell}_\sigma \dot{\ell}_\xi I_{\sigma\xi} & \dot{\ell}_\xi^2 I_{\xi\xi} \end{pmatrix}.$$

That is, conditions (ii) and (iii) of Theorem 2.3 become

$$\text{(ii)} \int \dot{\ell}_\sigma^2 (I_{\sigma\sigma} + I_{\sigma\xi})|x_\sigma|^2 K(dx) + \int \dot{\ell}_\xi^2 (I_{\xi\xi} + I_{\sigma\xi})|x_\xi|^2 K(dx) < \infty,$$

$$\begin{aligned} \text{(iii)} &\int (\dot{\ell}_{\sigma+s}^2 (I_{\sigma+s\sigma+s} + I_{\sigma+s\xi+s}) - \dot{\ell}_\sigma^2 (I_{\sigma\sigma} + I_{\sigma\xi}))|x_\sigma|^2 K(dx) \\ &+ \int (\dot{\ell}_{\xi+s}^2 (I_{\xi+s\xi+s} + I_{\sigma+s\xi+s}) - \dot{\ell}_\xi^2 (I_{\xi\xi} + I_{\sigma\xi}))|x_\xi|^2 K(dx) \to 0, \quad s \to 0. \end{aligned}$$

GEVD model: The scale-shape model GEVD$(0, \sigma, \xi)$ has error distribution $Q_\vartheta(y) = \exp\big(-(1 + \xi\frac{y}{\sigma})^{-\frac{1}{\xi}}\big)$. As mentioned, condition (i) of Theorem 2.3 is fulfilled as long as $\xi \in (-1/2, 0)$ or $\xi > 0$. This is reflected by the Fisher information matrix which reads

$$\mathscr{I}_{\sigma,\xi}^{\mathcal{Q}} = \xi^{-2} D \begin{pmatrix} I_{\sigma\sigma} & I_{\sigma\xi} \\ I_{\sigma\xi} & I_{\xi\xi} \end{pmatrix} D, \quad \text{where } D^{-1} = \operatorname{diag}(\sigma, \xi) \text{ and} \tag{3.3}$$

$$\begin{aligned} I_{\sigma\sigma} &= (\xi+1)^2 \Gamma(2\xi+1) - 2(\xi+1)\Gamma(\xi+1) + 1, \\ I_{\sigma\xi} &= -(\xi+1)^2\Gamma(2\xi+1) + (\xi^2+4\xi+3)\Gamma(\xi+1) + (\xi^2+\xi)\Gamma'(\xi)\Gamma(\xi+1) - \xi\Gamma'(1) - \xi - 1, \\ I_{\xi\xi} &= (\xi+1)^2\Gamma(2\xi+1) - 2\Gamma(\xi+3) - 2\xi\Gamma'(\xi)\Gamma(\xi+2) + 2\xi(\xi+1)\Gamma'(1) \\ &\quad + \xi^2(\Gamma''(1) + (\Gamma'(1))^2) + (\xi+1)^2 \end{aligned}$$

and has singularities in $\xi = 0$ and $\xi = -1/2$.

GPD model: The scale-shape model GPD$(0, \sigma, \xi)$, has a c.d.f. of $Q_\vartheta(y) = 1 - (1 + \xi\frac{y}{\sigma})^{-\frac{1}{\xi}}$ and here, for $\sigma > 0$ and $\xi > -\frac{1}{2}$ condition (i) is fulfilled with Fisher information matrix:

$$\mathscr{I}_{\sigma,\xi}^{\mathcal{Q}} = \frac{1}{1+2\xi} D \begin{pmatrix} 1, & 1 \\ 1, & 2(\xi+1) \end{pmatrix} D, \qquad D^{-1} = \operatorname{diag}(\sigma, \xi+1).$$

Again failure of condition (i) is reflected by a singularity at $\xi = -1/2$ of the Fisher information.

The canonical link function for the scale is log link $\ell_\sigma(x_\sigma^{\mathrm{T}}\beta_\sigma) = \exp(x_\sigma^{\mathrm{T}}\beta_\sigma)$, whereas due to a lack of equivariance in the shape, there is no such canonical link for this parameter. For our GEVD and GPD applications, however, (non-regression-based) empirical evidence speaks for shape ξ varying in $(0, 2)$. So a good link should not necessarily exclude values $\xi \notin (0, 2)$, but make them rather hard to attain. For this paper we even impose the sharp restriction $\xi > 0$.

Moreover, to use GLMs with GEVD and GPD errors in time series context to model parameter driven time dependences in the terminology of Cox (1981), a real challenge is to design (smooth and isotone) link functions such that the regressors may themselves follow a GEVD or a GPD distribution, as this implies very heavy tails against which we have to integrate.

More specifically, we aim at constructing an AR-type time series for the scale and shape of the form

$$X_t \sim \mathrm{GEVD}(\ell(X_{(t-1):(t-p_1)}{}^{\mathrm{T}}\beta_\sigma, X_{(t-1):(t-p_2)}{}^{\mathrm{T}}\beta_\xi)) \quad \text{for } X_{(t-1):(t-p)} = (X_{t-1}, \ldots, X_{t-p}). \tag{3.4}$$

In this model, negative values of β_ξ would dampen clustering of extremes, as then usually a large value stemming from a large positive shape parameter will be followed by an observation with low or even negative shape hence with much lighter tails, thus in general a smaller value; correspondingly β_ξ positive will foster clustering of extremes.

A straightforward guess would be to use the log link, but this does not work for GEVD or GPD time series, as then integrability (ii) fails. Thus, besides being smooth (for our theorem) and strictly increasing (for identifiability), an admissible link function must grow extremely slowly. To get candidates in case of the GEVD, note that all terms of the Fisher information matrix for GEVD are dominated by term $\Gamma(2\xi + 1)$, so conditions (ii) and (iii) are fulfilled if for large positive values θ_ξ, the link function grows so slowly to ∞ that $\Gamma(2\ell_\xi(\theta_\xi)) \approx \log(\theta_\xi)$, which for large x amounts to a behavior like the iterated logarithm $\log(\log(x))$; analogue arguments in case of the GPD suggest $\ell_\xi(\theta_\xi) \approx \log(\theta_\xi)$.

One possibility to achieve this for the GEVD for $p = 1$ is $\ell_\xi(\theta_\xi) = \log(f(\log(x_\xi)^{\mathrm{T}}\beta_\xi))$ where $f(x)$ for $x > 0$ is quadratic like $x^2/2 + x + 1$ and for $x < 0$ behaves like $a_1/(\log(a_2 - x))^2 + a_3$ for some $a_1, a_2, a_3 > 0$ such that f is continuously differentiable in 0 and $f(x) > e^{-1/2}$ always. As is shown in Appendix A.5, this choice indeed fulfills conditions (ii) and (iii).

With regard to the singularity in $\xi = 0$ of $\mathscr{I}^{\mathscr{Q}}_{\sigma,\xi}$ in (3.3), in many applications, it may turn out useful though to restrict shape ξ to lie in either $(-1/2, 0)$ or in $(0, \infty)$; correspondingly, one could suggest a rescaled binomial link $\ell = \ell^{\mathrm{Binom}}/2-1/2$ for the first case and shifting the link function ℓ_ξ sketched above to $\tilde{\ell}_\xi = \ell_\xi + 1/2$ in the second.

Of course, given an admissible link function, the next question would be whether for given starting values $x_{-1}, \ldots, x_{-\max(p_1,p_2)}$ a time series defined according to (3.4) for $t \geq 0$, using this link function is (asymptotically) stationary. This is out of scope for this paper and will be dealt with elsewhere.

Acknowledgments

This article is part of the Ph.D. thesis of Daria Pupashenko. All authors gratefully acknowledge financial support by the Volkswagen Foundation (I/85 429 - I/85 432) for the project "Robust Risk Estimation", http://www.mathematik.uni-kl.de/~wwwfm/RobustRiskEstimation. We also thank the associate editor for his helpful comments and drawing our attention to applications with α-stable error distributions.

Appendix. Proofs

A.1. Proof of Hájek's auxiliary result Proposition 2.2

Proof. Apparently, (H.1) and (H.2) are implied by continuous differentiability of the densities $dQ_\vartheta(y)$ w.r.t. ϑ. Hájek (1972) gives a proof of Proposition 2.2 for $dQ_\vartheta(y)$ Lebesgue densities and for $k = 1$, but our notion of absolute continuity for $k > 1$ from p. 4 reduces the problem to the situation of $k = 1$, which is possible here, as we require (H.1)–(H.3) on open neighborhoods. In addition, Hájek requires (H.1) for every y. Looking into his proof of his Lemma A.2, though, one does not need that $dQ_\vartheta(y)$ be Lebesgue densities, and in his Lemma A.3 one only needs absolute continuity for Q_{ϑ_0}-a.e. y. Finally, the asserted continuous L_2 differentiability (not mentioned in the cited reference) with regard to Definition 2.1 is just (H.3). A similar result, already for $k \geq 1$, but only for dominated $\mathscr{Q}$ and for continuous differentiability of $dQ_\vartheta(y)$ w.r.t. ϑ for Q_{ϑ_0}-a.e. y, is Witting (1985, Satz 1.194). □

A.2. Proof of the chain rule

Lemma A.1 (*Chain Rule*). *Let $\mathscr{Q} = \{Q_\vartheta \mid \vartheta \in \Theta\}$ a parametric model with open parameter domain $\Theta \subset \mathbb{R}^k$. Assume $\mathscr{Q}$ is L_2 differentiable in $\vartheta_0 \in \Theta$ with derivative $\Lambda^{\mathscr{Q}}_{\vartheta_0}$ and Fisher information $I^{\mathscr{Q}}_{\vartheta_0}$. Let $\ell\colon \Theta' \to \Theta$ with domain $\Theta' \subset \mathbb{R}^{k'}$ be differentiable in some $\theta_0 \in \Theta'$ such that $\ell(\theta_0) = \vartheta_0$ and with derivative denoted by $\dot{\ell}(\theta_0)$. Then $\tilde{\mathscr{Q}} = \{\tilde{Q}_\vartheta = Q_{\ell(\theta)} \mid \theta \in \Theta'\}$ is L_2 differentiable in θ_0 with derivative $\Lambda^{\tilde{\mathscr{Q}}}_\theta = (\dot{\ell}(\theta_0))^{\mathrm{T}}\Lambda^{\mathscr{Q}}_{\vartheta_0}$ and Fisher information $I^{\tilde{\mathscr{Q}}}_\theta = (\dot{l}(\theta_0))^{\mathrm{T}} I^{\mathscr{Q}}_{\vartheta_0}\dot{\ell}(\theta_0)$. If $\mathscr{Q}$ is continuously L_2 differentiable in ϑ_0, so is $\tilde{\mathscr{Q}}$ in θ_0.*

Proof. Let $h_n \to 0$, $n \to \infty$ in $\mathbb{R}^{k'}$, $|h_n| \neq 0$. We take $\vartheta_n := \ell(\theta_0 + h_n)$, $\vartheta_0 := \ell(\vartheta_0)$. Smoothness of link function ℓ implies:

$$\vartheta_n = \ell(\theta_0 + h_n) = \vartheta_0 + \dot{\ell}(\theta_0)h_n + r(\theta_0, h_n), \tag{A.1}$$

for some remainder function r such that

$$\lim_{n\to\infty} r(\theta_0, h_n)/|h_n| = 0. \tag{A.2}$$

Let Q_{ϑ_n} be dominated by some measure ν with density q_{ϑ_n}, i.e., $dQ_{\vartheta_n} = q_{\vartheta_n} d\nu$. By L_2 differentiability of model Q for $R_n := \int\Big(\sqrt{q_{\vartheta_n}} - \sqrt{q_{\vartheta_0}}\big(1 + \frac{1}{2}(\Lambda^{\mathscr{Q}}_{\vartheta_0})^{\mathsf{T}}(\vartheta_n - \vartheta_0)\big)\Big)^2 d\nu$, we have

$$\lim_{n\to\infty} R_n/|\vartheta_n - \vartheta_0|^2 = 0. \tag{A.3}$$

But by (A.1) we may write R_n as $R_n = \int (A_n - B_n)^2\, d\nu$ for

$$A_n := \sqrt{q_{\vartheta_n}} - \sqrt{q_{\vartheta_0}}\Big(1 + \frac{1}{2}(\Lambda^{\mathscr{Q}}_{\vartheta_0})^{\mathsf{T}}\dot{\mathrm{l}}(\vartheta_0)h_n\Big) \quad \text{and} \quad B_n := \frac{1}{2}\sqrt{q_{\vartheta_0}}(\Lambda^{\mathscr{Q}}_{\vartheta_0})^{\mathsf{T}} r(\vartheta_0, h_n).$$

Now, Cauchy–Schwarz entails that $A_n^2 \le 2(A_n - B_n)^2 + 2B_n^2$. Therefore

$$\begin{aligned}\int A_n^2 d\nu &\le 2\int (A_n - B_n)^2 d\nu + 2\int B_n^2 d\nu = 2R_n + 2\int B_n^2 d\nu \\ &\le 2R_n + \frac{1}{2}|r(\vartheta_0, h_n)|^2 \int q_{\vartheta_0}|\Lambda^{\mathscr{Q}}_{\vartheta_0}|^2 d\nu \le 2R_n + \frac{1}{2}|I^{\mathscr{Q}}_{\vartheta_0}|\,|r(\vartheta_0, h_n)|^2.\end{aligned}$$

Hence, using (A.1), (A.2), and (A.3)

$$\frac{1}{|h_n|^2}\int A_n^2 d\nu = \frac{2R_n}{|\vartheta_n - \vartheta_0|^2}\frac{\Big(\dot{\mathrm{l}}(\vartheta_0)h_n + r(\vartheta_0, h_n)\Big)^2}{|h_n|^2} + \frac{1}{2}|I^{\mathscr{Q}}_{\vartheta_0}|\frac{|r(\vartheta_0, h_n)|^2}{|h_n|^2} = \mathrm{o}(1).$$

That is, by Definition 2.1 model $\tilde{Q}$ is L_2 differentiable in $\vartheta_0 \in \Theta'$. □

A.3. Proof of Theorem 2.3

Let $s_n \to 0 \in \mathbb{R}^p$ for $n \to \infty$ such that $\tilde{s}_n = s_n/|s_n| \to \tilde{s}_0$ for some $\tilde{s}_0$ with $|\tilde{s}_0| = 1$. We take $\vartheta_s := \ell(\theta_s)$, $\theta_s := x^{\mathsf{T}}(\beta_0 + s)$, $\dot{\ell}_s = \dot{\ell}(\theta_s)$. Let $dQ_{\vartheta_n} = q_{\vartheta_n}\, d\nu$. By Definition 2.1 the GLM $\mathscr{P}$ is L_2 differentiable at every $\beta \in \mathbb{R}^p$ if $\lim_{n\to\infty} |s_n|^{-2} \iint \tilde{A}_n^2\, \nu(dy)\, K(dx) = 0$ for the A_n from Lemma A.1 now taking up the dependence on x, i.e.,

$$\tilde{A}_n = \tilde{A}_n(x, y) := \sqrt{q_{\vartheta_n}} - \sqrt{q_{\vartheta_0}}\,\Big(1 + \frac{1}{2}(\Lambda^{\mathscr{Q}}_{\ell(x^{\mathsf{T}}\beta_0)})^{\mathsf{T}}\dot{\ell}(x^{\mathsf{T}}\beta_0)\, \cdot_\pi\, x^{\mathsf{T}}s_n\Big). \tag{A.4}$$

Here (pointwise) existence (for P_β-a.e. (x, y)) and form of the L_2-derivative follow from (H.1) and the chain rule applied pointwise (in (x, y)). The proof of Lemma A.1 for K-a.e. x and s small enough provides some function $z(s) \to 0$ such that

$$\int \tilde{A}_n^2 \nu(dy) = |x^{\mathsf{T}}s_n|^2(z(x^{\mathsf{T}}s_n))^2.$$

Hence, for K-a.e. fixed x, $\tilde{A}'_n(x) := |s_n|^{-2}\int \tilde{A}_n^2\nu(dy) \to 0$. For Lebesgue measure λ, fixed $x \in \mathbb{R}^p$ and $u \in [0, 1]$ by the fundamental theorem of calculus for absolutely continuous functions, for K-a.e. fixed x we obtain

$$\begin{aligned}|s_n|^{-2}\int\Big(\sqrt{q_{\vartheta_{s_n}}} - \sqrt{q_{\vartheta_0}}\Big)^2 d\nu &= |s_n|^{-2}\int\Big(\int_0^1 \tfrac{1}{2}\sqrt{q_{\vartheta_{us_n}}}(\dot{\ell}^{\mathsf{T}}_{us_n}\Lambda^{\mathscr{Q}}_{\vartheta_{us_n}}\ \cdot_\pi\ x^{\mathsf{T}}s_n)\,\lambda(du)\Big)^2 d\nu \\ &\le \tfrac{1}{4|s_n|^2}\int\int_0^1 q_{\vartheta_{us_n}}(\dot{\ell}^{\mathsf{T}}_{us_n}\Lambda^{\mathscr{Q}}_{\vartheta_{us_n}}\ \cdot_\pi\ x^{\mathsf{T}}s_n)^2\,\lambda(du)\,d\nu \\ &= \frac{1}{4}\tilde{s}_n^{\mathsf{T}}\int_0^1 \mathscr{I}^{\mathscr{P}}_{\vartheta_{us_n}}(x)\,\lambda(du)\,\tilde{s}_n \\ &= \frac{1}{4|s_n|}\tilde{s}_n^{\mathsf{T}}\int_0^{|s_n|}\mathscr{I}^{\mathscr{P}}_{\vartheta_{u\tilde{s}_n}}(x)\,\lambda(du)\,\tilde{s}_n =: B_n(x).\end{aligned}$$

Now, introduce $B_0 = \tilde{s}_n^{\mathsf{T}}\mathscr{I}^{\mathscr{P}}_{\vartheta_0}(x)\tilde{s}_n/4$. By (ii) and (iii) $\int B_n(x)\,K(dx)$ is finite eventually in n, and by (iii) and Fubini

$$\int B_n(x)\,K(dx) = \frac{1}{4}\int_0^{|s_n|}\int |\mathscr{I}^{\mathscr{P}}_{\vartheta_{u\tilde{s}_n}}(x)|\,K(dx)\,\lambda(du) = \int B_0(x)\,K(dx) + \mathrm{o}(1).$$

Hence, by Vitali's Theorem (e.g. Rieder (1994, Proposition A.2.2)), B_n is uniformly integrable (w.r.t. K), and, as $\tilde{A}'_n(x) \le 2B_n(x) + 2B_0(x)$, so is $\tilde{A}'_n(x)$, and again by Vitali's Theorem, $\int \tilde{A}'_n(x)\,K(dx) \to 0$ which is (2.1). Continuity (2.2) with regard to Vitali's Theorem is just continuity of the Fisher information just shown.

The assertion of Remark 2.4 is shown similarly, replacing the B_n and B_0 from above with $|\mathscr{I}^{\mathscr{Q}}_{\vartheta_{st}}|\,|\dot{\ell}_{st}|^2\,|x|^2$ resp. $|\mathscr{I}^{\mathscr{Q}}_{\vartheta_0}|\,|\dot{\ell}_0|^2\,|x|^2$. □

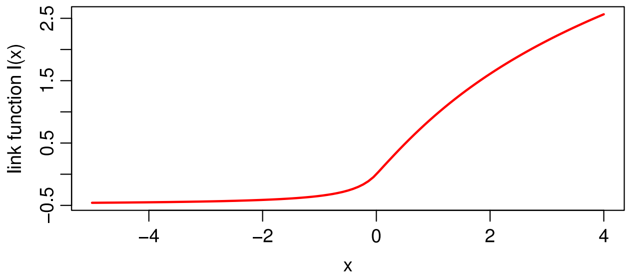

Fig. A.1. Link function for the shape of GEVD.

A.4. Proof of Theorem 2.6

For selfcontainedness, we reproduce the argument for condition (2.10) from Rieder (1994, Theorem 2.3.7). In model $\mathcal{Q}$, by (2.1), assuming ν-densities

$$\Big|\int \big(\sqrt{q_{\vartheta+h}} - \sqrt{q_\vartheta}\,\big(1+\tfrac{1}{2}(\Lambda_\vartheta^{\mathcal{Q}})^{\tau}h\big)\big)\sqrt{q_\vartheta}\,d\nu\Big|^2 \le \int \big|\sqrt{q_{\vartheta+h}} - \sqrt{q_\vartheta}\,\big(1+\tfrac{1}{2}(\Lambda_\vartheta^{\mathcal{Q}})^{\tau}h\big)\big|^2\,d\nu$$

where the RHS by L_2-differentiability of $\mathcal{Q}$ is $\mathrm{o}(|h|^2)$. Hence,

$$\begin{aligned}\mathrm{E}_\vartheta(\Lambda_\vartheta^{\mathcal{Q}})^{\tau}h + \mathrm{o}(|h|) &= \int(\sqrt{q_{\vartheta+h}}-\sqrt{q_\vartheta})\sqrt{q_\vartheta}\,d\nu = -\int(\sqrt{q_{\vartheta+h}}-\sqrt{q_\vartheta})^2\,d\nu/2\\ &= -h^{\tau}\mathcal{I}_\vartheta^{\mathcal{Q}}h/2 + \mathrm{o}(|h|^2) = \mathrm{o}(|h|).\end{aligned}$$

So $\mathrm{E}_\vartheta\Lambda_\vartheta^{\mathcal{Q}} = 0$, and hence also $\mathrm{E}_{n,i,\beta_0}\Lambda_{n,i,\beta_0}^{\mathcal{P}} = 0$. Lindeberg condition (2.11) is assumed without change, so it only remains to show condition (2.12). Let $N_{n,i}$ be the $Q_{\vartheta_{n,i,t_n}}$-null set such that both (H.1) and (H.2) hold for all $y \in N_{n,i}^c$. Let $N = \bigcup_n \bigcup_{i=1}^{i_n} N_{n,i}$. Then as in the case of stochastic regressors, from (H.1) and the chain rule applied pointwise (in $y \in N^c$) we obtain (pointwise) existence and form of the L_2-derivative. Let $\tilde{A}_n$ from (A.4) now take up the dependence on $x_{n,i}$, i.e., $\tilde{A}_{n,i} = \tilde{A}_n(x_{n,i})$ (with s_n from the preceding proof substituted by t_n) so that in particular, for every fixed i, $\tilde{A}'_{n,i} := \int \tilde{A}^2_{n,i}\nu(dy) \to 0$ as $t_n \to 0$. For condition (2.12) we have to show that $\lim_{n\to\infty}\sup_{|t|\le b}\sum_{i=1}^{i_n}\int \tilde{A}^2_{n,i}\nu(dy) = 0$. But, similarly as in the preceding proof for fixed i, by the fundamental theorem of calculus for absolutely continuous functions, we have

$$\tilde{A}'_{n,i} = \int\Big(\sqrt{q_{\vartheta_{n,i,t_n}}} - \sqrt{q_{\vartheta_{n,i,0}}}\Big)^2 d\nu \le \frac{1}{4|t_n|}\int_0^{|t_n|} t_n^{\tau}\mathcal{I}^{\mathcal{P}}_{n,i,ut}t_n\,\lambda(du) =: B_{n,i}.$$

Now, introduce $B_{0,i} = \frac{1}{4}t_n^{\tau}\mathcal{I}^{\mathcal{P}}_{n,i,0}t_n$ and note that $\sum_{i=1}^{i_n}\mathcal{I}^{\mathcal{P}}_{n,i,0} = \mathcal{I}^{\mathcal{P}}_{n,\beta_0}$, so $t_n^{\tau}\mathcal{I}^{\mathcal{P}}_{n,i,0}t_n = |t|^2 \le b$ and by (iii) $\sum_{i=1}^{i_n}B_{n,i} = \sum_{i=1}^{i_n}B_{0,i} + \mathrm{o}(1) = |t|^2/4 + \mathrm{o}(1)$. Hence, by Vitali's Theorem, $B_{n,i}$ is uniformly integrable (w.r.t. the counting measure), and, as $\tilde{A}'_{n,i} \le 2B_{n,i} + 2B_{0,i}$, so is $\tilde{A}'_{n,i}$, and again by Vitali's Theorem, $\sum_{i=1}^{i_n}\tilde{A}'_{n,i} \to 0$. Finally, continuity (2.13) again with regard to Vitali's Theorem is just continuity of the Fisher information just proven. □

A.5. Link function for the GEVD shape model

For GEVD for the shape we have chosen link function $\ell = \log(f(\beta\log(x_{t-1})))$, for

$$f(x) = (x^2/2 + x + 1)\mathbf{I}(x > 0) + (a_1(\log(a_2 - x))^{-2} + a_3)\mathbf{I}(x \le 0)$$

for some $a_1, a_2, a_3 > 0$. The constants a_1, a_2, a_3 are chosen so that f is continuously differentiable in 0 and $f(x) > e^{-1/2}$ always, i.e.

$$\frac{a_1}{(\log(a_2))^2} + a_3 = \frac{2a_1}{a_2(\log(a_2))^3} = 1, \qquad \frac{a_1}{(\log(a_2 - x))^2} + a_3 > e^{-1/2}, \quad \forall x < 0. \tag{A.5}$$

Since $a_1(\log(a_2 - x))^{-2} > 0$, to ensure the last inequality we let $a_3 = e^{-1/2} \approx 0.6063$. Solving the system of equations we get $a_2^{a_2} = e^{2(1-e^{-0.5})}$, so $a_2 \approx 1.624$ and $a_1 = 0.5a_2(\log(a_2))^3 \approx 0.00926$.

As said, shape is usually varying in (0, 2). As visible from Fig. A.1, this interval corresponds to an argument of the link function $x = \beta\log(x_{t-1})$ ranging in $(-\infty, \sqrt{1 - 2(1 - e^2)} - 1 \approx 2.712)$; hence, for $\beta = 1$, $\ell = \log(f(\beta\log(x_{t-1})))$ is smaller than 2 as long as $x_{t-1} < 15$ and $\ell < 3$ for $x_{t-1} < 193$.

To show that our choice of link function for GEVD, fulfills conditions (ii) and (iii), first we calculate its derivative $\dot{\ell} = \dot{f}/f$ and obtain $\dot{\ell} = (x+1)/(x^2/2 + x + 1)$ for $x > 0$ and $\dot{\ell} = 2a_1(a_2 - x)^{-1}(\log(a_2 - x))^{-3}$ for $x < 0$. Hence, for large x, $\dot{\ell}$ behaves like $2/x$, while for $x < 0$, it essentially behaves like $-x^{-1}(\log(-x))^{-3}$.

As mentioned, the term $\Gamma(2x)$ dominates all entries of all terms of $\mathscr{I}^{\mathscr{Q}}_{\sigma,\xi}$ in (3.3). Using the Stirling approximation, i.e., $\Gamma(x) \approx \sqrt{2\pi}\,\exp(x(\log(x) - 1/2))$, due to the double application of the logarithm in the link function we get that $\Gamma(2\ell_\xi(\theta_\xi))$ is approximately $\beta_\xi \log(x_\xi)$. By equivariance in μ and σ, therefore the integral of condition (ii) turns into: $B_1(\xi) := 4\beta_\xi^{-1} \int \log(x)\, K(dx) < \infty$ for $\beta_\xi > 0$ and, for $\beta_\xi < 0$, to $B_2(\xi) := \beta_\xi^{-1} \int \log(x)\big((\log(-\beta_\xi)) + \log(\log(x))\big)^{-6} K(dx)$. Finiteness of $B_1(\xi)$ and $B_2(\xi)$ follows from finiteness of $\mathrm{E}(\min\{1, (\log x)^k\})$ for $x \sim \mathrm{GEVD}(0, 1, \xi)$, $k \in \mathbf{N}$. Reconsidering $B_1(\xi), B_2(\xi)$ at $\xi + s$, for $|s| < h, h < 1$ we see that $\sup_{|s|<h} B_i(\xi + s) < \infty$ for $i = 1, 2$, hence, condition (iii) is a consequence of dominated convergence and continuity of Fisher information $I_{\xi\xi}$ in ξ.

References

Andrews, B., Calder, M., Davis, R.A., 2009. Maximum likelihood estimation for α-stable autoregressive processes. Ann. Statist. 37, 1946–1982.
Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Inference in Semiparametric Models. Springer, New York.
Cox, D.R., 1981. Statistical analysis of time series: some recent developments. Scand. J. Stat. 8, 93–115.
Dahen, H., Georges, D., 2010. Scaling models for the severity and frequency of external operational loss data. J. Banking Finance 34 (7), 1484–1496.
DuMouchel, W.H., 1973. On the asymptotic normality of the maximum likelihood estimate when sampling from a stable distribution. Ann. Statist. 1, 948–957.
Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer, New York.
Fahrmeir, L., 1990. Maximum likelihood estimation in misspecified generalized linear models. Statistics 21 (4), 487–502.
Fahrmeir, L., Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13 (1), 342–368.
Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models, second ed. Springer, New York.
Gouriéroux, C., Monfort, A., Trognon, A., 1984. Pseudo maximum likelihood methods: theory. Econometrica 52 (3), 681–700.
Haberman, S.J., 1974. Log-linear models for frequency tables with ordered classifications. Biometrics 30 (4), 589–600.
Haberman, S.J., 1977. Maximum likelihood estimates in exponential response models. Ann. Statist. 5 (5), 815–841.
Hájek, J., 1972. Local asymptotic minimax and admissibility in estimation. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1, pp. 175–194.
Hallin, M., Swan, Y., Verdebout, T., Veredas, D., 2011. Rank-based testing in linear models with stable errors. J. Nonparametr. Stat. 23, 305–320.
Huber, P., 1981. Robust Statistics. Wiley, New York.
Le Cam, L., 1970. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. Ann. Math. Statist. 41 (3), 802–828.
Le Cam, L., 1986. Asymptotic Methods in Statistical Decision Theory. Springer, New York.
Le Cam, L., Yang, G.L., 2000. Asymptotics in Statistics. Some Basic Concepts, second ed. Springer, New York.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman Hall, London.
Nelder, J.A., Pregibon, D., 1987. An extended quasi-likelihood function. Biometrika 74 (2), 221–232.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. Roy. Statist. Soc. Ser. A 370–384.
Rieder, H., 1994. Robust Asymptotic Statistics. Springer, New York.
Rieder, H., Ruckdeschel, P., 2001. Short proofs on L_r-differentiability. Statist. Decisions 19, 419–425.
Rudin, W., 1986. Real and Complex Analysis, third ed. McGraw-Hill Inc., New York.
Schlather, M., 1994. Glattheit von Generalisierten Linearen Modellen und statistische Folgerungen (Diplomarbeit) Universität Bayreuth.
van der Vaart, A., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.
Witting, H., 1985. Mathematische Statistik. Parametrische Verfahren bei festem Stichprobenumfang. Teubner, Stuttgart.