Abstract

We derive conditions for $L^2$ differentiability of generalized linear models with error distributions not necessarily belonging to an exponential family, covering both cases of stochastic and deterministic regressors. These conditions allow a formulation of smoothness and integrability conditions for corresponding time series models with innovations from generalized Pareto distributions or generalized extreme value error distributions. In these models, time dependence is induced by linking the respective scale and shape parameters to the own past observations.

Keywords Generalized linear models; $L^2$-differentiability; shape scale model; time series model for shape

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 Fahrmeir (1990); Fahrmeir and Kaufmann (1985); Haberman (1974, 1977), or uses quasi-likelihood or pseudo-likelihood techniques to account for over/under-dispersion effects, see Gouriéroux et al. (1984); McCullagh and Nelder (1989); Nelder and Pregibon (1987). 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, but rather in local asymptotic normality in the sense of Hájek (1972), LeCam (1970), and hence, 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, Sec. 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, Lem. A.1–A.3), 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.

The respective conditions (i)–(iii) for the cases of stochastic and deterministic regressors, respectively, are explicated in examples including—for reference and comparison—linear regression, Poisson, and Binomial regression, as well as scale-shape regression for the generalized Pareto distributions and the generalized extreme value distributions.

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 where the tail weight respectively, the shape parameter depends on past observations in an autoregressive way. Therefore, making it "over-informative" for the present.

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 2.11 (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 $(\Omega,\mathcal{A})$ be a measurable space and $\mathcal{M}(\mathcal{A})$ the set of all probability measures on $\mathcal{A}$. Consider $\mathfrak{Q} = \{Q_\theta | \theta \in \Theta \} \subset \mathcal{M}(\mathcal{A})$ a parametric model with open parameter domain $\Theta \subset \mathbb{R}^k$. Following LeCam and Rieder, we write $dQ_\theta$ for the densities w.r.t. some dominating measure $\nu$ on $\mathcal{A}$ and denote the norm in the respective $L_2(\nu)$ space by $\|\cdot\|_{L_2}$; as usual, $\nu$ is suppressed from notation as the choice of $\nu$ 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 $\mathfrak{Q}$ is called $L_2$ differentiable at $\theta \in \Theta$ if there exists a function $\Lambda_\theta^2 \in L_2^0(\mathbb{P}_0)$ such that, as $h \to 0 \in \mathbb{R}^k$

$$\left\| \sqrt{dQ_{\theta+h}} - \sqrt{dQ_\theta}(1 + \frac{1}{2}(\Lambda_\theta^2)^T)h \right\|_{L_2} = o(h)$$  \hspace{1cm} (2.1)

Then, $\Lambda_\theta^2$ is the $L_2$ derivative and the $k \times k$ matrix $\mathcal{F}_\theta = E_\theta \Lambda_\theta^2(\Lambda_\theta^2)^T$ is the Fisher information of $\mathfrak{Q}$ at $\theta$.

We say that $\mathfrak{Q}$ is continuously $L_2$ differentiable at $\theta$ if, for any $h \to 0 \in \mathbb{R}^k$, $\sup_{r \in \mathbb{R}^k: \|r\| \leq 1} \left\| \sqrt{dQ_{\theta+h}}(\Lambda_\theta^2)^T - \sqrt{dQ_\theta}(\Lambda_\theta^2)^T \right\|_{L_2} = o(1)$ \hspace{1cm} (2.2)

**Remark 2.2** Rieder (1994, Def. 2.3.6) in addition requires the local identifiability condition $I_\theta^0 = 0$ for $L_2$ differentiability. We drop this condition, as it turns out to be too strict when we are only interested in some lower dimensional, smooth aspect $\tau(\theta)$ of the parameter: As in the Gauss-Markov Theorem on best linear unbiased estimation, we then only have to require that $\partial \tau(\theta) \in \{I_\theta^0 \}$ for $\partial \tau(\theta)$ the Jacobian of $\tau$ in $\theta$.

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

$$T_{\pi} : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^k, \quad (a,b) \mapsto T_{\pi}(a,b) =: a^T \pi b = (\sum_{j=1}^{p_\beta} a_{h,b_j} b_{h,j})_{h=1,\ldots,k}$$  \hspace{1cm} (2.3)

$$\rho_{\pi} : \mathbb{R}^k \times \mathbb{R}^p \rightarrow \mathbb{R}^p, \quad (c,a) \mapsto \rho_{\pi}(c,a) =: c^T a = (c_{a,h})_{h=1,\ldots,k}$$  \hspace{1cm} (2.4)

$$M_{\pi} : \mathbb{R}^{k \times k} \times \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^{p \times p}, \quad (C,a,b) \mapsto M_{\pi}(C,a,b) = (C_{h_1,b_{h_1,j_1} b_{h_2,j_2} j_2} b_{h_2,j_2+1,\ldots,p_\beta})_{h_1=1,\ldots,k \beta=1,\ldots,p_\beta}$$  \hspace{1cm} (2.5)

We also write $C_{\cdot \cdot} a$ for a $k \times m$ matrix $C$, meaning that we apply $\rho_{\pi}$ to $C$ column by column as first argument, so that the result will be the respective $p \times m$ matrix $(c_{a,h})_{j=1,\ldots,p_\beta \beta=1,\ldots,p_\beta}$. Then, the case of a $k$-dimensional parameter $\vartheta$ in Model $\mathfrak{Q}$ 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 $\ell$, so that for a $p$-dimensional regressor $X$ and $p$-dimensional regression parameter $\beta$ we obtain a regression as $y = \ell(\theta) = X^T \beta$. Applying the chain rule, the candidate $L_2$ derivative in this regression model is

$$\Lambda_\beta^\vartheta(x,y) = \ell(\theta)^T \Lambda_\beta^\vartheta(y) \cdot x \quad (2.6)$$

The case of the linear regression model treated in [Rieder (1994) Sec. 2.4] is obtained as a special case for $\vartheta$ an $L_2$-differentiable $k = 1$-dimensional location model and $\ell$ the identity. As in [Rieder (1994) Sec. 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) chap. 6]).

For later reference we recall the results of [ Hájek (1972) Lem. A.1–A.3]:

**Proposition 2.3 (Hájek)** Assume that in some $\vartheta_0 \in \Theta$ surrounded by some open neighborhood $U$, model $\mathcal{D}$ satisfies

(H.1) The densities $dQ_\vartheta(y)$ are absolutely continuous in each $\vartheta \in U$ for $Q_{\vartheta_0}$-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 $Q_{\vartheta_0}$-a.e. $y$.

(H.3) The Fisher information $\mathcal{I}_\vartheta = \int \Lambda_\vartheta(y) \Lambda_\vartheta(y)^T Q_\vartheta(dy)$ exists, (i.e., the integral is finite) and is continuous in $\vartheta$ on $U$.

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

**Remark 2.4** Apparently, (H.1) and (H.2) are implied by continuous differentiability of the densities $dQ_\vartheta(y)$ w.r.t. $\vartheta$. In the cited reference, the proof is given for $dQ_{\vartheta}(y)$ Lebesgue densities and for $k = 1$. Our notion of absolute continuity for $k > 1$ from above 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 $Q_{\vartheta_0}$-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 $\mathcal{D}$ and for continuous differentiability of $dQ_\vartheta(y)$ w.r.t. $\vartheta$ for $Q_{\vartheta_0}$-a.e. $y$, is [Witting (1985) Satz 1.194].

### 2.1 Random Carriers

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

$$\mathcal{P} = \left\{ P_\beta(dx,dy) = Q_{\ell(x^T \beta)}(dy|x) K(dx) \mid \beta \in \mathbb{R}^p; Q_\vartheta \in \mathcal{D} \right\} \quad (2.7)$$

We state the following result.

**Theorem 2.5** Let $\beta_0 \in \mathbb{R}^p$ and $\vartheta_i = \ell(\theta_i)$ for $\theta_i = x^T (\beta_0 + t)$ as well as $\ell_i = \ell(\theta_i)$; further define $\mathcal{I}_{\vartheta_i}^\beta(x) := M_\mathcal{D} \left( \ell_i^T \mathcal{I}_{\vartheta_i}^\beta \ell_i, x, x \right)$. Then model $\mathcal{P}$ from (2.7) is $L_2$ differentiable in $\beta_0$ if subsequent conditions (i)–(iii) hold.

1. Model $\mathcal{D}$ fulfills (H.1)–(H.3) with “$Q_{\vartheta_0}$-a.e. $y$” replaced by “$P_{\beta_0}$-a.e. $(x,y)$” in (H.1) and (H.2).

2. 

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

3. 

$$\lim_{t \to 0} \sup_{\|p\| \leq b} \int |\mathcal{I}_{\vartheta_0}^\beta(x) - |\mathcal{I}_{\theta_i}^\beta(x)|| K(dx) = 0. \quad (2.9)$$

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


Then model $\mathcal{P}$ is continuously $L_2$ differentiable in $\beta_0$ with derivative $\Lambda_{\beta_0}(x,y) = \ell_{\beta_0}^T \Lambda_{\beta_0}(y) \cdot x$ and Fisher information

$$
\mathcal{F}_{\beta_0} = \mathbb{E}_{\beta_0} \Lambda_{\beta_0}^T (\Lambda_{\beta_0})^T = \int \mathcal{F}_{\beta_0}(x) K(dx)
$$

**Remark 2.6** Sufficient conditions for (2.8) and (2.9) are

$$
\int |\mathcal{F}_{\beta_0}(x)| K(dx) < \infty,
$$

and for every $b \in (0, \infty)$,

$$
\limsup_{n \to \infty} \sup_{|t| \leq b} \int \left( |\mathcal{F}_{n,b}(x)|^2 - |\mathcal{F}_{\beta_0}(x)|^2 \right) K(dx) = 0.
$$

**Remark 2.7** As just seen, the general GLM case comes with additional conditions for the link function $\ell$ 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 $\ell \equiv 1$ and $\mathcal{F}$ does not depend on the parameter—compare Rieder (1994, Thm. 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, Def 2.3.8) and define a corresponding notion of $L_2$-differentiability:

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

**Definition 2.8** The parametric array $\mathcal{P} = (\otimes_{i=1}^n \mathcal{P}_{n,i})$ is called $L_2$ differentiable at $\beta_0 \in \mathbb{R}^p$ if there exists an array of functions $\Lambda_{\beta_0} \in L_2^2(P_{n,i,\beta_0})$ such that for all $i = 1, \ldots, n$ and $n \geq 1$ the following conditions (2.12)–(2.14) are fulfilled:

$$
E_{n,i,\beta_0} \Lambda_{\beta_0} = 0
$$

Let $\mathcal{F}_{n,i,\beta_0} = E_{n,i,\beta_0} \Lambda_{n,i,\beta_0}^T (\Lambda_{n,i,\beta_0})^T$ and $\mathcal{F}_{\beta_0} = \sum_{i=1}^n \mathcal{F}_{n,i,\beta_0}$ and for $t \in \mathbb{R}^k$, we define $t_n = (\mathcal{F}_{n,\beta_0})^{-1} t$ and $U_{n,i} = U_{n,i,\beta_0}(t) = t_n^T \Lambda_{n,i,\beta_0}^T$. Then, for all $\varepsilon \in (0, \infty)$ and all $t \in \mathbb{R}^k$ we require

$$
\lim_{n \to \infty} \sum_{i=1}^n \int \{ |U_{n,i}| > \varepsilon \} U_{n,i}^2 dP_{n,i,\beta_0} = 0
$$

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

$$
\limsup_{n \to \infty} \sup_{|t| \leq b} \sum_{i=1}^n \left( \left| dP_{n,i,\beta_0} + t_n - dP_{n,i,\beta_0}(1 + \frac{1}{2} U_{n,i,\beta_0}(t)) \right| \right)^2 = 0
$$

Then, in $\beta_0$ and at time $n$, $\mathcal{P}$ has $L_2$ derivative $(\Lambda_{n,\beta_0})$ and Fisher information $\mathcal{F}_{\beta_0}$.

**Remark 2.9** As in Remark 2.2, we drop the local identifiability condition $\mathcal{I}_{\beta_0} > 0$ from Rieder (1994, Def. 2.3.8), again for the same reasons.

Our GLM with deterministic regressors $x_{n,i} \in \mathbb{R}^p$ correspondingly is defined as

$$
\mathcal{P} = \otimes_{i=1}^n \mathcal{P}_{n,i}
$$

with

$$
\mathcal{P}_{n,i} = \left\{ P_{n,i,\beta_0}(dy) = Q_{\theta_{n,i}}(dy) \right\} \beta_0 \in \mathbb{R}^p; \ \theta_{n,i} = \ell(x_{n,i}^T \beta_0), Q_{\theta_{n,i}} \in \mathcal{P}
$$
Remark 2.10 Rieder (1994, Theorem 2.4.2) shows that in the linear regression case, conditions (2.13) and (2.14) follow from the (uniform) smallness of the hat matrix \( H_n = H_{n,i,j} = x_{n,i}^T (\sum_{\ell = 1}^{b_n} x_{n,\ell} x_{n,\ell}^T)^{-1} x_{n,j} \), which, as \( H_n \) is a projector, reduces to the Feller type condition
\[
\lim_{n \to \infty} \max_{i,j} H_{n,i,j} = 0
\]  
(2.18)
In our more general framework, one may still define a corresponding projector \( H_n \) locally (i.e., in \( \beta_0 \)) as
\[
H_n = H_{n,i,j,\beta_0} = L_{n,i,\beta_0}^{-1} L_{n,j,\beta_0} \quad \text{and, locally, (the changes in the)} \quad \text{parameters of a corresponding Fisher scoring procedure) then can analogously be written as}
\[
\hat{\phi}_{n,i}^{(0)} = \hat{\phi}_{n,i} + \sum_{j=1}^{n} \left( \hat{I}_{n,j,\beta_0}^{-1} H_{n,i,j,\beta_0}^{-1} \hat{I}_{n,j,\beta_0} \right) \Lambda_{n,j,\beta_0} (y_{n,j})
\]  
(2.19)
However, contrary to the linear regression case, in the general GLM case, the distribution of the standardized scores \( (\hat{I}_{n,i,\beta_0}^{-1/2} \Lambda_{n,i,\beta_0} (y_{n,i})) \) is not invariant in \( \beta_0 \). Therefore, the proof for the linear regression fails at this point and condition (2.18) is not sufficient—compare for instance the one-dimensional GLM \( \mathcal{P} \) at \( \beta_0 = 1 \) induced by the one-dimensional Poisson model \( \mathcal{P} \) with parameter \( \lambda > 0 \), \( p = n \), the identity as link function and regressors \( x_{n,i} = i/n \); 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 the Lindeberg condition (2.13) entails condition (2.18).

Theorem 2.11 Model \( \mathcal{P} \) from (2.16) is continuously \( L_2 \) differentiable in \( \beta_0 \in \mathbb{R}^p \) with \( L_2 \) derivative \( \Lambda^\varphi_{n,i,\beta_0} = \Lambda^\varphi_{n,i,\beta_0} (x_{n,i}, y) \) with \( \Lambda^\varphi_{n,i,\beta_0} \) from (2.16) and Fisher information \( \mathcal{I}^\varphi_{n,\beta_0} \) as given in Definition 2.8 if the following conditions (i)-(iii) are fulfilled.

(i) Model \( \mathcal{P} \) fulfills (H.1)–(H.3).

(ii) The Lindeberg condition (2.13) holds for \( U_{n,i} \) defined as in Definition 2.8.

(iii) Let \( \hat{\phi}_{n,i} = \hat{\ell}(y_{n,i}, \beta_0) \) for \( \theta_{n,i} = x_{n,i}^T \beta_0 + (\mathcal{I}^\varphi_{n,i,\beta_0})^{-1/2} \) and introduce the abbreviations \( \mathcal{I}^\varphi_{n,i,\beta_0} = \mathcal{I}^\varphi_{\theta_{n,i}} \), \( \hat{\ell}_{n,i} = \hat{\ell}(y_{n,i}, \beta_0) \), and \( \mathcal{I}^\varphi_{n,i,\beta_0} = \mathcal{I}^\varphi_{\theta_{n,i}} \). Then, for every \( b \in (0, \infty) \) it must hold
\[
\lim_{n \to \infty} \sup_{|t| \leq b} \sum_{i=1}^{n} \left( \mathcal{I}^\varphi_{n,i,\beta_0} \right)^2 |t_{n,i}|^2 = 0
\]  
(2.21)

Remark 2.12 Corresponding analogues to conditions (2.10) and (2.11) are
\[
\lim_{n \to \infty} \sup_{|t| \leq b} \left| |t_{n,i}|^2 \sum_{i=1}^{n} \mathcal{I}^\varphi_{n,i,\beta_0} \right| |t_{n,0}|^2 < \infty
\]  
(2.22)
and
\[
\lim_{n \to \infty} \sup_{|t| \leq b} \left| \sum_{i=1}^{n} \left( \mathcal{I}^\varphi_{n,i,\beta_0} - \mathcal{I}^\varphi_{n,i,\beta_0} \right)^2 |x_{n,i}|^2 \right| = 0
\]  
(2.23)
Note that the analogue to condition (2.8) is automatically fulfilled.

3 Examples

Example 3.1 (Linear regression) It is obvious that Theorem 2.5 can be applied to the linear regression model
\[
\mathcal{P} = \{ P_\beta(dx, dy) = F(dy - x^T \beta) K(dx) \}
\]  
(3.1)
about the one-dimensional location model
\[
\mathcal{P} = \{ Q_\theta(dy = F(dy - \theta) \}
\]  
(3.2)
for some probability \( F \) on \( (\mathbb{R}, \mathcal{B}) \) with finite Fisher information of location \( \sup_\phi (|\phi'(x)| dF)/(|\phi'^2 dF|) \) where \( \phi \) varies in the set \( \mathcal{X}_0^2(\mathbb{R} \to \mathbb{R}) \) of all continuously differentiable functions with compact support. see Huber, 1981—finite Fisher information of location settles condition (i) of Theorem 2.5, 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 $	ext{Binom}(n, p)$ for known size $n \in \mathbb{N}$, usually $n = 1$, and unknown success probability $p \in (0, 1)$ has error distribution with counting density $q_p(y) = \binom{n}{y} p^y (1 - p)^{n-y}$ (on $y \in \{0, \ldots, n\}$), hence condition (i) of Theorem 2.5 is obviously fulfilled with Fisher information $\mathcal{F}_p^2(\ell(\theta))^2 = mn(1-p)^{-1}$. Choosing a logit link, i.e., $\ell(\theta) = \theta/(1 + \theta)$, $\mathcal{F}_p^2(\ell(\theta))^2 = mp(1-p)$, conditions (ii) and (iii) become

\[
\int \frac{e^{\ell^T \beta}}{(1+e^{\ell^T \beta})^2} |x|^2 K(dx) < \infty, \quad \int \frac{e^{\ell^T \beta} (e^{\ell^T x} - 1)(1 - e^{\ell^T (2\beta + \theta)})}{(1+e^{\ell^T (\beta + \theta)})^2} |x|^2 K(dx) \to 0, \quad s \to 0.
\]

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 $\text{Pois}(\lambda)$ $(\lambda \in (0, \infty))$ has error distribution with counting density $q_p(y) = e^{-\lambda} \lambda^y / y!$ (on $y \in \mathbb{N}$), hence condition (i) of Theorem 2.5 is obviously fulfilled with Fisher information $\mathcal{F}_\lambda^2 = \lambda^{-1}$. Choosing log link, i.e., $\ell(\theta) = \theta$, $\mathcal{F}_\lambda^2(\ell(\theta))^2 = \lambda$, conditions (ii) and (iii) become

\[
\int e^{\ell^T |x|^2} K(dx) < \infty, \quad \int e^{\ell^T (e^{\ell^T x} - 1)} |x|^2 K(dx) \to 0, \quad s \to 0.
\]

Hence if $e^{\ell^T |x|^2}$ is integrable w.r.t. $K$ then the Poisson GLM with log-link, is continuously $L_2$ differentiable.

These two conditions, i.e., $|x| \in L_2(K)$ for Binomial logit and $e^{\ell^T |x|^2}$ 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 $(\mu, \sigma, \xi)$ 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 $\mu$ known in both models, and, for simplicity, $\mu = 0$. Then, parameter $\theta$ consists of scale $\sigma$ and shape $\xi$. In both models, the scores $A_{\sigma, \xi}$ on the quantile scale, i.e., $A_{\sigma, \xi}^{-1}(u)$ for $F_{\sigma, \xi}^{-1}(u)$ the respective quantile function, include terms of order $(1 - u)^{\xi}$. Hence for condition (ii), 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 $\xi > 0$; similarly, if we restrict attention to the special case of Fréchet distributions for GEV distributions, $\xi > 0$ is a natural restriction.

For parameter $\theta$, we consider a continuously differentiable componentwise link function $\ell: \mathbb{R}^2 \to \mathbb{R}$, i.e., the link function is of the form $\ell(\theta) = (\ell_\sigma(\xi, \beta_1, \beta_2), \ell_\xi(\sigma, \beta_1, \beta_2))$ where we partition the $p$-dimensional regressor $x$ and parameter $\beta$ accordingly to $x = (x_\sigma, x_\xi)$ and $\beta = (\beta_1, \beta_2)$ so that $\theta = x^T \beta = (x_\sigma^T \beta_1, x_\xi^T \beta_2)$. Then, based on the $2 \times 2$ Fisher information matrix $\mathcal{F}_{\sigma, \xi}^2$ for joint scale and shape with entries $I_{\sigma, \sigma}, I_{\sigma, \xi}$ and $I_{\xi, \xi}$, we obtain

\[
\ell^T \mathcal{F}_{\sigma, \xi}^2 \ell = \begin{pmatrix} I_{\sigma, \sigma}^2 & I_{\sigma, \xi}^2 \\ I_{\sigma, \xi}^2 & I_{\xi, \xi}^2 \end{pmatrix}
\]

That is, conditions (ii) and (iii) of Theorem 2.5 become

\[
\int (I_{\sigma, \sigma}^2 + I_{\sigma, \xi}^2) |x_\sigma|^2 K(dx) + \int (I_{\xi, \sigma}^2 + I_{\xi, \xi}^2) |x_\xi|^2 K(dx) < \infty, \\
\int (I_{\sigma, \sigma}^2 + I_{\sigma, \xi}^2 + I_{\sigma, \xi}^2 + I_{\xi, \xi}^2) |x_\sigma|^2 K(dx) + \\
+ \int (I_{\xi, \sigma}^2 + I_{\xi, \xi}^2 + I_{\sigma, \xi}^2 + I_{\xi, \xi}^2) |x_\xi|^2 K(dx) \to 0, \quad s \to 0.
\]

GEVD model: The scale shape model GEVD$(0, \sigma, \xi)$ has error distribution $Q_\theta(y) = \exp\left(-\left(1 + \frac{y}{\xi} - \frac{y^\xi}{\xi} \right)\right)$. As mentioned, condition (i) of Theorem 2.5 is fulfilled as long as $\xi \in (-1/2, 0)$ or $\xi > 0$. This is reflected by the Fisher information matrix which reads

\[
\mathcal{F}_{\sigma, \xi}^2 = \xi^{-2} D \begin{pmatrix} I_{\sigma, \sigma} & I_{\sigma, \xi} \\ I_{\sigma, \xi} & I_{\xi, \xi} \end{pmatrix} D, \quad \text{where} \quad D^{-1} = \text{diag}(\sigma, \xi)
\]

$D = \begin{pmatrix} \xi + 1 & 0 \\ 0 & \xi \end{pmatrix}^2 \Gamma(2\xi + 1) - (\xi + 1) \Gamma(\xi + 1) + 1,$

$I_{\sigma, \sigma} = - \xi (\xi^2 + 4 + 3 \xi) \Gamma(\xi + 1) + (\xi^2 + \xi) \Gamma(\xi + 1) - \xi \Gamma(\xi + 1) - \xi - 1,$

$I_{\sigma, \xi} = \xi (\xi + 1)^2 \Gamma(\xi + 1) - 2 \Gamma(\xi + 1) \Gamma(\xi + 1) + 1,$

$I_{\xi, \xi} = \xi (\xi + 1)^2 \Gamma(\xi + 1) - 2 \Gamma(\xi + 1) \Gamma(\xi + 1) + 1,$
and has singularities in $\xi = 0$ and $\xi = -1/2$.

**GPD model:** For the scale-shape model GPD$(0, \sigma, \xi)$, the error distribution is $Q_0(y) = 1 - (1 + \frac{y}{\sigma})^{-\frac{1}{\xi}}$ and here, for $\sigma > 0$ and $\xi > -\frac{1}{2}$ condition (i) is fulfilled with Fisher information matrix:

$$\mathcal{F}_0^{\xi} = \frac{1}{1 + 2\xi^2} D  \begin{pmatrix} 1 & 1 \\ 1 & 2(\xi + 1) \end{pmatrix} D, \quad D^{-1} = \text{diag}(\sigma, \xi + 1).$$

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

The canonical choice of the link function for the scale is log link $\ell_0(x_0^\xi \beta_0) = \exp(x_0^\xi \beta_0)$ 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 $\xi$ 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 dependencies 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 autoregressive-type time series for the scale and shape of the form

$$X_t \sim \text{GEVD}((\ell(X_{t-1};\beta_0),X_{t-1};\beta_0)) \quad \text{for} \quad X_{t-1;\beta_0} = (X_{t-1},\ldots,X_{t-p})$$

(3.3)

In this model, negative values of $\beta_2$ 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 $\beta_2$ positive will foster clustering of extremes.

A straightforward guess would be to use the log link, but this does not work for GEVD or GLM time series, as then integrability (ii) fails. Thus, besides being smooth (for our theorem) and strictly increasing (for identifyability), 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 $\beta_2$, the link function grows so slowly to $\infty$ that $\Gamma(2(\xi(\beta_2))) \approx \log(\beta_2)$, which for large $x$ amounts to a behavior like the iterated logarithm $\log(\log(\log(x)))$: analogue arguments in case of the GPD suggest $\ell_x(\theta) \approx \log(\theta)$.

One possibility to achieve this for the GEVD for $p = 1$ is $\ell_1(\theta) = \log(f(x)^2/\beta_2^2)$ where $f(x)$ for $x > 0$ is quadratic like $x^2/2 + x + 1$ and for $x < 0$ behaves like $a_1/(\log(x^2 - 1)) + a_2$ for some $a_1, a_2 > 0$ such that $f$ is continuously differentiable in $0$ and $f(x) > e^{-1/2}$ always. As is shown in appendix [A.4], this choice indeed fulfills conditions (ii) and (iii).

**Remark 3.4** Of course given this admissible link function, the next question would be whether for given starting values $X_{-1,\ldots,-\max(p,0)}$ a time series defined according to (3.3) for $r \geq 0$, using this link function is (asymptotically) stationary. This is out of scope for this paper and will be dealt with elsewhere.

## A Proofs

### A.1 Proof of the Chain rule

**Lemma A.1 (Chain rule)** Let $\mathcal{Q} = (Q_0 \mid \theta \in \Theta)$ a parametric model with open parameter domain $\Theta \subset \mathbb{R}^k$. Assume $\mathcal{Q}$ is $L_2$ differentiable in $\theta_0 \in \Theta$ with derivative $\Lambda_{\theta_0}$ and Fisher information $I_{\theta_0}$. Let $\ell : \Theta \to \Theta$ with domain $\Theta' \subset \mathbb{R}^k$ be differentiable in some $\theta_0 \in \Theta'$ such that $\ell(\theta_0) = \theta_0$ and with derivative denoted by $\ell'(\theta_0)$. Then $\mathcal{Q} = (Q_{\ell(\theta)} \mid \theta \in \Theta')$ is $L_2$ differentiable in $\theta_0$ with derivative $\Lambda_{\theta_0} = (\ell'(\theta_0))' \Lambda_{\theta_0}$ and Fisher information $I_{\ell(\theta_0)} = (\ell'(\theta_0))' I_{\theta_0} (\ell'(\theta_0))$. If $\mathcal{Q}$ is continuously $L_2$ differentiable in $\theta_0$, so is $\mathcal{Q}$ in $\theta_0$.

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

$$\theta_0 = \ell(\theta_0 + h_n) = \theta_0 + \ell'(\theta_0)h_n + r(\theta_0, h_n),$$

(A.1)

for some remainder function $r$ such that

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

(A.2)

Let $Q_{\theta_0}$ be dominated by some measure $\nu$ with density $q_{\theta_0}$, i.e., $dQ_{\theta_0} = q_{\theta_0} d\nu$. By $L_2$ differentiability of model $Q$ for

$$R_n := \left(\sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}}(1 + \frac{1}{2}(\Lambda_{\theta_0}')(\theta_0 - \theta_0))\right)^2 d\nu, \quad \text{we have} \quad \lim_{n \to \infty} \frac{R_n}{|\theta_n - \theta_0|^2} = 0.$$  

(A.3)

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

$$A_n := \sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}}(1 + \frac{1}{2}(\Lambda_{\theta_0}')(\theta_0)) \quad \text{and} \quad B_n := \frac{1}{2} \sqrt{q_{\theta_0}}(\Lambda_{\theta_0}')\ell(\theta_0, h_n).$$


Now Cauchy-Schwarz entails that $A_n^2 = (A_n - B_n + B_n)^2 \leq 2(A_n - B_n)^2 + 2B_n^2$. Therefore
\[
\int A_n^2 \, dv = 2\int (A_n - B_n)^2 \, dv + 2\int B_n^2 \, dv = 2R_n + 2\int B_n^2 \, dv \leq 2R_n + \frac{1}{2}r(0, h_n)^2 \int q_n |A_n^2| \, dv \leq 2R_n + \frac{1}{2}l_n^0 |r(0, h_n)|^2.
\]
Hence using (A.1), (A.2), and (A.3)
\[
\frac{1}{|h_n|^2} \int A_n^2 \, dv = 2R_n \left( \frac{\bar{t}(0)_n h_n + r(0, h_n)}{|h_n|^2} \right)^2 + \frac{1}{2}l_n^0 |r(0, h_n)|^2 = o(1).
\]
That is, by Definition 2.1 model $\tilde{Q}$ is $L_2$ differentiable in $\theta_0 \in \Theta'$.

### A.2 Proof of Theorem 2.5
Let $s_n \to 0$ in $\mathbb{R}^p$ for $n \to \infty$ such that $s_n = s_n/|s_n| \to \tilde{s}_0$ for some $\tilde{s}_0$ with $|\tilde{s}_0| = 1$. We take $\theta_n := \ell(\theta), \theta_n := x(\tilde{s}_0 + s), \ell_n := \ell(\theta_n)$. Let $Q_{\theta_0}$ have $v$ density $q_{\theta_0}$. By Definition 2.1 the GLM $\mathcal{P}$ is $L_2$ differentiable at every $\tilde{\theta} \in \mathbb{R}^p$ if $\lim_{n \to \infty} |s_n|^{-2} \int A_n^2 \, dv(\tilde{\theta}) K(\tilde{\theta}) = 0$ for the $A_n$ from Lemma A.1 now taking up the dependence on $x$, i.e.,
\[
\tilde{A}_n \equiv \tilde{A}_n(x, y) := \sqrt{q_{\theta_n}} - \sqrt{q_{\theta_0}} \left( 1 + \frac{1}{2} (A_n^2 \tilde{s}_0 + \tilde{s}_n)^2 \tilde{t}(x^T \tilde{s}_0 - y^T) \right) \tilde{s}_n.
\]
Here (pointwise) existence (for $P_{\tilde{\theta}}$-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 $\tilde{z}(s) \to 0$ such that
\[
\int A_n^2 \, dv(\tilde{\theta}) = |x^T s_n|^2 (\tilde{z}(x^T s_n))^2.
\]
Hence, for K-a.e. fixed $x, \tilde{A}_n^2(x) := |s_n|^{-2} \int A_n^2 \, dv(\tilde{\theta}) \to 0$. For Lebesgue measure $\lambda$, 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
\[
|s_n|^{-2} \int \left( \sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}} \left( 1 + \frac{1}{2} (A_n^2 \tilde{s}_0 + \tilde{s}_n)^2 \tilde{t}(x^T \tilde{s}_0 - y^T) \right) \tilde{s}_n \right)^2 \, dv(\tilde{\theta}) \leq \frac{1}{4|s_n|^2} \int_0^1 q_{\theta_0} \tilde{A}_n(x, y) \lambda(du) \, dv(\tilde{\theta}) \tilde{s}_n = \frac{1}{4|s_n|^2} \int_0^1 \tilde{A}_n(x, y) \lambda(du) \, dv(\tilde{\theta}) \tilde{s}_n =: B_n(x)
\]
Now introduce $B_0 = \frac{1}{4} |\tilde{A}_n(x, y)| \tilde{s}_n$. By (ii) and (iii) $B_n(x) K(\tilde{\theta})$ is finite eventually in $n$, and by (iii) and Fubini
\[
\int B_n(x) K(\tilde{\theta}) = \int B_n(x) K(\tilde{\theta}) \, dv(\tilde{\theta}) = \int B_0(x) K(\tilde{\theta}) + o(1)
\]
Hence by Vitali’s Theorem (e.g., Rieder (1994) Prop. A.2.2), $B_0$ is uniformly integrable (w.r.t. $K$), and, as $\tilde{A}_n(x) \leq 2B_n(x) + 2B_0(x)$, so is $\tilde{A}_n(x)$, and again by Vitali’s Theorem, $\int A_n^2 \, dv(\tilde{\theta}) \to 0$ which is (2.7). Continuity (2.2) with regard to Vitali’s Theorem is just continuity of the Fisher information just shown.

The assertion of Remark 2.6 is shown similarly, replacing the $B_n$ and $B_0$ from above with $|\tilde{A}_n^2| (|\tilde{A}_n^2| |x|^2)^2$ resp. $|\tilde{A}_n^2| (|\tilde{A}_n^2| |x|^2)^2$.

### A.3 Proof of Theorem 2.11
For self-containedness, we reproduce the argument for condition (2.12) from Rieder (1994) Thm. 2.3.7). In model $\mathcal{P}$ by (2.1) assuming $v$-densities
\[
|\left( \sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}} \left( 1 + \frac{1}{2} (A_n^2) \right) \right) \sqrt{q_{\theta_0}} |^2 \leq \int |\sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}} \left( 1 + \frac{1}{2} (A_n^2) \right) |^2 \, dv = o(|h|^2)
\]
Hence
\[
E \tilde{A}_n^2 = \int \left( \sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}} \left( 1 + \frac{1}{2} (A_n^2) \right) \right) \sqrt{q_{\theta_0}} \, dv + o(|h|) = \int q_{\theta_0} \sqrt{q_{\theta_0}} \, dv + o(|h|) = -1/2 \int \left( \sqrt{q_{\theta_0}} - \sqrt{q_{\theta_0}} \right) \, dv + o(|h|) = -1/(2h^T \tilde{A}_n \tilde{A}_n \, dv + o(|h|^2) + o(|h|)
\]
So $E \tilde{A}_n^2 = 0$, and hence also $E_n, E_{Q_{\theta_0}(y)}^2 = 0$. Lindeberg condition (2.13) is assumed without change, so it only remains to show condition (2.14). Let $N_{n,t}$ be the $Q_{\theta_0}^{-1}$-null set such that both (H.1) and (H.2) hold for all $y \in N_{n,t}$. Let $N =$
Then as in the case of stochastic regressors, from (H.1) and the chain rule applied pointwise (in $y \in \mathbb{N}$) we obtain (pointwise) existence and form of the $\ell_2$-derivative. Let $\tilde{A}_i$ from (A.4) now take up the dependence on $x_{n,i}$, i.e., $\tilde{A}_i = A_i(x_{n,i})$ (with $s_0$ from the preceding proof substituted by $t_n$) so that in particular, for every fixed $i$, $\tilde{A}_i := \int \tilde{A}_i^2 \nu(dy) \to 0$ as $t_n \to 0$. For condition (2.14) we have to show that $\lim_{n \to \infty} \sup_{|t| \leq 1} \int \tilde{A}_i^2 \nu(dy) = 0$. But, similarly as in the preceding proof for fixed $i$, by the fundamental theorem of calculus for absolutely continuous functions, we have

$$\tilde{A}_i = \int \left( \sqrt{q_{0,i,n}} - \sqrt{q_{0,i,n}} \right)^2 dy \leq \frac{1}{4|t_n|} \int_{|t|} |t_n| L_{n,i} \tilde{A}_i^2 \nu(dy) =: B_{n,i}$$

Now introduce $B_{0,i} = \frac{1}{4|t_n|} \int_{|t|} |t_n| L_{n,i} \tilde{A}_i^2 \nu(dy)$ and note that $\sum_{i=1}^{n} \tilde{A}_i^2 = \tilde{A}_i^2(\tilde{A}_i^2) + \tilde{A}_i^2(\tilde{A}_i^2) = |t|^2 \leq b$ and by (iii)

$$\sum_{i=1}^{n} B_{n,i} = \sum_{i=1}^{n} B_{0,i} + o(1) = |t|^2 / 4 + o(1)$$

Hence by Vitali’s Theorem, $B_{n,i}$ is uniformly integrable (w.r.t. the counting measure), and, as $\tilde{A}_i^2 \leq 2B_{n,i} + 2B_{0,i}$, so is $\tilde{A}_i^2$, and again by Vitali’s Theorem, $\sum_{i=1}^{n} \tilde{A}_i^2 \to 0$. Finally, continuity (2.15) again with regard to Vitali’s Theorem is just continuity of the Fisher information just proven.

Again, the assertion of Remark 2.12 is shown similarly, replacing the $B_{n,i}$ and $B_{0,i}$ from above with corresponding terms involving $|t_{n,i}|^2$, $|\tilde{A}_{n,i}|^2$, and $|\tilde{A}_{n,i}|^2$.

A.4 Link function for GEVD joint shape-scale model

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

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

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.

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

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 $a_2 = e^{2(1-e^{-a_1})^{-1}}$, so $a_2 \approx 1.624$ and $a_1 = 0.5a_2(\log(a_2)) \approx 0.00926$.

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

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

As we mentioned, all terms of the Fisher information matrix for GEVD are dominated by the Gamma term $\Gamma(2x)$. Using the Stirling approximation, i.e., $\Gamma(x) \approx \sqrt{2\pi} x^{x-1/2} e^{-x}$, the double application of the logarithm in the link function we get that $\Gamma(2\ell(x_{-1}^2))$ is approximately $\beta \log(x_{-1})$. By equivariance in $\mu$ and $\sigma$, therefore the integral of condition (ii) turns into:
\[ B_1(\xi) := \frac{4}{\beta \xi} \int \log(x \xi) K(dx) < \infty \quad \text{for } \beta \xi > 0 \quad \text{(A.5)} \]

and
\[ B_2(\xi) := \frac{1}{\beta \xi} \int \frac{\log(x \xi)}{\log(-\beta \xi) + \log(\log(x \xi))} K(dx) \quad \text{for } \beta \xi < 0 \quad \text{(A.6)} \]

Finiteness of (A.5) and (A.6) follow from finiteness of \( E(\min \{1, (\log x)^k\}) \) for \( x \sim \text{GEVD}(0,1,\xi), k \in \mathbb{N} \). Reconsidering (A.5), (A.6) 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 \( \xi \).

Acknowledgement

This article is part of the PhD of Daria Pupashenko. All authors gratefully acknowledge financial support by the Volkswagen Foundation for the project “Robust Risk Estimation”, http://www.mathematik.uni-kl.de/~wwwfm/RobustRiskEstimation.

References

Cox, D.R. (1981): Statistical Analysis of Time Series: Some Recent Developments. Scand. J. Statist. 8: 93–115.
Dahen, H., and Georges D. (2010): Scaling models for the severity and frequency of external operational loss data. Journal of Banking & Finance 34(7): 1484–1496.
Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997): Modelling extremal events for insurance and finance. Springer.
Fahrmeir, L. (1990): Maximum likelihood estimation in misspecified generalized linear models. Statistics 21(4): 487–502.
Fahrmeir, L., and Kaufmann, H. (1985): Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics 13(1) 342–368.
Fahrmeir, L. and Tutz, G. (2001): Multivariate Statistical Modelling Based on Generalized Linear Models. 2nd Edn. Springer.
Gouriéroux, C., Monfort, A., and Trognon, A. (1984): Pseudo maximum likelihood methods: Theory. Econometrica 52(3): 681–700.
McCullagh, P., and Nelder, J.A. (1989): Generalized linear models. Chapman Hall.
Nelder, J.A. and Pregibon, D. (1987): An Extended Quasi-Likelihood Function. Biometrika 74(2): 221–232.
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. The Annals of Statistics 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.
Huber, P. (1981): Robust Statistics. Wiley.
Haberman, S.J. (1970): On the assumptions used to prove asymptotic normality of maximum likelihood estimates. The Annals of Mathematical Statistics 41(3): 802–828.
Nelder, J.A., and Wedderburn, R.W.M. (1972): Generalized linear models. Journal of the Royal Statistical Society. Series A, 370–384.
Witting H. (1985): Mathematische Statistik. Parametrische Verfahren bei festem Stichprobenumfang. Teubner Stuttgart.
Rieder H. (1994): Robust Asymptotic Statistics. Springer. New York.
Rudin, W. (1986): Real and complex analysis. 3rd Edn. McGraw-Hill Inc. New York.
Schlather M. (1994): Glattheit von Generalisierten Linearen Modellen und statistische Folgerungen. Diplomarbeit Universität Bayreuth.