Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Parameter estimation in time series analysis

Vladimir Spokoiny

submitted: March 2nd, 2009

Weierstrass Institute
for Applied Analysis and Stochastics
Mohrenstr. 39
10117 Berlin Germany
E-Mail: spokoiny@wias-berlin.de

No. 1404
Berlin 2009

2000 *Mathematics Subject Classification.* 62F10. Secondary 62J12, 62F25, 62H12.

Key words and phrases. exponential risk bounds, rate function, quasi maximum likelihood autoregression, generalized linear models, quantile regression.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax:	+ 49 30 2044975
E-Mail:	preprint@wias-berlin.de
World Wide Web:	http://www.wias-berlin.de/

Abstract

The paper offers a novel unified approach to studying the accuracy of parameter estimation for a time series. Important features of the approach are: (1) The underlying model is not assumed to be parametric. (2) The imposed conditions on the model are very mild and can be easily checked in specific applications. (3) The considered time series need not to be ergodic or stationary. The approach is equally applicable to ergodic, unit root and explosive cases. (4) The parameter set can be unbounded and non-compact. (5) No conditions on parameter identifiability are required. (6) The established risk bounds are nonasymptotic and valid for large, moderate and small samples. (7) The results describe confidence and concentration sets rather than the accuracy of point estimation. The whole approach can be viewed as complementary to the classical one based on the asymptotic expansion of the log-likelihood. In particular, it claims a consistency of the considered estimate in a rather general sense, which usually is assumed to be fulfilled in the asymptotic analysis. In standard situations under ergodicity conditions, the usual rate results can be easily obtained as corollaries from the established risk bounds. The approach and the results are illustrated on a number of popular time series models including autoregressive, Generalized Linear time series, ARCH and GARCH models and meadian/quantile regression.

1 Introduction

Estimation of parameters of a time series is one of the most popular statistical problems which is included as an important building block in almost any econometric analysis. It is well known that statistical inference for time series is much more involved than the similar problem in the i.i.d. or regression set-up. The established results require quite strong conditions which are rather difficult to check in particular applications; see e.g. Brockwell and Davis (1991), Fan and Yao (2003). The aim of this paper is to offer a rather general and unified approach to measuring the quality in statistical estimation problem for time series which delivers meaningful and informative results under mild assumptions. We focus on the parametric modeling. It is however worth noting that any parametric assumption is only an approximation of reality and it is not precisely fulfilled in many particular situations. One can say that, in long run, any fixed parametric specification is not flexible enough to describe the real structure of the data. The presented approach continues to apply in the cases when the underlying model *does not follow the parametric specification.* In some sense this approach refocuses the statistical paradigm: in many situations it might be useful and reasonable to apply a misspecified model with nice geometric properties rather than trying to precisely mimic the underlying model specifications. Typical examples of this pragmatic procedure are given by least squares, least absolute deviation or quintile regression: all of them can be viewed as quasi maximum likelihood estimates with a specific parametric structure.

One more nice feature of the presented approach is that the model assumptions are very general and non-restrictive and can be easily checked in specific applications. In particular, there is no any identifiability requirements, the results apply even if the parameter of the model is not identifiable. The parameter set can be unbounded and non-compact. Also no conditions like mixing, ergodicity, stationarity etc. are required: the observed time series can be *non-stationary*, *non-ergodic*, *non-mixing*, etc. The approach equally applies to ergodic, unit root and explosive time series. This enables, for instance, to analyze the quality of estimation and testing procedures for the unit root or cointegration analysis in a unified way; cf. Brockwell and Davis (1991), Johansen (1995), Johansen (2002). The required conditions are very mild and can be easily checked in particular applications.

The established risk bounds are *nonasymptotic* can be used for large, moderate and small samples. The results describe nonasymptotic *confidence and concentration sets* rather than the accuracy of point estimation. In the most of examples, the usual consistency and rate results can be easily obtained as corollaries from the established risk bounds.

The obtained exponential bound have been already used in various econometric studies. Spokoiny (2007) offered a local change point volatility estimation method, Čížek et al. (2007) discussed the estimation problem for varying coefficient ARCH and GARCH models, while Giacomini et al. (2007) focused on time varying copulae, Chen and Spokoiny (2007) considered the problem of robust risk management for non-normal and non-stationary market using stagewise aggregation procedure. All these and many other procedures are based on the multiple model check. The crucial issue in practical applications of such methods is the choice of the related parameters like thresholds or critical values in a data-driven way. This choice as well as the related theoretical analysis require to bound from above the probability of a wrong choice which can be done by the results presented below.

The paper is organized as follows. The next section describes the considered time series framework. Particularly, possible violations of the parametric assumption are discussed.

Section 3 presents the main results in form of penalized exponential bounds on the (quasi) maximum likelihood. Section 3.2 demonstrates some implications of the obtained results. We especially focus on the concentration properties of the estimates and on the likelihood based confidence sets. Section 4 illustrates how the general results can be specified for a number of popular time series models like Generalized Linear time series regression, linear autoregression, median and quantile regression. The main result given in Theorem 3.3 is obtained as a specification of general penalized exponential bound for the maximum of a random field from Section 5.

2 Modeling approach

This section describes the considered model and the modeling approach. Let the observed process $Y_t, t = 1, \ldots, n$ be progressively measurable w.r.t. a filtration $\mathcal{F} = (\mathcal{F}_t)$. Typically $\mathcal{F}_t$ stands for the information available at the moment t. One way of describing the joint distribution of the sample $\boldsymbol{Y}$ is by specifying the conditional distribution $Q_t = \mathcal{L}(Y_t \big| \mathcal{F}_{t-1})$ of every observation of Y_t given the "past" $\mathcal{F}_{t-1}$. The parametric approach discussed below allows to reduce the whole description of the model to a few parameters which have to be estimated from the data.

2.1 A parametric model

The parametric time series modeling usually includes two important components, see e.g. Anderson (1994), Brockwell and Davis (1991), Kedem and Fokianos (2002), Fan and Yao (2003). One of them describes the type of conditional distribution Q_t of Y_t given the "past" $\mathcal{F}_{t-1}$ and the other one explains the dynamics of the corresponding parameter. The standard approach assumes that the conditional distribution Q_t belongs to some given parametric family $\mathcal{P} = (P_{\upsilon}, \upsilon \in \mathcal{U})$, but the corresponding parameter υ may change in time and even be a random predictable process $f_t \sim \mathcal{F}_{t-1}$. We write this relation in the form

$$Q_t \overset{\text{def}}{=} \mathcal{L}\bigl(Y_t \big| \mathcal{F}_{t-1}\bigr) = P_{f_t} \in \mathcal{P}. \tag{2.1}$$

The second structural component of the parametric modeling concerns the driving (dynamic) process f_t. Namely, it is assumed that this process is uniquely described by a finite dimensional parameter $\boldsymbol{\theta}_0 \in \Theta \subseteq I\!\!R^p$, that is, $f_t = f_t(\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$. These two assumptions lead to the parametric model in the form

$$Q_t = P_{f_t(\boldsymbol{\theta}_0)}. \tag{2.2}$$

Some typical examples of such parametric specifications are given in Section 2.3 and continued in Section 4.

For estimating $\boldsymbol{\theta}_0$, we apply the quasi maximum likelihood (quasi-MLE) approach. Let the family $\mathcal{P}$ be dominated by a measure P_0. Denote by $\ell(y, \upsilon)$ the corresponding log-density

$$\ell(y, \upsilon) = \log \frac{dP_\upsilon}{dP_0}(y) .$$

The quasi log-likelihood $L(\boldsymbol{Y}, \boldsymbol{\theta})$ for the model (2.1)–(2.2) can be represented in the form

$$L(\boldsymbol{Y}, \boldsymbol{\theta}) = \sum_t \ell\bigl(Y_t, f_t(\boldsymbol{\theta})\bigr).$$

Here in in what follows, $\sum_t$ means summation over the whole time interval $t = 1, \ldots, n$. We define the quasi-MLE estimate $\widetilde{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$ by maximizing the log-likelihood $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{Y}, \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_t \ell\bigl(Y_t, f_t(\boldsymbol{\theta})\bigr), \tag{2.3}$$

and denote by $L(\boldsymbol{Y}, \widetilde{\boldsymbol{\theta}})$ the corresponding maximum.

It is important to stress that the parametric assumption (2.2) is only an approximation of the underlying data distribution $\boldsymbol{P}$ which justifies the estimation procedure (2.3). In reality, the modeling assumption (2.2) can be violated in one or even both parts. One of the aims of our study is to address the questions of what is estimated and with which accuracy if the parametric assumption $Q_t = P_{f_t(\boldsymbol{\theta}_0)}$ is not precisely fulfilled.

2.2 Violation of the parametric assumption

The parametric model (2.2) can be violated by two different reasons. One is due to misspecified conditional distribution and the other one due to a wrong parametric dynamics $f_t = f_t(\boldsymbol{\theta})$.

2.2.1 Misspecified conditional distribution

The model (2.2) assumes that the conditional distribution Q_t of Y_t given $\mathcal{F}_{t-1}$ belongs to the given family $\mathcal{P}$. This assumption can be well justified for many examples from categorical data analysis, for instance, for binary or discrete observations; see Fokianos and Kedem (2003). However, this assumption could be too restrictive for many other applications. We present here a couple of examples of this sort. First consider a stochastic dynamic system described by the equation $Y_t = f_t + \varepsilon_t$ whose drift f_t is a predictable process and innovations ε_t are martingale differences. In the case of conditionally standard normal innovations ε_t, the distribution Q_t is normal with the mean f_t leading to the log-density function $\ell(y, v) = -(y - v)^2/2$ (up to an unimportant constant term $-\frac{1}{2}\log(2\pi)$). Parametric dynamics $f_t = f_t(\boldsymbol{\theta})$ leads to the log-likelihood $L(\boldsymbol{\theta}) = -\sum_t |Y_t - f_t(\boldsymbol{\theta})|^2/2$. In the case of non-normal innovations, this expression is a quasi log-likelihood leading to the least squares solution.

Another typical example is given by the volatility modeling. The log-returns r_t are described by the conditional heteroscedasticity model: $r_t = \sigma_t \varepsilon_t$. The case of standard Gaussian innovations ε_t and the parametric dynamics $f_t = f_t(\boldsymbol{\theta})$ for the volatility $f_t = \sigma_t^2$ leads to the log-likelihood $L(\boldsymbol{\theta}) = -\frac{1}{2}\sum_t \bigl\{ r_t^2/f_t(\boldsymbol{\theta}) + \log(2\pi f_t(\boldsymbol{\theta})) \bigr\}$. In the case of, say, heavy tailed innovations, one can still try to maximize this expression which becomes a quasi log-likelihood.

2.2.2 Misspecified parametric dynamics and the best parametric fit

Suppose for a moment that the conditional distribution Q_t belongs to the given family $\mathcal{P}$ almost surely for all t. Then the data $\boldsymbol{Y}$ can be described by the model $\mathcal{L}\bigl(Y_t \big| \mathcal{F}_{t-1}\bigr) = P_{f_t} \in \mathcal{P}$ for some predictable process f_t. The parametric assumption means that the process f_t belongs to a parametric family of processes $\bigl(f_t(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset I\!\!R^p\bigr)$. This assumption is very useful for the analysis but it is usually only an idealization of reality. An interesting question in this respect is what is estimated in the situation when the process f_t does not follow the parametric dynamics $f_t = f_t(\boldsymbol{\theta})$ whatever $\boldsymbol{\theta}$ is. Below we show that in such cases the quasi log-likelihood approach leads to an estimate of the projection of the given model on the parametric subspace of models. One also speaks about the best parametric fit $f_t(\boldsymbol{\theta})$ to the model f_t which can be defined as a solution of the optimization problem

$$\boldsymbol{\theta}_0 \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \boldsymbol{E} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{K}(\boldsymbol{P}, \boldsymbol{P}_{\boldsymbol{\theta}}),$$

where $\boldsymbol{P} = \prod_t Q_t = \prod_t P_{f_t}$ is the true measure, $\boldsymbol{P}_{\boldsymbol{\theta}} = \prod_t P_{f_t(\boldsymbol{\theta})}$ is its parametric counterpart, and $\mathcal{K}(\boldsymbol{P}, \boldsymbol{P}') \stackrel{\text{def}}{=} \boldsymbol{E} \log(d\boldsymbol{P}/d\boldsymbol{P}')$ is the Kullback-Leibler divergence between two measures $\boldsymbol{P}$ and $\boldsymbol{P}'$.

The interpretation of $\boldsymbol{\theta}_0$ as the “best parametric fit” continues to apply even in the case when also the assumption $Q_t \in \mathcal{P}$ a.s. is violated. However, $\boldsymbol{\theta}_0$ cannot be defined as a minimizer of the Kullback-Leibler divergence anymore.

2.3 Examples

This section presents some popular time series models. Later in Section 4 we illustrate the obtained results on these examples.

2.3.1 Linear autoregression

Let the family $\mathcal{P}$ be a Gaussian shift. This case corresponds to the model $Y_t = f_t + \varepsilon_t$ in which the innovations ε_t are assumed to be i.i.d Gaussian: $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The corresponding (quasi) maximum likelihood approach leads back to the least square estimate $\widetilde{\boldsymbol{\theta}}$: with $L(\boldsymbol{\theta}) = -(2\sigma^2)^{-1} \sum_t \bigl\{ Y_t - f_t(\boldsymbol{\theta}) \bigr\}^2$

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_t \bigl\{ Y_t - f_t(\boldsymbol{\theta}) \bigr\}^2.$$

Linear autoregression means the structural equation $f_t(\boldsymbol{\theta}) = \alpha_1 Y_{t-1} + \ldots + \alpha_p Y_{t-p}$ for $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_p)^\top$ leading to maximization of the quadratic functional $\sum_t \bigl(Y_t - \Psi_t^\top \boldsymbol{\theta} \bigr)^2$ with $\Psi_t = (Y_{t-1}, \ldots, Y_{t-p})^\top$ which admits a closed form solution:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \Bigl(\sum_t \Psi_t \Psi_t^\top \Bigr)^{-1} \sum_t Y_t \Psi_t = B^{-1} \sum_t Y_t \Psi_t \\ L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) &= \frac{1}{2\sigma^2} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top B (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \end{aligned}$$

with $B \overset{\text{def}}{=} \sum_t \Psi_t \Psi_t^\top$. Note, however, that through the closed form solution for $\widetilde{\boldsymbol{\theta}}$ is available, the analysis in time series context remains a difficult task, especially in the non-ergodic case, cf. Dickey and Fuller (1981), Basawa and Brockwell (1984), Chan and Wei (1988), Fountis and Dickey (1989), Cox and Llatas (1991), Koul and Saleh (1993), Phillips and Xu (2006). All the mentioned papers studied the asymptotic properties of the estimates. There are very few papers concerning nonasymptotic results, see Chan and Wei (1988). Fixed accuracy sequential procedures are discussed in Lai and Siegmund (1983), Sriram (1987), Shiryaev and Spokoiny (1997), Konev and Pergamenshchikov (1997). More robust estimates like minimum distance, M- or quantile estimates have been considered in Wang (1986), Chan and Wei (1988).

Our approach based on the exponential bounds for the (quasi) likelihood continues to apply. Moreover, it does not assume that the innovations are independent or conditionally Gaussian. Neither we require that the underlying process f_t follows the linear structural equation $f_t = \Psi_t^\top \boldsymbol{\theta}$.

2.3.2 GARCH(1,1) estimation

GARCH-modeling introduced in Bollerslev (1986) is very popular in analysis of financial time series. A number of GARCH extensions is proposed to make the model even more flexible; for example, EGARCH Nelson (1991), QGARCH Sentana (1995), among many others. We focus on the classical GARCH(1,1) model although most of conclusions can be extended to more general specifications. The underlying modeling assumption is that the observed squared log-returns R_t follows the conditional heteroscedasticity equation:

$$R_t = X_t \varepsilon_t^2,$$

where ε_t are standardized innovations satisfying $\boldsymbol{E}\bigl(\varepsilon_t \big| \mathcal{F}_{t-1}\bigr) = 0$, $\boldsymbol{E}\bigl(\varepsilon_t^2 \big| \mathcal{F}_{t-1}\bigr) = 1$, and X_t is a predictable volatility process. The parametric GARCH(1,1) assumption means that the volatility process X_t follows the equation

$$X_t = \omega + \alpha R_{t-1} + \beta X_{t-1}. \tag{2.4}$$

For simplicity we assume that the initial value X_0 is fixed, e.g. $X_0 = R_0$. Then for every $\boldsymbol{\theta} = (\omega, \alpha, \beta)^\top$ we can recursively apply the structural equation (2.4) yielding the process $X_t(\boldsymbol{\theta})$ with

$$X_t(\boldsymbol{\theta}) = \omega + \alpha R_{t-1} + \beta X_{t-1}(\boldsymbol{\theta}), \qquad t \ge 1, \qquad X_0(\boldsymbol{\theta}) = X_0.$$

With this process we associate a (quasi) log likelihood

$$L(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{t=1}^{n} \bigl\{ \log(2\pi X_t(\boldsymbol{\theta})) + R_t / X_t(\boldsymbol{\theta}) \bigr\}.$$

This expression becomes the log-likelihood if the innovations ε_t are conditionally on $\mathcal{F}_{t-1}$ standard normal and the structural equation (2.4) is fulfilled for some combination of parameters. Asymptotic properties of such estimates are well studied, see e.g. Lee and Hansen (1994), Fan and Yao (2003), Sun and Stengos (2006), Francq and Zakoian (2007), and references therein.

2.3.3 Median and quantile time series estimation

Median or more generally quantile estimation is known to be more robust and stable against outliers and it is frequently used in econometric studies; see Koenker (2005), Koenker and Xiao (2006). The corresponding approach explains the observations Y_t by the regression equation $Y_t = f_t + \varepsilon_t$ where the individual errors ε_t are not assumed to fulfill $\boldsymbol{E}(\varepsilon_t \big| \mathcal{F}_{t-1}) = 0$. Instead, one imposes the constraint $\boldsymbol{P}(\varepsilon_t > 0 \big| \mathcal{F}_{t-1}) = \alpha$ for a given α. The median regression corresponds to $\alpha = 1/2$. Under the parametric assumption $f_t = f_t(\boldsymbol{\theta}_0)$ for a given parametric class of predictable processes $f_t(\boldsymbol{\theta})$, the corresponding estimate can be defined by maximizing the quasi log-likelihood

$$L(\boldsymbol{\theta}) = \sum_t \ell_\alpha\bigl(Y_t - f_t(\boldsymbol{\theta})\bigr)$$

with $\ell_\alpha(x) = (1-\alpha)x_- - \alpha x_+$.

In this case, $\mathcal{P}$ is the family with the log-density $\ell(y, \upsilon) = \ell_\alpha(y - \upsilon)$. In particular, the median regression for $\alpha = 1/2$ corresponds to the Laplacian shift family.

2.3.4 Categorical time series

Let $\mathcal{P}$ be an exponential family with the canonical parametrization (EFC) which means that the corresponding log-likelihood function can be written in the form

$$\ell(y, \upsilon) = y\upsilon - d(\upsilon) + \ell(y)$$

where $d(\cdot)$ is a given convex function; see e.g. McCullagh and Nelder (1989), Green and Silverman (1994). The term $\ell(y)$ is unimportant and it cancels in the log-likelihood ratio. Such families are often used in the categorical time series analysis for describing the conditional distribution Q_t of the observed data; see Fokianos and Kedem (2003). The corresponding model can be written as

$$\mathcal{L}\bigl(Y_t \big| \mathcal{F}_{t-1}\bigr) = P_{f_t} \in \mathcal{P}. \tag{2.5}$$

Parametric modeling assumes a specific structure of the driving process f_t leading to the parametric log-likelihood function $L(\boldsymbol{\theta}) = \sum_t \ell(Y_t, f_t(\boldsymbol{\theta}))$:

$$L(\boldsymbol{\theta}) = \sum_t \ell(Y_t, f_t(\boldsymbol{\theta})) = \sum_t \bigl\{Y_t f_t(\boldsymbol{\theta}) - d\bigl(f_t(\boldsymbol{\theta})\bigr)\bigr\}. \tag{2.6}$$

Usually $f_t(\boldsymbol{\theta})$ can be represented in the form $f_t(\boldsymbol{\theta}) = m(X_t, \boldsymbol{\theta})$ for some *regression function* $m(\cdot,\cdot)$, an explanatory process X_t and a parameter vector $\boldsymbol{\theta} \in \Theta \subset I\!R^p$. One popular example of linear regression is discussed in the next section.

Our approach allows to account for the both kinds of model misspecification: $Q_t \notin \mathcal{P}$ and/or $f_t \neq f_t(\boldsymbol{\theta})$. However, to be more specific, we consider below the case when $Q_t \in \mathcal{P}$. Then the focus of analysis is the best parametric approximation of the true regression function f_t by a parametric model $f_t(\boldsymbol{\theta})$.

2.3.5 Generalized Linear time series

Let $\mathcal{P}$ be again an EFC. A parametric *generalized linear* specification for the model (2.5) is given by the following set of structural equations:

$$\mathcal{L}\bigl(Y_t \big| \mathcal{F}_{t-1}\bigr) = P_{f_t} \in \mathcal{P}, \qquad f_t = g(\Psi_t), \qquad \Psi_t = A(\boldsymbol{\theta})\Psi_{t-1} \tag{2.7}$$

where Ψ_t is a predictable R^d-dimensional explanatory process, $g(\cdot)$ is a given mapping from $I\!R^d$ to $I\!R$, $A(\boldsymbol{\theta})$ is a given $d \times d$-matrix linearly depending on the parameter vector $\boldsymbol{\theta} \in \Theta \subset I\!R^p$. Such models are widely used in statistical modeling. A popular example is given by the equations $f_t = g(X_t)$ and

$$X_t = \omega + \alpha_1 Y_{t-1} + \ldots \alpha_p Y_{t-p} + \beta_1 X_{t-1} + \ldots \beta_q X_{t-q}. \tag{2.8}$$

Here $\Psi_t = (X_t, \ldots, X_{t-q+1}, Y_t, \ldots, Y_{t-p+1}, 1)^\top$, $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_q, \alpha_1, \ldots, \alpha_p, \omega)^\top$, and the first row of $A(\boldsymbol{\theta})$ is just $\boldsymbol{\theta}$. For $\beta_1 = \ldots = \beta_q = 0$ the value X_t is a linear combination of the past observations and (2.7) becomes an autoregressive type model. If there is at least one coefficient $\beta_j \neq 0$, then X_t is an unobservable (hidden/latent/exogeneous) component and (2.7) is of ARMA type; see e.g. Fokianos and Kedem (2003).

Let $\boldsymbol{\theta}$ be the parameter vector. Then, given $\boldsymbol{\theta}$, the observations $Y_1, \ldots, Y_t$, and the pre-history Ψ^0, Y^0, one can uniquely reconstruct the process $\Psi_t = \Psi_t(\boldsymbol{\theta})$ and then $f_t(\boldsymbol{\theta}) = g(\Psi_t(\boldsymbol{\theta}))$ for $t \geq 1$ by recurrently applying the relation $\Psi_t = A(\boldsymbol{\theta})\Psi_{t-1}$. The process $f_t(\boldsymbol{\theta})$ leads to the (quasi) log-likelihood $L(\boldsymbol{\theta}) = \sum_t \bigl\{Y_t f_t(\boldsymbol{\theta}) - d\bigl(f_t(\boldsymbol{\theta})\bigr)\bigr\}$. Inference for GLM's has been discussed in many papers and books. We only mention Green and Silverman (1994), Chen (1995), Chen et al. (1999), Sun et al. (2000), Fokianos and Kedem (2003), Kedem and Fokianos (2002), Fan and Yao (2003), among many others. Our analysis further in Section 4 essentially differs from all the mentioned studies. In particular, it does not assume that any of imposed parametric specifications from (2.7) is really fulfilled. The methods and results are non-asymptotic.

In all the examples, the true model is still given by (2.1). By $\boldsymbol{\theta}_0$ we denote the parameter corresponding to the best parametric fit: $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta}} \boldsymbol{E} L(\boldsymbol{\theta})$. The parameter $\boldsymbol{\theta}_0$ is estimated by maximizing the objective function $L(\boldsymbol{\theta})$.

3 Exponential risk bounds

In this section we first introduce the basic notions and conditions on the model and then state the main results in form of general exponential bounds for the supremum of the quasi log-likelihood function $L(\boldsymbol{\theta})$. The quality of estimation of $\boldsymbol{\theta}$ is measured in terms of the maximum $L(\widetilde{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$ rather than the point of maximum $\widetilde{\boldsymbol{\theta}}$, where $L(\boldsymbol{\theta})$ from (2.5). More precisely, we define the point

$$\boldsymbol{\theta}_0 \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \boldsymbol{E} L(\boldsymbol{\theta})$$

which is the true value in the parametric situation and can be viewed as the parameter of the best parametric fit in the general case. Now the aim of our study is to establish some exponential bounds on the supremum in $\boldsymbol{\theta}$ of the random field

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}_0) = \sum_t \bigl\{ \ell(Y_t, f_t(\boldsymbol{\theta})) - \ell(Y_t, f_t(\boldsymbol{\theta}_0)) \bigr\}.$$

Later in Section 3.2 we comment how the accuracy of estimation of $\boldsymbol{\theta}_0$ by $\widetilde{\boldsymbol{\theta}}$ relates to the value $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$. We will also see that the bound for $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ yields the confidence sets for the parameter $\boldsymbol{\theta}_0$ and concentration sets for the estimate $\widetilde{\boldsymbol{\theta}}$.

Define for $\boldsymbol{\theta} \in \Theta$

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} \sum_t \mathfrak{m}_t\bigl(\mu, f_t(\boldsymbol{\theta}), f_t(\boldsymbol{\theta}_0)\bigr). \tag{3.1}$$

where for $\upsilon, \upsilon' \in \mathfrak{U}$

$$\mathfrak{m}_t(\mu, \upsilon, \upsilon') \stackrel{\mathrm{def}}{=} -\log \boldsymbol{E}\bigl[\exp\bigl\{\mu \ell(Y_t, \upsilon) - \mu \ell(Y_t, \upsilon')\bigr\} \big| \mathcal{F}_{t-1}\bigr].$$

This definition assumes the following condition:

$(\boldsymbol{E})$ *There exists some $\mu > 0$ such that for all $\boldsymbol{\theta} \in \Theta$ and all t the value $\mathfrak{m}_t\bigl(\mu, f_t(\boldsymbol{\theta}), f_t(\boldsymbol{\theta}_0)\bigr)$ is finite.*

Note that this condition is automatically fulfilled with $\mu \le 1$ if $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{\theta}_0}$ and $L(\boldsymbol{\theta})$ is indeed a log-likelihood function.

The main observation behind the definition (3.1) is that

$$\boldsymbol{E} \exp\bigl\{\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)\bigr\} = 1.$$

Our main goal is to get an exponential bound for the maximum of the random field $\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ over $\boldsymbol{\theta} \in \Theta$. Unfortunately, this maximum may explode and we consider the penalized expression $\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta})$ where the penalty function $\operatorname{pen}(\boldsymbol{\theta})$ should provide some bounded exponential moments for

$$\sup_{\boldsymbol{\theta} \in \Theta} \bigl[\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta})\bigr].$$

More precisely, we present a penalty function $\operatorname{pen}(\boldsymbol{\theta})$ that ensures under rather general conditions for every $\varrho < 1$ that the value

$$\mathfrak{Q}(\varrho) \overset{\text{def}}{=} \boldsymbol{E} \sup_{\boldsymbol{\theta} \in \Theta} \exp\bigl\{ \varrho \bigl[\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta}) \bigr] \bigr\}$$

is bounded by a fixed constant.

We consider the following decomposition of the log-likelihood process $L(\boldsymbol{\theta})$ into the martingale-difference and predictable parts:

$$L(\boldsymbol{\theta}) \overset{\text{def}}{=} M(\boldsymbol{\theta}) + \zeta(\boldsymbol{\theta})$$

where

$$\begin{aligned} M(\boldsymbol{\theta}) &\overset{\text{def}}{=} \sum_t m_t\bigl(f_t(\boldsymbol{\theta})\bigr), \\ \zeta(\boldsymbol{\theta}) &\overset{\text{def}}{=} \sum_t \zeta_t\bigl(Y_t, f_t(\boldsymbol{\theta})\bigr) \end{aligned}$$

with $m_t(\upsilon) \overset{\text{def}}{=} \boldsymbol{E}\bigl[\ell(Y_t, \upsilon) \big| \mathcal{F}_{t-1}\bigr]$, $\zeta_t(Y_t, \upsilon) \overset{\text{def}}{=} \ell(Y_t, \upsilon) - m_t(\upsilon)$ for $\upsilon \in \mathcal{U}$. Below we assume that the (random) functions $m_t(\upsilon)$ and $\zeta_t(\upsilon)$ are differentiable w.r.t. υ and denote $\dot{m}_t(\upsilon) = dm_t(\upsilon)/d\upsilon$ and $\dot{\zeta}_t(y, \upsilon) = d\zeta_t(y, \upsilon)/d\upsilon$.

Suppose also that the random function $f_t(\boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$ and denote $\nabla f_t(\boldsymbol{\theta}) = \partial f_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \in I\!R^p$. Define

$$\begin{aligned} \nabla M(\boldsymbol{\theta}) &\overset{\text{def}}{=} \sum_t \dot{m}_t\bigl(f_t(\boldsymbol{\theta})\bigr) \nabla f_t(\boldsymbol{\theta}) \\ \nabla \zeta(\boldsymbol{\theta}) &\overset{\text{def}}{=} \sum_t \dot{\zeta}_t\bigl(Y_t, f_t(\boldsymbol{\theta})\bigr) \nabla f_t(\boldsymbol{\theta}). \end{aligned}$$

Condition (E) assumes that the quasi log-likelihood has bounded exponential moments. We also assume a similar property for its gradient.

$(\boldsymbol{ED})$ *There exist some deterministic symmetric matrix $V(\boldsymbol{\theta})$ and a constant $\lambda^* > 0$ such that for all $\lambda \le \lambda^*$*

$$\sup_{\boldsymbol{\gamma} \in \mathcal{S}^p} \sup_{\boldsymbol{\theta} \in \Theta} \log \boldsymbol{E} \exp\biggl\{ 2\lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\theta})}{\sqrt{\boldsymbol{\gamma}^\top V(\boldsymbol{\theta}) \boldsymbol{\gamma}}} \biggr\} \le 2\lambda^2, \tag{3.2}$$

and

$$\sup_{\boldsymbol{\gamma} \in \mathcal{S}^p} \sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{E} \exp\biggl\{ 2\lambda \frac{\boldsymbol{\gamma}^\top [\nabla M(\boldsymbol{\theta}) - \boldsymbol{E} \nabla M(\boldsymbol{\theta})]}{\sqrt{\boldsymbol{\gamma}^\top V(\boldsymbol{\theta}) \boldsymbol{\gamma}}} \biggr\} \le 2\lambda^2. \tag{3.3}$$

This condition is usually simple to check. Below we present some simple sufficient conditions for (3.2) and (3.3).

Lemma 3.1. *Suppose that there exist a constant* $\lambda_1^* > 0$ *and a random function* $\mathfrak{n}_t(\upsilon) \sim \mathcal{F}_{t-1}$ *such that for all* t *and* $\lambda \le \lambda_1^*$

$$\log \boldsymbol{E}\Bigl[\exp\Bigl\{2\lambda \frac{\dot{\zeta}_t(Y_t, \upsilon)}{\mathfrak{n}_t(\upsilon)}\Bigr\} \Bigm| \mathcal{F}_{t-1}\Bigr] \le 2\lambda^2, \qquad \upsilon \in \mathcal{U}. \tag{3.4}$$

Let also there exist a deterministic matrix function $V(\boldsymbol{\theta}) \ge I$ *and the value* $\lambda^* \ge \lambda_1^*$ *such that it holds almost surely for any* $\boldsymbol{\theta} \in \Theta$ *and* $\boldsymbol{\gamma} \in \mathcal{S}^p$

$$B(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_t \mathfrak{n}_t^2(f_t(\boldsymbol{\theta}))\, \nabla f_t(\boldsymbol{\theta}) \nabla f_t(\boldsymbol{\theta})^\top \le V(\boldsymbol{\theta}),$$

$$\mathfrak{n}_t(f_t(\boldsymbol{\theta}))\, \bigl|\boldsymbol{\gamma}^\top \nabla f_t(\boldsymbol{\theta})\bigr| \le \frac{\lambda_1^*}{\lambda^*} \sqrt{\boldsymbol{\gamma}^\top V(\boldsymbol{\theta}) \boldsymbol{\gamma}}. \tag{3.5}$$

Then (3.2) is fulfilled with this $V(\boldsymbol{\theta})$ *for* $\lambda \le \lambda^*$ *.*

Proof. By definition

$$2\lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta(\boldsymbol{\theta})}{\sqrt{\boldsymbol{\gamma}^\top V(\boldsymbol{\theta}) \boldsymbol{\gamma}}} - 2\lambda^2 \frac{\boldsymbol{\gamma}^\top B(\boldsymbol{\theta}) \boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top V(\boldsymbol{\theta}) \boldsymbol{\gamma}} = \sum_t \Bigl\{ 2\lambda c_t \frac{\dot{\zeta}_t(Y_t, f_t(\boldsymbol{\theta}))}{\mathfrak{n}_t(f_t(\boldsymbol{\theta}))} - 2\lambda^2 c_t^2 \Bigr\}$$

where $c_t = \mathfrak{n}_t(f_t(\boldsymbol{\theta}))\, \boldsymbol{\gamma}^\top \nabla f_t(\boldsymbol{\theta}) \bigl[\boldsymbol{\gamma}^\top V(\boldsymbol{\theta}) \boldsymbol{\gamma}\bigr]^{-1/2}$ so that $\lambda \le \lambda^*$ implies $\lambda c_t \le \lambda_1^*$ in view of (3.5). Now by (3.4)

$$\boldsymbol{E}\Bigl[\exp\Bigl\{2\lambda c_t \frac{\dot{\zeta}_t(Y_t, f_t(\boldsymbol{\theta}))}{\mathfrak{n}_t(f_t(\boldsymbol{\theta}))} - 2\lambda^2 c_t^2\Bigr\} \Bigm| \mathcal{F}_{t-1}\Bigr] \le 1$$

and the result follows by induction arguments starting from $t = n$. □

Lemma 3.2. *Let for some* $\overline{\lambda} > 0$ *, the function*

$$\boldsymbol{E} \exp\bigl\{2\overline{\lambda} \boldsymbol{\gamma}^\top [\nabla M(\boldsymbol{\theta}) - \boldsymbol{E} \nabla M(\boldsymbol{\theta})]\bigr\}$$

be uniformly continuous in $(\boldsymbol{\theta}, \boldsymbol{\gamma}) \in \Theta \times \mathcal{S}^p$ *. Let also a matrix* $V_0(\boldsymbol{\theta})$ *be uniformly continuous* $\boldsymbol{\theta} \in \Theta$ *and satisfy* $V_0(\boldsymbol{\theta}) \ge I$ *and*

$$\operatorname{Var}\bigl[\nabla M(\boldsymbol{\theta})\bigr] \le V_0(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta.$$

Then for every $\lambda^* < \overline{\lambda}$ *there exists a constant* $C_1 = C_1(\lambda^*, \overline{\lambda})$ *such that (3.3) is fulfilled with* $V(\boldsymbol{\theta}) = C_1 V_0(\boldsymbol{\theta})$ *.*

The result is an easy corollary of Lemma 5.8 from Golubev and Spokoiny (2009).

Define for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $u = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ and $\boldsymbol{\gamma} = (\boldsymbol{\theta}' - \boldsymbol{\theta})/u$

$$\mathfrak{S}^2(\boldsymbol{\theta}, \boldsymbol{\theta}') \stackrel{\text{def}}{=} u^2 \int_0^1 \boldsymbol{\gamma}^\top V(\boldsymbol{\theta} + tu\boldsymbol{\gamma}) \boldsymbol{\gamma}\, dt.$$

Next, introduce for every $\boldsymbol{\theta}^\circ \in \Theta$ the local vicinity $\mathcal{B}(\epsilon, \boldsymbol{\theta}^\circ)$ such that $\mathfrak{S}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \le \epsilon$ for all $\boldsymbol{\theta} \in \mathcal{B}(\epsilon, \boldsymbol{\theta}^\circ)$.

Let also the matrix function $V(\cdot)$ satisfy the following regularity condition:

$(\boldsymbol{V})$ *There exist constants* $\epsilon > 0$ *and* $\nu_1 < 1$ *such that*

$$\sup_{\boldsymbol{\theta},\boldsymbol{\theta}^{\circ} \in \Theta \colon \mathfrak{S}(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}) \le \epsilon} \, \sup_{\boldsymbol{\gamma} \in S^p} \frac{\boldsymbol{\gamma}^{\top} V(\boldsymbol{\theta}) \boldsymbol{\gamma}}{\boldsymbol{\gamma}^{\top} V(\boldsymbol{\theta}^{\circ}) \boldsymbol{\gamma}} \le \nu_1 \, .$$

The next result presents the claimed exponential bound. It is a specification of a more general result from Theorem 5.5 in Section 5.

Theorem 3.3. *Assume* (E), (ED) *with some* $\lambda^* > 0$ *and* (V) *with some* ν_1 *and* ϵ. *Let* $\varrho < 1$ *be such that* $\varrho\epsilon/(1-\varrho) \le \lambda^*$. *If the function* $\operatorname{pen}(\boldsymbol{\theta})$ *fulfills*

$$\mathfrak{H}_{\epsilon}(\varrho) \stackrel{\text{def}}{=} \log\Bigl\{ \omega_p^{-1} \epsilon^{-p} \int_{\Theta} \sqrt{\det(V(\boldsymbol{\theta}))} \exp\bigl\{-\varrho \operatorname{pen}_{\epsilon}(\boldsymbol{\theta})\bigr\} d\boldsymbol{\theta} \Bigr\} < \infty \tag{3.6}$$

with $\operatorname{pen}_{\epsilon}(\boldsymbol{\theta}^{\circ}) = \inf_{\boldsymbol{\theta} \in \mathcal{B}(\epsilon,\boldsymbol{\theta}^{\circ})} \operatorname{pen}(\boldsymbol{\theta})$ *and* ω_p *being the volume of the unit ball in* $I\!\!R^p$, *then*

$$\boldsymbol{E} \exp\Bigl\{ \sup_{\boldsymbol{\theta} \in \Theta} \varrho \bigl[\mu L(\boldsymbol{\theta},\boldsymbol{\theta}_0) + \mathfrak{M}(\mu,\boldsymbol{\theta},\boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta}) \bigr] \Bigr\} \le \mathfrak{Q}(\varrho), \tag{3.7}$$

with

$$\log \mathfrak{Q}(\varrho) = \frac{2\epsilon^2 \varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}_p + \mathfrak{H}_{\epsilon}(\varrho) + p \log(\nu_1)$$

where $\mathbb{Q}_p$ *is the usual entropy number for the Euclidean ball in* $I\!\!R^p$.

3.1 Penalty via the norm $\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|$

The choice of the penalty function $\operatorname{pen}(\boldsymbol{\theta})$ can be made more precise if the condition (ED) can be checked with a constant matrix $V(\boldsymbol{\theta}) \equiv V^*$ for a fixed matrix V^* and all $\boldsymbol{\theta}$. This section describes how the penalty function can be defined in terms of the norm $\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|$.

Theorem 3.4. *Let the conditions* (E) *and* (ED) *be fulfilled with* $V(\boldsymbol{\theta}) \equiv V^*$ *for some matrix* V^* *for all* $\boldsymbol{\theta} \in \Theta$. *Let* $\varrho \in (0,1)$ *and* $\epsilon > 0$ *be fixed to ensure* $\varrho\epsilon/(1-\varrho) \le \lambda^*$. *Suppose that* $\varkappa(\mathfrak{r})$ *is a monotonously decreasing positive function on* $[0,+\infty)$ *satisfying*

$$\mathfrak{P}^* \stackrel{\text{def}}{=} \omega_p^{-1} \int_{I\!\!R^p} \varkappa(\|\boldsymbol{\theta}\|) d\boldsymbol{\theta} = p \int_0^{\infty} \varkappa(t) t^{p-1} dt < \infty. \tag{3.8}$$

Define

$$\operatorname{pen}(\boldsymbol{\theta}) = -\varrho^{-1} \log \varkappa\bigl(\epsilon^{-1} \bigl\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\bigr\| + 1\bigr). \tag{3.9}$$

Then the assertion (3.7) holds with

$$\log \mathfrak{Q}(\varrho) = \frac{2\epsilon^2 \varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}_p + \log(\mathfrak{P}^*).$$

Proof. This result is a straightforward corollary of Theorem 3.3 applied with $V(\boldsymbol{\theta}) \equiv V^*$ and thus, condition (V) is fulfilled with $\nu_1 = 1$. □

Here two natural ways of defining the penalty function $\operatorname{pen}(\boldsymbol{\theta})$: quadratic or logarithmic in $\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|$. The functions $\varkappa(\cdot)$ and the corresponding $\mathfrak{P}^*$-values are:

$$\begin{array}{rclcrcl} \varkappa_1(u) & = & e^{-\delta_1 (t-1)_+^2}, & & \mathfrak{P}_1^* & = & 1 + \omega_p^{-1}(\pi/\delta_1)^{p/2}, \\ \varkappa_2(t) & = & (t+1)^{-p-\delta_2}, & & \mathfrak{P}_2^* & = & p/\delta_2\,, \end{array} \tag{3.10}$$

where $\delta_1, \delta_2 > 0$ are some constant and $[a]_+$ means $\max\{a, 0\}$. The corresponding penalties read as:

$$\begin{array}{rcl} \operatorname{pen}_1(\boldsymbol{\theta}) & = & \varrho^{-1}\delta_1\,\epsilon^{-2}\bigl\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\bigr\|^2. \\ \operatorname{pen}_2(\boldsymbol{\theta}) & = & -\varrho^{-1}(p+\delta_2)\log\bigl(\epsilon^{-1}\bigl\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\bigr\| + 2\bigr). \end{array}$$

3.2 Some corollaries

The result of Theorem 3.3 means that the value $\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta})$ is uniformly in $\boldsymbol{\theta} \in \Theta$ stochastically bounded. In particular, one can plug the estimate $\widetilde{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$: with some $\varrho < 1$

$$\boldsymbol{E}\exp\Bigl\{\varrho\bigl[\mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}})\bigr]\Bigr\} \le \mathfrak{Q}(\varrho). \tag{3.11}$$

Below we present some corollaries of this result.

To simplify the presentation, we consider the case when there is a deterministic function $\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ such that the following bound holds almost sure:

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0), \qquad \boldsymbol{\theta} \in \Theta \tag{3.12}$$

3.2.1 Concentration properties of the estimator $\widetilde{\boldsymbol{\theta}}$

Define for every subset A of the parameter set Θ the value

$$\mathfrak{z}(A) \stackrel{\text{def}}{=} \inf_{\boldsymbol{\theta} \notin A}\{\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta})\}. \tag{3.13}$$

The next result shows that the estimator $\widetilde{\boldsymbol{\theta}}$ deviates out of the set A with an exponentially small probability of order $\exp\{-\varrho\mathfrak{z}(A)\}$.

Corollary 3.5. *Suppose (3.11). Then for any set* $A \subset \Theta$

$$\boldsymbol{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin A\bigr) \le \mathfrak{Q}(\varrho)\mathrm{e}^{-\varrho\mathfrak{z}(A)}.$$

Proof. If $\widetilde{\boldsymbol{\theta}} \notin A$, then $\mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}}) \geq \mathfrak{z}(A)$. As $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \geq 0$, it follows

$$\begin{aligned} \mathfrak{Q}(\varrho) &\geq \boldsymbol{E} \exp\Bigl\{ \varrho \bigl[\mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}}) \bigr] \Bigr\} \\ &\geq \boldsymbol{E} \exp\Bigl\{ \varrho \bigl[\mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}}) \bigr] \Bigr\} \geq \mathrm{e}^{\mathfrak{z}(A)} \boldsymbol{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin A\bigr) \end{aligned}$$

as required. □

Two particular choices of the set A can be mentioned:

$$\begin{aligned} A &= \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \leq \mathfrak{r}\}, \\ A &= \mathcal{A}'(\mathfrak{r}, \boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta}) \leq \mathfrak{r}\}, \end{aligned}$$

For the set $\mathcal{A}'(\mathfrak{r}, \boldsymbol{\theta}_0)$, Corollary 3.5 yields

$$\boldsymbol{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}'(\mathfrak{r}, \boldsymbol{\theta}_0)\bigr) = \boldsymbol{P}\bigl(\mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}}) \geq \mathfrak{r}\bigr) \leq \mathfrak{Q}(\varrho) \mathrm{e}^{-\varrho \mathfrak{r}}.$$

For the set $\mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)$, define additionally the value $\mathfrak{b}(\mathfrak{r})$ by the relation

$$\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta}) \geq \mathfrak{r} - \mathfrak{b}(\mathfrak{r}), \qquad \boldsymbol{\theta} \in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0),$$

or, equivalently,

$$\mathfrak{b}(\mathfrak{r}) = \sup_{\boldsymbol{\theta} \in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)} \bigl\{ \mathfrak{r} + \operatorname{pen}(\boldsymbol{\theta}) - \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \bigr\}.$$

Corollary 3.6. *Suppose (3.11). Then for any* $\mathfrak{r} > 0$

$$\boldsymbol{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)\bigr) = \boldsymbol{P}\bigl(\overline{\mathfrak{M}}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \geq \mathfrak{r}\bigr) \leq \mathfrak{Q}(\varrho) \mathrm{e}^{-\varrho[\mathfrak{r} - \mathfrak{b}(\mathfrak{r})]}.$$

In typical situations the value $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ and thus, $\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is nearly proportional to the sample size n and is nearly quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ so that and each set $\mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)$ corresponds to a root-n neighborhood of the point $\boldsymbol{\theta}_0$, and the concentration property becomes a non-asymptotic analog of root-n consistency. See below Section 3.4 for a precise formulation.

It is important to stress that for applying the result of Corollary 3.6, it is not required to compute the rate function $\mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ and the penalty function $\operatorname{pen}(\boldsymbol{\theta})$. It only suffices to obtain some rough upper bound for the penalty function and deterministic lower bound for the rate function. The result claims that the estimate well localizes on a vicinity $\mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)$ of the point $\boldsymbol{\theta}_0$.

Another remark concerns the identifiability issue. It was already mentioned in the introduction that the results do not require any identifiability condition. However, if the model parameter is not well identifiable this leads to the situation that the rate function

$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is very flat and its level sets are quite big. Therefore, a poor parametrization leads to a less informative concentration property. In particular, the set $\mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)$ can be unbounded or disconnected.

The concentration property is very useful in combination with a more fine analysis based on the Taylor expansion of the (quasi) log-likelihood. Indeed, it ensures that the estimate belongs with a high probability to a small vicinity of $\boldsymbol{\theta}_0$ and in this vicinity the classical asymptotic technique based on the second order approximation of the process $L(\boldsymbol{\theta})$ can be used to address the issues of asymptotic distribution and asymptotic efficiency.

3.2.2 Confidence sets based on $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$

Next we discuss how the exponential bound can be used for establishing some risk bounds and for constructing the confidence sets for the target $\boldsymbol{\theta}_0$ based on the maximized value $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$. The inequality (3.11) claims that $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ is stochastically bounded with finite exponential moments.

Define

$$\mathfrak{b} \overset{\text{def}}{=} \mathfrak{b}(0) = \sup_{\boldsymbol{\theta}} [\operatorname{pen}(\boldsymbol{\theta}) - \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)]_+ . \tag{3.14}$$

Corollary 3.7. *Suppose (3.11) and let* $\mathfrak{b}$ *from (3.14) be finite. Then*

$$\boldsymbol{E} \exp\bigl\{ \varrho \mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \bigr\} \le e^{\varrho \mathfrak{b}} \mathfrak{Q}(\varrho).$$

Proof. Observe that

$$\boldsymbol{E} \exp\bigl\{ \varrho \mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \bigr\} \le e^{\varrho \mathfrak{b}} \boldsymbol{E} \exp\Bigl\{ \varrho \bigl[\mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}}) \bigr] \Bigr\} \le e^{\varrho \mathfrak{b}} \mathfrak{Q}(\varrho).$$

This obviously yields the assertion. □

By the same reasons, one can construct confidence sets based on the (quasi) likelihood process. Define

$$\mathcal{E}(\mathfrak{z}) = \bigl\{ \boldsymbol{\theta} \in \Theta : L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z} \bigr\}.$$

The bound for $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$ ensures that $\boldsymbol{\theta}_0$ belongs to this set with a high probability provided that $\mathfrak{z}$ is large enough. The next result claims that $\mathcal{E}(\mathfrak{z})$ does not cover the true value $\boldsymbol{\theta}_0$ with a probability which decreases exponentially with $\mathfrak{z}$.

Corollary 3.8. *Suppose (3.11). For any* $\mathfrak{z} > 0$

$$\boldsymbol{P}\bigl(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z}) \bigr) \le \mathfrak{Q}(\varrho) \exp\bigl\{ -\varrho \mu \mathfrak{z} + \varrho \mathfrak{b} \bigr\}.$$

Proof. The bound (3.11) implies for the event $\{\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\} = \{L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \mathfrak{z}\}$

$$\begin{aligned} \boldsymbol{P}\bigl\{\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\bigr\} &\leq \boldsymbol{P}\bigl\{\varrho\bigl[\mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}})\bigr] > \varrho\mu\mathfrak{z} - \varrho\mathfrak{b}\bigr\} \\ &\leq \exp\bigl\{-\varrho\mu\mathfrak{z} + \varrho\mathfrak{b}\bigr\}\boldsymbol{E}\exp\bigl\{\varrho\mu L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) - \operatorname{pen}(\widetilde{\boldsymbol{\theta}})\bigr\} \\ &\leq \mathfrak{Q}(\varrho)\exp\bigl\{-\varrho\mu\mathfrak{z} + \varrho\mathfrak{b}\bigr\} \end{aligned}$$

as required. □

The result of Corollary 3.8 only presents an upper bound for the coverage probability of the value $\boldsymbol{\theta}_0$ by the set $\mathcal{E}(\mathfrak{z})$. The given exponential bound contains some implicit constant and is rather rough, and therefore, it can hardly be used for computing the coverage probability and for fixing the constant $\mathfrak{z}_\alpha$ which ensures the coverage level $1 - \alpha$. It would be unrealistic to obtain a universal non-asymptotic sharp bound for the coverage level which applies in such a general situation. However, the result is meaningful because it suggests the form of the confidence set and guarantees that the choice a sufficiently big but fixed threshold $\mathfrak{z}$ ensures the prescribed coverage probability. A precise value can be found by the Monte-Carlo simulations, see e.g. Spokoiny (2007) for some examples.

3.3 Identifiability condition

Until this point no any identifiability condition on the model has been used, that is, the presented results apply even for a very poor parametrization. Actually, a particular parametrization of the parameter set plays no role as long as the value of maximum is considered. If we want to derive any quantitative result on the point of maximum $\widetilde{\boldsymbol{\theta}}$, then the parametrization matters and an identifiability condition is really necessary. Here we follow the usual path by applying the quadratic lower bound for the rate function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ in a vicinity of the point $\boldsymbol{\theta}_0$.

Finally we discuss a risk bound in terms of the classical loss $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. The idea is to apply the quadratic lower bound for the rate function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ in a vicinity of the point $\boldsymbol{\theta}_0$ and to use the concentration property of the estimator $\widetilde{\boldsymbol{\theta}}$. To explain the conditions imposed below suppose that the log-likelihood function is two times continuously differentiable in $\boldsymbol{\theta}$. This implies the differentiability in $\boldsymbol{\theta}$ of the moment generating function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = -\sum_t \mathfrak{m}_t\{\mu, f_t(\boldsymbol{\theta}), f_t(\boldsymbol{\theta}_0)\}$. Obviously $\mathfrak{M}(\mu, \boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0$ and a simple algebra yields for the gradient $\nabla\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = d\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)/d\boldsymbol{\theta}$:

$$\boldsymbol{E}\nabla\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\mu\boldsymbol{E}\nabla L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\mu\nabla\boldsymbol{E}L(\boldsymbol{\theta}_0) = 0$$

because $\boldsymbol{\theta}_0$ is the point of maximum of $\boldsymbol{E}L(\boldsymbol{\theta})$. The same holds automatically for the lower bound $\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$. The Taylor expansion of the second order in a vicinity of $\boldsymbol{\theta}_0$ yields for all $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$ the following approximation:

$$\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \approx \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top I(\mu, \boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

with the matrix $I(\mu, \boldsymbol{\theta}_0) = \boldsymbol{E}\nabla^2 \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. This and the concentration property from Corollary 3.6 lead to the following bound on $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$:

Corollary 3.9. *Let (3.11) hold. Suppose that for some positive symmetric matrix D and some $\mathfrak{r} > 0$, the function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ fulfills almost surely*

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top D^2 (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \qquad \boldsymbol{\theta} \in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0), \tag{3.15}$$

Then for any $\mathfrak{z} \le \mathfrak{r}$

$$\boldsymbol{P}\bigl(\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 > \mathfrak{z}\bigr) \le \mathfrak{Q}(\varrho)\mathrm{e}^{-\varrho[\mathfrak{z} - \mathfrak{b}(\mathfrak{z})]}.$$

Proof. It is obvious that

$$\begin{aligned}
\bigl\{\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 > \mathfrak{z}\bigr\} &\subseteq \bigl\{\|D(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 > \mathfrak{z}, \, \widetilde{\boldsymbol{\theta}} \in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)\bigr\} \cup \bigl\{\widetilde{\boldsymbol{\theta}} \not\in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)\bigr\} \\
&\subseteq \bigl\{\mathfrak{M}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \mathfrak{z}, \, \widetilde{\boldsymbol{\theta}} \in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)\bigr\} \cup \bigl\{\overline{\mathfrak{M}}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \mathfrak{z}\bigr\} \\
&= \bigl\{\overline{\mathfrak{M}}(\mu, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) > \mathfrak{z}\bigr\}
\end{aligned}$$

and the result follows from Corollary 3.6. □

In the next theorem we assume the lower bound (3.15) to be fulfilled on the whole parameter set Θ. The general case can be reduced to this one by using once again the concentration property of Corollary 3.6.

Theorem 3.10. *Suppose* (*E*), (*ED*) *with $V(\boldsymbol{\theta}) \le V^*$ for a matrix V^*. Let also for some $\mathfrak{a} > 0$ and a deterministic function $\overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ hold*

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \mathfrak{a}^2 (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top V^* (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \qquad \boldsymbol{\theta} \in \Theta. \tag{3.16}$$

Fix some $\mathfrak{a}_1 \le \mathfrak{a}$ and define $\operatorname{pen}(\boldsymbol{\theta})$ *by*

$$\operatorname{pen}(\boldsymbol{\theta}) = \mathfrak{a}_1^2 (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top V^* (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{3.17}$$

Then with $s = 1 - \mathfrak{a}_1^2/\mathfrak{a}^2$ it holds

$$\begin{aligned}
\mathfrak{Q}(\varrho, s) &\overset{\text{def}}{=} \log \boldsymbol{E} \exp\bigl\{\varrho \sup_{\boldsymbol{\theta}} \bigl[\mu L(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) - \operatorname{pen}(\boldsymbol{\theta})\bigr]\bigr\} \\
&\le 2\varrho + (1 - \varrho)\mathbb{Q}_p + \log\left(1 + \frac{\omega_p^{-1}\pi^{p/2}}{(1 - \varrho)^{p/2}\mathfrak{a}_1^p}\right) \\
&\le pC(\varrho) + p\log\bigl(|\mathfrak{a}^2(1 - s)(1 - \varrho)|^{-1/2}\bigr)
\end{aligned} \tag{3.18}$$

for some fixed constant $C(\varrho)$. In addition, $\mathfrak{b}(\mathfrak{r}) = 0$ for all $\mathfrak{r} \ge 0$ yielding for any $\mathfrak{z} > 0$ the concentration property and confidence bound:

$$\begin{aligned} \boldsymbol{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_0)\bigr) &\le \mathfrak{Q}(\varrho, s)\mathrm{e}^{-\varrho s \mathfrak{z}}, \qquad & \mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_0) &= \{\boldsymbol{\theta} : \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \le \mathfrak{z}\}, \\ \boldsymbol{P}\bigl(\boldsymbol{\theta}_0 \notin \mathcal{E}(\mathfrak{z})\bigr) &\le \mathfrak{Q}(\varrho, 0)\mathrm{e}^{-\varrho \mathfrak{z}}, & \mathcal{E}(\mathfrak{z}) &= \{\boldsymbol{\theta} : L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z}\}. \end{aligned}$$

Proof. We apply Theorem 3.4 with

$$\varkappa(t) = \exp\bigl\{-(1-\varrho)\mathfrak{a}_1^2 (t-1)_+^2\bigr\}$$

leading for $\epsilon^2 = (1-\varrho)/\varrho$ and $t = \epsilon^{-1}\bigl\|\sqrt{V^*}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\bigr\|$ to the formula (3.17) for $\operatorname{pen}(\boldsymbol{\theta})$. By simple algebra

$$\mathfrak{P}^* = \omega_p^{-1} \int_{I\!R^p} \varkappa(\|\boldsymbol{\theta}\|) d\boldsymbol{\theta} = 1 + \omega_p^{-1} \frac{\pi^{p/2}}{(1-\varrho)^{p/2}\mathfrak{a}_1^p};$$

cf. the bound (3.10) for $\mathfrak{P}^*$ with $\delta_1 = (1-\varrho)\mathfrak{a}_1^2$. This implies the bound (3.18) for the $\mathfrak{Q}(\varrho)$ because $p^{-1}\mathbb{Q}_p$ and $p^{-1}\log\omega_p^{-1}$ are bounded by some fixed constants.

The inequality (3.16) ensures for $\mathfrak{r} = \overline{\mathfrak{M}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ that $\operatorname{pen}(\boldsymbol{\theta}) \le \mathfrak{a}_1^2/\mathfrak{a}^2\mathfrak{r}$, i.e. $\mathfrak{b}(\mathfrak{r}) \le \mathfrak{a}_1^2/\mathfrak{a}^2\mathfrak{r}$ and $\mathfrak{b} = \mathfrak{b}(0) = 0$. Finally, the concentration and coverage bounds follow from Corollaries 3.6 and 3.8. □

3.4 Sub-ergodocity and root-n consistency

Consider for every $\boldsymbol{\theta} \in \Theta$ a $p \times p$ random matrix $B(\boldsymbol{\theta}) = \sum_t \mathfrak{n}_t^2(f_t(\boldsymbol{\theta}))\, \nabla f_t(\boldsymbol{\theta}) \nabla f_t(\boldsymbol{\theta})^\top$; see (3.5). Condition (3.4) implies that for any $\boldsymbol{\gamma} \in \mathbb{S}^p$ and $|\lambda| \le \lambda^*$

$$\boldsymbol{E}\exp\bigl\{2\lambda\boldsymbol{\gamma}^\top\nabla\zeta(\boldsymbol{\theta}) - 2\lambda^2\boldsymbol{\gamma}^\top B(\boldsymbol{\theta})\boldsymbol{\gamma}\bigr\} \le 1.$$

The usual ergodicity condition for the sum $B(\boldsymbol{\theta})$ means that $n^{-1}B(\boldsymbol{\theta})$ converges to some deterministic matrix $b(\boldsymbol{\theta})$ for every $\boldsymbol{\theta}$ as n grows. Sub-ergodicity can be understood in the sense that $n^{-1}B(\boldsymbol{\theta})$ is bounded by some deterministic matrix $v(\boldsymbol{\theta})$ with a high probability. We define $V(\boldsymbol{\theta}) = nv(\boldsymbol{\theta})$ and suppose that conditions (ED) and (V) are fulfilled for such defined $V(\boldsymbol{\theta})$.

Similarly the sub-ergodicity applied to the random quantity $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ means that there is a deterministic positive function $\overline{\mathfrak{m}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ such that $n^{-1}\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \overline{\mathfrak{m}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ with a high probability. In this situation one can rewrite the main corollaries from Section 3.2 in terms of the functions $v(\boldsymbol{\theta})$ and $\overline{\mathfrak{m}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$. In particular, the concentration set $\mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)$ can be replaced by

$$\overline{\mathcal{A}}(\mathfrak{r}, \boldsymbol{\theta}_0) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\theta} : n\overline{\mathfrak{m}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \le \mathfrak{r}\}.$$

In addition we assume similarly to (3.15) that for some fixed symmetric positive matrix D_1 and some $\mathfrak{r} > 0$, it holds in the vicinity $\mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0)$ of the point $\boldsymbol{\theta}_0$:

$$\overline{\mathfrak{m}}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top D_1^2 (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \qquad v(\boldsymbol{\theta}) \leq \mathfrak{a}^2 D_1^2 \quad \boldsymbol{\theta} \in \mathcal{A}(\mathfrak{r}, \boldsymbol{\theta}_0). \tag{3.19}$$

Corollary 3.11. *Assume (3.11) and (3.19) for some* $\mathfrak{r} > 0$. *Then for any* $\mathfrak{z} \leq \mathfrak{r}$

$$\boldsymbol{P}\bigl(\|D_1(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|^2 > \mathfrak{z}/n\bigr) \leq \mathfrak{Q}(\varrho) \exp\{-\varrho(\mathfrak{z} - \mathfrak{b})\}.$$

4 Applications and examples

This section illustrates how the general results can be applied to some popular examples of parametric time series models which we already mentioned in Section 2.3. For all examples we assume that the two components of the parametric modeling are fixed: a parametric family $\mathcal{P} = (P_v, v \in \mathcal{U})$ and a family of parameter processes $\{f(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$. Moreover, we assume that both families are sufficiently regular, in particular, the functions $f_t(\boldsymbol{\theta})$ are differentiable w.r.t. $\boldsymbol{\theta}$. The corresponding gradient is denoted by $\nabla f_t(\boldsymbol{\theta}) = \partial f_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \in I\!R^p$.

In all the examples, the real model dynamics is described by some predictable process f_t accepting that the parametric assumption $f_t = f_t(\boldsymbol{\theta})$ is not precisely fulfilled whatever $\boldsymbol{\theta} \in \Theta$ is. By $\boldsymbol{\theta}_0$ we denote the parameter corresponding to the best parametric fit: $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta}} \boldsymbol{E} L(\boldsymbol{\theta})$. Such defined vector parameter $\boldsymbol{\theta}_0$ is estimated by maximizing the objective function $L(\boldsymbol{\theta})$.

4.1 Linear autoregression

Assume the model $Y_t = f_t + \varepsilon_t$ in which the innovations ε_t are martingale differences: $\boldsymbol{E}\bigl[\varepsilon_t | \mathcal{F}_{t-1}\bigr] = 0$ possible heterogeneous with bounded exponential moments: $\log \boldsymbol{E} \exp\bigl\{\lambda \varepsilon_t \big| \mathcal{F}_{t-1}\bigr\} \leq \varkappa_1 \lambda^2$ for $\lambda \leq \lambda^*$ and some fixed $\varkappa_1$.

With $L(\boldsymbol{\theta}) = -(2\sigma^2)^{-1} \sum_t \bigl\{Y_t - f_t(\boldsymbol{\theta})\bigr\}^2$, the least square estimate $\widetilde{\boldsymbol{\theta}}$ reads as

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_t \bigl\{Y_t - f_t(\boldsymbol{\theta})\bigr\}^2.$$

Linear autoregression means the structural equation $f_t(\boldsymbol{\theta}) = \alpha_1 Y_{t-1} + \ldots + \alpha_p Y_{t-p}$ for $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_p)^\top$ leading to maximization of the quadratic functional $\sum_t \bigl(Y_t - \Psi_t^\top \boldsymbol{\theta}\bigr)^2$ with $\Psi_t = (Y_{t-1}, \ldots, Y_{t-p})^\top$ which admits a closed form solution:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \Bigl(\sum_t \Psi_t \Psi_t^\top\Bigr)^{-1} \sum_t Y_t \Psi_t = B^{-1} \sum_t Y_t \Psi_t \\ L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) &= \frac{1}{2\sigma^2} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top B (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \end{aligned}$$

with $B \stackrel{\text{def}}{=} \sum_t \Psi_t \Psi_t^\top$.

The matrix $B(\boldsymbol{\theta})$ from (3.5) does not depend on $\boldsymbol{\theta}$: $B(\boldsymbol{\theta}) \equiv B$. Next, the formulae for $\nabla\zeta(\boldsymbol{\theta})$ and $\nabla M(\boldsymbol{\theta})$ simplifies to

$$\begin{aligned} \nabla\zeta(\boldsymbol{\theta}) &= \sigma^{-2} \sum_t \{Y_t - f_t\} \Psi_t = \sigma^{-2} \sum_t \varepsilon_t \Psi_t , \\ \nabla M(\boldsymbol{\theta}) &= \sigma^{-2} \sum_t \{ f_t - \Psi_t^\top \boldsymbol{\theta} \} \Psi_t . \end{aligned}$$

A natural choice for the matrix $V(\boldsymbol{\theta})$ is

$$V(\boldsymbol{\theta}) = C_1 \sigma^{-2} \boldsymbol{E} B + C_2 \operatorname{Var} \nabla M(\boldsymbol{\theta})$$

for some $C_1, C_2 \ge 1$. Such defined matrix $V(\boldsymbol{\theta})$ is a quadratic function of $\boldsymbol{\theta}$ and condition (V) is straightforward. Condition (ED) is also easy to check in typical situations.

The value $\boldsymbol{\theta}_0$ which maximizes $\boldsymbol{E} L(\boldsymbol{\theta})$ can be found by the following optimization problem:

$$\begin{aligned} \boldsymbol{\theta}_0 &= \operatorname*{argmax}_{\boldsymbol{\theta}} \boldsymbol{E} L(\boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \boldsymbol{E} \sum_t \bigl(Y_t - \Psi_t \boldsymbol{\theta} \bigr)^2 \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \boldsymbol{E} \sum_t \bigl(f_t - \Psi_t \boldsymbol{\theta} \bigr)^2 = \Bigl(\boldsymbol{E} \sum_t \Psi_t \Psi_t^\top \Bigr)^{-1} \boldsymbol{E} \sum_t f_t \Psi_t . \qquad (4.1) \end{aligned}$$

Theorem 3.3 and Corollary 3.7 claim that $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = (2\sigma^2)^{-1} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top B (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is stochastically bounded with finite polynomial and exponential moments:

$$\boldsymbol{E} \bigl| (2\sigma^2)^{-1} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top B (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \bigr|^r \le \mathfrak{R}(r).$$

This also justifies the use of confidence sets in the form

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \bigl\{ \boldsymbol{\theta} : (2\sigma^2)^{-1} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top B (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \le \mathfrak{z} \bigr\}.$$

By Corollary 3.8, this set does not contain the target $\boldsymbol{\theta}_0$ with a probability exponentially decreasing in $\mathfrak{z}$. The result is valid simultaneously for stationary, stable (unit root) and explosive cases, also mixed structure is allowed. The only essential assumption is about exponential moments of the errors.

Finally we briefly discuss the concentration property of the estimate $\widetilde{\boldsymbol{\theta}}$. For applying Corollary 3.6 one has to compute or evaluate the rate function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Suppose for simplicity that the errors ε_t are conditionally normal with zero mean and the conditional variance σ_t^2, that is, $Q_t = \mathcal{N}(0, \sigma_t^2)$. Then for any υ, υ'

$$\begin{aligned} \mathfrak{m}_t(\mu, \upsilon, \upsilon') &\stackrel{\text{def}}{=} -\log \boldsymbol{E} \bigl[\exp\bigl\{ \mu \ell(Y_t, \upsilon) - \mu \ell(Y_t, \upsilon') \bigr\} \bigm| \mathcal{F}_{t-1} \bigr] \\ &= \frac{\mu}{2\sigma^2} \bigl\{ (f_t - \upsilon)^2 - (f_t - \upsilon')^2 \bigr\} - \frac{\mu^2 \sigma_t^2}{2\sigma^4} (\upsilon' - \upsilon)^2. \end{aligned}$$

Therefore,

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{\mu}{2\sigma^2} \sum_t \bigl\{ (f_t - \Psi_t \boldsymbol{\theta})^2 - (f_t - \Psi_t \boldsymbol{\theta}_0)^2 - \frac{\mu \sigma_t^2}{\sigma^2} (\Psi_t \boldsymbol{\theta} - \Psi_t \boldsymbol{\theta}_0)^2 \bigr\}.$$

In particular, under the parametric assumption $f_t = \Psi_t^\top \boldsymbol{\theta}_0$

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{\mu}{2\sigma^2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \sum_t \bigl(1 - \mu \sigma_t^2 / \sigma^2 \bigr) \Psi_t \Psi_t^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

If, in addition, the variance σ_t is homogeneous, $\sigma_t \equiv \sigma_0$, then

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \frac{\mu}{2\sigma^2} \bigl(1 - \mu \sigma_0^2 / \sigma^2 \bigr) B (\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Optimizing w.r.t. μ yields $\mu = \sigma^2/(2\sigma_0^2)$ and $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top B (\boldsymbol{\theta} - \boldsymbol{\theta}_0)/(4\sigma_0^2)$. In the ergodic case, when the matrix $n^{-1}B$ is close to the stationary limit b, the result of Corollary 3.6 claims that $\widetilde{\boldsymbol{\theta}}$ concentrates in the root-n neighborhood $\mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|b^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 \le 4\sigma_0^2 \mathfrak{z}/n\}$ of $\boldsymbol{\theta}_0$.

If the parametric assumption $f_t = \Psi_t^\top \boldsymbol{\theta}$ is not fulfilled whatever $\boldsymbol{\theta}$ is, then the calculations become only slightly more complicated. Namely,

$$\begin{aligned} \mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \frac{\mu}{2\sigma^2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \sum_t \bigl(1 - \mu \sigma_t^2 / \sigma^2 \bigr) \Psi_t \Psi_t^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &\quad + \frac{\mu}{\sigma^2} \sum_t \bigl(f_t \Psi_t - \Psi_t \Psi_t^T \boldsymbol{\theta}_0 \bigr) (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \end{aligned}$$

Due to (4.1), the expectation of the second sum is zero and this term is typically smaller in order than the first one. So, $\mathcal{A}(\mathfrak{z}, \boldsymbol{\theta}_0)$ remains to be a concentration set for $\widetilde{\boldsymbol{\theta}}$.

4.2 Categorical time series

Let $\mathcal{P} = (P_v, v \in \mathcal{U})$ be a given exponential family with the canonical parameter. The conditional distribution $Q_t = \mathcal{L}\bigl(Y_t \big| \mathcal{F}_{t-1}\bigr)$ of every observation Y_t given the past $\mathcal{F}_{t-1}$ is assumed to be in $\mathcal{P}$ and described by the varying stochastic parameter f_t: $Q_t = P_{f_t} \in \mathcal{P}$. The parametric assumption $Q_t = P_{f_t(\boldsymbol{\theta})}$ leads to the log-likelihood $L(\boldsymbol{\theta}) = \sum_t \ell(Y_t, f_t(\boldsymbol{\theta}))$ where $\ell(y, v) = yv - d(v)$ is the log-likelihood function for $\mathcal{P}$:

$$L(\boldsymbol{\theta}) = \sum_t \ell(Y_t, f_t(\boldsymbol{\theta})) = \sum_t \bigl\{ Y_t f_t(\boldsymbol{\theta}) - d\bigl(f_t(\boldsymbol{\theta})\bigr) \bigr\}. \tag{4.2}$$

We use the well known properties of the canonical exponential families: $E_v Y = \dot{d}(v)$ and $\log E_v \exp\bigl\{\mu Y\bigr\} = d(v + \mu) - d(v)$. This yields for every $v, v' \in \mathcal{U}$ and all t

$$\begin{aligned} m_t(v) &\overset{\text{def}}{=} \boldsymbol{E}\bigl[\ell(Y_t, v) \big| \mathcal{F}_{t-1}\bigr] = \dot{d}(f_t) v - d(v), \\ \zeta_t(Y_t, v) &\overset{\text{def}}{=} \boldsymbol{E}\bigl[\ell(Y_t, v) - m_t(v) \big| \mathcal{F}_{t-1}\bigr] = \bigl\{ Y_t - \dot{d}(f_t) \bigr\} v, \\ \mathfrak{m}_t(\mu, v, v') &\overset{\text{def}}{=} -\log \boldsymbol{E}\bigl[\exp\{\mu \ell(Y_t, v, v')\} \big| \mathcal{F}_{t-1}\bigr] \\ &= d(f_t) - d(f_t + \mu(v - v')) + \mu d(v) - \mu d(v'). \end{aligned}$$

It is easy to see that condition (E) is fulfilled if $f_t + \mu\bigl\{f_t(\boldsymbol{\theta}) - f_t(\boldsymbol{\theta}_0)\bigr\} \in \mathcal{U}$ for all t and all $\boldsymbol{\theta} \in \Theta$.

Next, $\dot{\zeta}_t(Y_t, \upsilon) \overset{\mathrm{def}}{=} \partial\zeta_t(Y_t, \upsilon)/\partial\upsilon = Y_t - \dot{d}(f_t)$. In particular, this expression does not depend on υ. Let $\mathfrak{n}(\upsilon)$ be a function of υ which ensures for some fixed $\lambda_1^* > 0$ that

$$\log E_\upsilon \exp\Bigl\{2\lambda \frac{Y - \dot{d}(\upsilon)}{\mathfrak{n}(\upsilon)}\Bigr\} \le 2\lambda^2, \qquad \lambda \le \lambda_1^*$$

Define

$$\begin{aligned} M(\boldsymbol{\theta}) &\overset{\mathrm{def}}{=} \sum_t m_t\bigl(f_t(\boldsymbol{\theta})\bigr) = \sum_t \bigl\{\dot{d}(f_t) f_t(\boldsymbol{\theta}) - d(f_t(\boldsymbol{\theta}))\bigr\}, \\ \zeta(\boldsymbol{\theta}) &\overset{\mathrm{def}}{=} \sum_t \zeta_t\bigl(Y_t, f_t(\boldsymbol{\theta})\bigr) = \sum_t \bigl\{Y_t - \dot{d}(f_t)\bigr\} f_t(\boldsymbol{\theta}), \\ B(\boldsymbol{\theta}) &\overset{\mathrm{def}}{=} \sum_t \mathfrak{n}^2(f_t)\, \nabla f_t(\boldsymbol{\theta}) \bigl\{\nabla f_t(\boldsymbol{\theta})\bigr\}^\top. \end{aligned}$$

Then under simple conditions on the parameter process $f_t(\boldsymbol{\theta})$, the derivatives

$$\begin{aligned} \nabla M(\boldsymbol{\theta}) &\overset{\mathrm{def}}{=} \sum_t \bigl\{\dot{d}(f_t) - \dot{d}(f_t(\boldsymbol{\theta}))\bigr\} \nabla f_t(\boldsymbol{\theta}), \\ \nabla\zeta(\boldsymbol{\theta}) &\overset{\mathrm{def}}{=} \sum_t \bigl\{Y_t - \dot{d}(f_t)\bigr\} \nabla f_t(\boldsymbol{\theta}), \end{aligned}$$

fulfill the conditions (ED) and (V) with $V(\boldsymbol{\theta}) = C_1 \boldsymbol{E} B(\boldsymbol{\theta}) + C_2 \operatorname{Var} \nabla M(\boldsymbol{\theta})$ for some $C_1, C_2 \ge 1$.

Sub-ergodic properties are strictly related to the behavior of the parameter process $f_t(\boldsymbol{\theta})$ and its derivative $\nabla f_t(\boldsymbol{\theta})$. Usually it suffices to assume that the both processes remains bounded, at least with a dominating probability. Root-n consistency can be shown under the identifiability condition that

$$n^{-1}\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = n^{-1} \sum_t \mathfrak{m}_t\bigl(\mu, f_t(\boldsymbol{\theta}), f_t(\boldsymbol{\theta}_0)\bigr) \ge \overline{\mathfrak{m}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$$

where $\overline{\mathfrak{m}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is two times continuously differentiable and satisfies $\overline{\mathfrak{m}}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) > 0$ for $\boldsymbol{\theta} \ne \boldsymbol{\theta}_0$. Then Corollary 3.9 ensures root-n consistency of the estimate $\widetilde{\boldsymbol{\theta}}$.

4.3 Estimation for Generalized Linear time series

Here we consider the case when $\mathcal{P}$ is again an exponential family with the canonical parameter and the parametric function $f_t(\boldsymbol{\theta})$ is a transformation of another function which linearly depends on the parameter $\boldsymbol{\theta}$. However, in the contrary to the previous example, we admit that the true conditional data distribution is not in $\mathcal{P}$.

A generalized linear specification for this model is given by the following set of structural equations:

$$\mathcal{L}\bigl(Y_t \big| \mathcal{F}_{t-1}\bigr) = P_{f_t} \in \mathcal{P}, \qquad f_t = g(\varPsi_t), \qquad \varPsi_t = A(\boldsymbol{\theta})\varPsi_{t-1} \tag{4.3}$$

where $\varPsi_t$ is a predictable R^d-dimensional explanatory process, $g(\cdot)$ is a given mapping from $I\!R^d$ to $I\!R$, $A(\boldsymbol{\theta})$ is a given $d \times d$-matrix linearly depending on the parameter vector $\boldsymbol{\theta} \in \varTheta \subset I\!R^p$.

Let $\boldsymbol{\theta}$ be the parameter vector. Then, given $\boldsymbol{\theta}$, the observations $Y_1, \ldots, Y_t$, and the pre-history $\varPsi^0, Y^0$, one can uniquely reconstruct the process $\varPsi_t = \varPsi_t(\boldsymbol{\theta})$ and then $f_t(\boldsymbol{\theta}) = g(\varPsi_t(\boldsymbol{\theta}))$ for $t \ge 1$ by recurrently applying the relation (4.3). This function $f_t(\boldsymbol{\theta})$ leads to the (quasi) log-likelihood $L(\boldsymbol{\theta}) = \sum_t \bigl\{ Y_t f_t(\boldsymbol{\theta}) - d\bigl(f_t(\boldsymbol{\theta})\bigr) \bigr\}$.

A useful feature of models of type (4.3) is that the gradient $\nabla f_t(\boldsymbol{\theta})$ also follows the linear structural equation:

$$\begin{aligned} \nabla f_t &= \nabla \varPsi_t(\boldsymbol{\theta}) g'(\varPsi_t(\boldsymbol{\theta})), \\ \nabla \varPsi_t(\boldsymbol{\theta}) &= \nabla A \cdot X_{t-1}(\boldsymbol{\theta}) + A(\boldsymbol{\theta}) \nabla X_{t-1}(\boldsymbol{\theta}) \end{aligned}$$

where $g'(\cdot)$ means the gradient of $g(\cdot)$, and ∇A is the gradient of of $A(\boldsymbol{\theta})$ which does not depend on $\boldsymbol{\theta}$ because $A(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$.

Define

$$\begin{aligned} b_t &\overset{\text{def}}{=} \boldsymbol{E}\bigl[Y_t \big| \mathcal{F}_{t-1}\bigr], \\ m_t(\upsilon) &\overset{\text{def}}{=} \boldsymbol{E}\bigl[\ell(Y_t, \upsilon) \big| \mathcal{F}_{t-1}\bigr] = b_t \upsilon - d(\upsilon), \\ \zeta_t(Y_t, \upsilon) &\overset{\text{def}}{=} \boldsymbol{E}\bigl[\ell(Y_t, \upsilon) - m_t(\upsilon) \big| \mathcal{F}_{t-1}\bigr] = (Y_t - b_t)\upsilon, \\ \mathfrak{m}_t(\mu, \upsilon, \upsilon') &\overset{\text{def}}{=} \log \boldsymbol{E}\bigl[\exp\{\mu \ell(Y_t, \upsilon, \upsilon')\} \big| \mathcal{F}_{t-1}\bigr]. \end{aligned} \tag{4.4}$$

All these quantities are computed for the underlying data distribution $\boldsymbol{P}$ for which both the assumption on the conditional distribution $Q_t \in \mathcal{P}$ and the parametric dynamics $f_t = f_t(\boldsymbol{\theta}_0)$ can be violated.

We suppose the following condition to be satisfied:

(*Yt*) *There exist a constant* $\lambda_1^* > 0$ *and a predictable process* $\mathfrak{n}_t$ *such that for all* t *and* $\lambda \le \lambda_1^*$

$$\log \boldsymbol{E} \exp\bigl\{2\lambda \mathfrak{n}_t^{-1}(Y_t - b_t) \big| \mathcal{F}_{t-1}\bigr\} \le 2\lambda^2.$$

This condition ensures that for all $\upsilon, \upsilon' \in \mathcal{U}$ and $\mu > 0$ with $\mu(\upsilon - \upsilon') \le 2\lambda_1^*/\mathfrak{n}_t$ the quantity $\mathfrak{m}_t(\mu, \upsilon, \upsilon')$ from (4.4) is well defined.

The description of the process $\nabla\varPsi_t(\boldsymbol{\theta})$ also yields the representation for $\nabla\zeta(\boldsymbol{\theta})$, $\nabla M(\boldsymbol{\theta})$, $B(\boldsymbol{\theta})$; see (3.5):

$$\begin{aligned}
\nabla\zeta(\boldsymbol{\theta}) &= \sum_t (Y_t - b_t)\nabla f_t(\boldsymbol{\theta}) = \sum_t \{Y_t - b_t\}\nabla\varPsi_t(\boldsymbol{\theta})g'(\varPsi_t(\boldsymbol{\theta})), \\
\nabla M(\boldsymbol{\theta}) &= \sum_t \bigl\{b_t - \dot{d}(f_t(\boldsymbol{\theta}))\bigr\}\nabla\varPsi_t(\boldsymbol{\theta})g'(\varPsi_t(\boldsymbol{\theta})). \\
B(\boldsymbol{\theta}) &= \sum_t \mathfrak{n}_t^2\, \nabla\varPsi_t(\boldsymbol{\theta})g'(\varPsi_t(\boldsymbol{\theta}))\bigl\{\nabla\varPsi_t(\boldsymbol{\theta})g'(\varPsi_t(\boldsymbol{\theta}))\bigr\}^\top.
\end{aligned}$$

This suggests to take the matrix $V(\boldsymbol{\theta})$ in the form

$$V(\boldsymbol{\theta}) = C_1\boldsymbol{E}B(\boldsymbol{\theta}) + C_2 \operatorname{Var}\nabla M(\boldsymbol{\theta}) \tag{4.5}$$

with some $C_1, C_2 \geq 1$. Checking the condition (ED) is quite straightforward in the most of situations and all the results of Sections 3, 3.2 apply.

We now specify the above expressions for the important special case of $d = 1$ when the model is given by the equations $f_t = g(X_t)$ for a univariate link function g and a univariate process X_t following the linear dynamic equation

$$X_t = \omega + \alpha_1 Y_{t-1} + \ldots \alpha_p Y_{t-p} + \beta_1 X_{t-1} + \ldots \beta_q X_{t-q}.$$

Here $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_q, \alpha_1, \ldots, \alpha_p, \omega)^\top$. One easily computes

$$\begin{aligned}
\nabla X_t(\boldsymbol{\theta}) &= (1, Y_{t-1}, \ldots Y_{t-p}, X_{t-1}(\boldsymbol{\theta}), \ldots, X_{t-q}(\boldsymbol{\theta})) \\
&\qquad + \biggl(0, \ldots, 0, \beta_1\frac{\partial X_{t-1}(\boldsymbol{\theta})}{\partial\beta_1}, \ldots, \beta_q\frac{\partial X_{t-q}(\boldsymbol{\theta})}{\partial\beta_q}\biggr).
\end{aligned}$$

So, given the initial values $X_{1-q}, \ldots, X_0$, $\nabla X_{1-q}, \ldots, \nabla X_0$ and the observations $Y_1, \ldots, Y_n$, one can recurrently construct for every $\boldsymbol{\theta}$ the hidden process $X_t(\boldsymbol{\theta})$ and its gradient $\nabla X_t(\boldsymbol{\theta})$. This yields the representation $f_t(\boldsymbol{\theta}) = g\bigl(X_t(\boldsymbol{\theta})\bigr)$ and $\nabla f_t(\boldsymbol{\theta}) = g'(X_t(\boldsymbol{\theta}))\nabla X_t(\boldsymbol{\theta})$ for the parameter process $f_t(\boldsymbol{\theta})$. The formulae for $\nabla\zeta(\boldsymbol{\theta})$, $\nabla M(\boldsymbol{\theta})$, $B(\boldsymbol{\theta})$ can also be specified:

$$\begin{aligned}
\nabla\zeta(\boldsymbol{\theta}) &= \sum_t (Y_t - b_t)\nabla f_t(\boldsymbol{\theta}) = \sum_t \{Y_t - b_t\}g'(X_t(\boldsymbol{\theta}))\nabla X_t(\boldsymbol{\theta}), \\
\nabla M(\boldsymbol{\theta}) &= \sum_t \bigl\{b_t - \dot{d}(f_t(\boldsymbol{\theta}))\bigr\}g'(X_t(\boldsymbol{\theta}))\nabla X_t(\boldsymbol{\theta}). \\
B(\boldsymbol{\theta}) &= \sum_t \mathfrak{n}_t^2\, \bigl|g'(X_t(\boldsymbol{\theta}))\bigr|^2 \nabla X_t(\boldsymbol{\theta})\bigl\{\nabla X_t(\boldsymbol{\theta})\bigr\}^\top.
\end{aligned}$$

Further one can proceed as in the general case with $d > 1$.

4.4 GARCH(1,1) estimation

GARCH-model with the parameter $\boldsymbol{\theta} = (\omega, \alpha, \beta)^\top$ can be described by structural equation:

$$X_t(\boldsymbol{\theta}) = \omega + \alpha R_{t-1} + \beta X_{t-1}(\boldsymbol{\theta}), \qquad t \ge 1, \qquad X_0(\boldsymbol{\theta}) = X_0, \tag{4.6}$$

yielding the (quasi) log likelihood

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_t \bigl\{ \log(2\pi X_t(\boldsymbol{\theta})) + R_t / X_t(\boldsymbol{\theta}) \bigr\} \\ &= \frac{1}{2}\sum_t \bigl\{ \log(2\pi f_t(\boldsymbol{\theta})) - R_t f_t(\boldsymbol{\theta}) \bigr\} \end{aligned}$$

with $f_t(\boldsymbol{\theta}) = 1/X_t(\boldsymbol{\theta})$.

Below we assume that the innovations ε_t^2's have bounded conditional exponential moments: with some $\lambda^* > 0$ and $\mathfrak{n}$, it holds almost surely

$$\log \boldsymbol{E} \exp\bigl\{ -\frac{\lambda}{\mathfrak{n}} (\varepsilon_t^2 - 1) \big| \mathcal{F}_{t-1} \bigr\} \le 2\lambda^2, \qquad |\lambda| \le \lambda^*.$$

If every ε_t is conditionally standard normal then $\log \boldsymbol{E} \exp\bigl\{ -\lambda(\varepsilon_t^2 - 1) \big| \mathcal{F}_{t-1} \bigr\} = \lambda - \frac{1}{2}\log(1 + 2\lambda) \le 2\lambda^2$ for $|\lambda| \le 1/3$.

As before, $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \boldsymbol{E} L(\boldsymbol{\theta})$ denotes the "true" parameter vector. By some technical reason we assume that the constant term ω of the parameter vector $\boldsymbol{\theta}$ is not smaller than given value $\delta > 0$. Then the equation (4.6) ensures for every $\boldsymbol{\theta}$ and every $t \ge 1$ the lower bound $X_t(\boldsymbol{\theta}) \ge \delta$.

The gradient $\nabla X_t(\boldsymbol{\theta})$ of the process $X_t(\boldsymbol{\theta})$ satisfies the equation

$$\nabla X_t(\boldsymbol{\theta}) = (1, R_{t-1}, X_{t-1}(\boldsymbol{\theta}))^\top + \beta \nabla X_{t-1}(\boldsymbol{\theta}).$$

For the canonical parameter $f_t(\boldsymbol{\theta}) = 1/X_t(\boldsymbol{\theta})$ this yields

$$\nabla f_t(\boldsymbol{\theta}) = X_t^{-2}(\boldsymbol{\theta}) \nabla X_t(\boldsymbol{\theta}) = f_t^2(\boldsymbol{\theta}) \nabla X_t(\boldsymbol{\theta}).$$

Next, one easily computes

$$\begin{aligned} \nabla \zeta(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_t (R_t - X_t) \nabla f_t(\boldsymbol{\theta}), \\ \nabla M(\boldsymbol{\theta}) &= \frac{1}{2}\sum_t \bigl\{ X_t - X_t(\boldsymbol{\theta}) \bigr\} \nabla f_t(\boldsymbol{\theta}). \end{aligned}$$

With $\mathfrak{n}_t \equiv \mathfrak{n}$ and $Y_t = -R_t/2$, one has for $\lambda \le \lambda^*$

$$\log \boldsymbol{E}\bigl[\exp\bigl\{ 2\lambda \mathfrak{n}_t^{-1} (Y_t - b_t) \bigr\} \big| \mathcal{F}_{t-1} \bigr] = \log \boldsymbol{E}\bigl[\exp(-\lambda \varepsilon_t^2 + \lambda) \big| \mathcal{F}_{t-1} \bigr] \le 2\lambda^2$$

for $|\lambda| \le \lambda_1^*$, and condition (Yt) is verified. Next, condition (ED) can be checked with $V(\boldsymbol{\theta})$ from (4.5), (3.5). The results of Section 3 and 3.2 describe the accuracy of estimation in terms of the function

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \;=\; \sum_t \mathfrak{m}_t\bigl\{\mu, f_t(\boldsymbol{\theta}), f_t(\boldsymbol{\theta}_0)\bigr\},$$

where for υ, υ' with $|\mu X_t^{-1}(\upsilon - \upsilon')|\mathfrak{n} \le \lambda^*/2$ that

$$\begin{aligned}
\mathfrak{m}_t\{\mu, \upsilon, \upsilon'\} &\overset{\text{def}}{=} -\log \boldsymbol{E}\bigl[\exp\bigl\{\mu\ell(R_t, \upsilon, \upsilon')\bigr\}\big|\mathfrak{F}_{t-1}\bigr] \\
&= \frac{\mu}{2}\log(\upsilon'/\upsilon) + \frac{\mu(\upsilon - \upsilon')}{2X_t} + \frac{1}{2}\log \boldsymbol{E}\exp\bigl\{\mu X_t^{-1}(\upsilon - \upsilon')(\varepsilon_t^2 - 1)\bigr\} \\
&\ge \frac{\mu}{2}\log(\upsilon'/\upsilon) + \frac{\mu(\upsilon - \upsilon')}{2X_t} - \frac{\mu^2(\upsilon - \upsilon')^2\mathfrak{n}^2}{4X_t^2}
\end{aligned}$$

yielding

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \sum_t \Bigl\{\frac{\mu}{2}\log\Bigl(\frac{f_t(\boldsymbol{\theta}_0)}{f_t(\boldsymbol{\theta})}\Bigr) + \frac{\mu\bigl\{f_t(\boldsymbol{\theta}) - f_t(\boldsymbol{\theta}_0)\bigr\}}{2X_t} - \frac{\mu^2|f_t(\boldsymbol{\theta}) - f_t(\boldsymbol{\theta}_0)|^2\mathfrak{n}^2}{4X_t^2}\Bigr\}$$

In the parametric situation $X_t = 1/f_t(\boldsymbol{\theta}_0)$ and with $\delta_t(\boldsymbol{\theta}) = f_t(\boldsymbol{\theta})/f_t(\boldsymbol{\theta}_0) - 1$

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) \ge \sum_t \Bigl\{\frac{\mu}{2}\log(1 + \delta_t(\boldsymbol{\theta})) + \frac{\mu\delta_t(\boldsymbol{\theta})}{2} - \frac{\mu^2\delta_t(\boldsymbol{\theta})^2\mathfrak{n}^2}{4}\Bigr\}$$

So, $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ can be viewed as a kind of distance between two functions $f_t(\boldsymbol{\theta})$ and $f_t(\boldsymbol{\theta}_0)$. In the stationary case $\alpha + \beta < 1$ for any $\boldsymbol{\theta} = (\omega, \alpha, \beta)^\top \in \Theta$, the process $X_t(\boldsymbol{\theta})$ is ergodic and the normalized sum $n^{-1}\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ converges to the integral of every summand w.r.t. the stationary measure. One can easily seen that this integral is nearly quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ in a neighborhood of $\boldsymbol{\theta}_0$ yielding the root-n consistency of estimation by Corollary 3.9.

4.5 Median and quantile time series estimation

The median or more generally quantile estimation can be defined by maximizing the quasi log-likelihood

$$L(\boldsymbol{\theta}) = \sum_t \ell_\alpha\bigl(Y_t - f_t(\boldsymbol{\theta})\bigr)$$

with $\ell_\alpha(x) = (1 - \alpha)x_- - \alpha x_+$. Define

$$\begin{aligned}
m_t(\boldsymbol{\theta}) &\overset{\text{def}}{=} \boldsymbol{E}\bigl\{\ell_\alpha\bigl(Y_t - f_t(\boldsymbol{\theta})\bigr)\big|\mathfrak{F}_{t-1}\bigr\}, \\
q_t(\boldsymbol{\theta}) &\overset{\text{def}}{=} \boldsymbol{P}\bigl(Y_t - f_t(\boldsymbol{\theta}) \le 0\big|\mathfrak{F}_{t-1}\bigr), \\
M(\boldsymbol{\theta}) &\overset{\text{def}}{=} \sum_t m_t(\boldsymbol{\theta}) = \sum_t \boldsymbol{E}\bigl\{\ell_\alpha\bigl(Y_t - f_t(\boldsymbol{\theta})\bigr)\big|\mathfrak{F}_{t-1}\bigr\}, \\
\zeta(\boldsymbol{\theta}) &\overset{\text{def}}{=} \sum_t \bigl\{\ell_\alpha\bigl(Y_t - f_t(\boldsymbol{\theta})\bigr) - m_t(\boldsymbol{\theta})\bigr\}.
\end{aligned}$$

Simple derivations yield

$$\begin{aligned}\nabla\zeta(\boldsymbol{\theta}) &= \sum_t \bigl\{ \mathbf{1}(Y_t - f_t(\boldsymbol{\theta}) \le 0) - q_t(\boldsymbol{\theta}) \bigr\} \nabla f_t(\boldsymbol{\theta}), \\ \nabla M(\boldsymbol{\theta}) &= \sum_t q_t(\boldsymbol{\theta}) \nabla f_t(\boldsymbol{\theta}).\end{aligned}$$

Here it is natural to set $\mathfrak{n}_t(\upsilon) \equiv 1$ leading to

$$B(\boldsymbol{\theta}) = \sum_t \nabla f_t(\boldsymbol{\theta}) \nabla f_t(\boldsymbol{\theta})^\top.$$

As previously, the condition of Theorem 3.3 are easy to verify with $V(\boldsymbol{\theta}) = C_1 \boldsymbol{E} B(\boldsymbol{\theta}) + C_2 \operatorname{Var} \nabla M(\boldsymbol{\theta})$ for some $C_1, C_2 \ge 1$. The function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ from (3.1) describing the quality of estimation can be represented as

$$\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \sum_t \mathfrak{m}_t(\mu, f_t(\boldsymbol{\theta}), f_t(\boldsymbol{\theta}_0))$$

with

$$\mathfrak{m}_t(\mu, \upsilon, \upsilon') = -\log \boldsymbol{E}\bigl[\exp\bigl\{\mu \ell_\alpha(Y_t - \upsilon) - \mu \ell_\alpha(Y_t - \upsilon')\bigr\} \big| \mathcal{F}_{t-1}\bigr]$$

4.6 Risk bounds. Summary

In this section we summarize what we obtained in all the previous examples and what can be stated in each particular case.

One can see that in every application there is a straightforward expression for the quasi likelihood function $L(\boldsymbol{\theta})$ and its components $\zeta(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta})$ and for their gradient in $\boldsymbol{\theta}$ in terms of $f_t(\boldsymbol{\theta})$ and $\nabla f_t(\boldsymbol{\theta})$. Moreover, all the conditions required for the main results of Sections 3 and 3.2 can be easily verified with the matrix $V(\boldsymbol{\theta})$ given in (4.5). Therefore, all these theorems and corollaries are applicable. In particular, the concentration property and the structure of confidence sets is given.

To understand what the results exactly imply in particular cases, one has to bound the penalty function $\operatorname{pen}(\boldsymbol{\theta})$ from above and the rate function $\mathfrak{M}(\mu, \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ from below. The penalty function can be defined in the most of cases via the linear ranking $\|\sqrt{V(\boldsymbol{\theta})}(\boldsymbol{\theta} - \boldsymbol{\theta})\|$. The rate admits typically a quadratic in $\boldsymbol{\theta} - \boldsymbol{\theta}$ lower bound which immediately yields the classical root-n accuracy by Corollary 3.11. The general case of Sections 3 and 3.2 apply in each of considered example without any significant change of formulation.

5 A penalized exponential bound for a random field

Let $(\mathcal{Y}(\boldsymbol{\upsilon}), \boldsymbol{\upsilon} \in \varUpsilon)$ be a random field on a probability space $(\varOmega, \mathcal{F}, \boldsymbol{P})$, where $\varUpsilon$ is a separable locally compact space. For any $\boldsymbol{\upsilon} \in \varUpsilon$ we assume the following exponential moment condition to be fulfilled:

$(\mathcal{E})$ *For every* $\boldsymbol{\upsilon} \in \varUpsilon$

$$\boldsymbol{E} \exp\{\mathcal{Y}(\boldsymbol{\upsilon})\} = 1.$$

The aim of this section is to establish a similar exponential bound for a supremum of $\mathcal{Y}(\boldsymbol{\upsilon})$ over $\boldsymbol{\upsilon} \in \varUpsilon$. A trivial corollary of the condition $(\mathcal{E})$ is that if the set $\varUpsilon$ is finite with $N = \#\varUpsilon$, then

$$\boldsymbol{E} \exp\Bigl\{\sup_{\boldsymbol{\upsilon} \in \varUpsilon} \mathcal{Y}(\boldsymbol{\upsilon})\Bigr\} \le N.$$

Unfortunately, in the general case the supremum of $\mathcal{Y}(\boldsymbol{\upsilon})$ over $\boldsymbol{\upsilon}$ does not necessarily fulfill the condition of bounded exponential moments. We therefore, consider a penalized version of the process $\mathcal{Y}(\boldsymbol{\upsilon})$, that is, we try to bound the exponential moment of $\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})$ for some *penalty* function $\operatorname{pen}(\boldsymbol{\upsilon})$. The goal is to find a possibly minimal such function $\operatorname{pen}(\boldsymbol{\upsilon})$ which provides

$$\boldsymbol{E} \exp\Bigl\{\sup_{\boldsymbol{\upsilon} \in \varUpsilon} \bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr]\Bigr\} \le 1.$$

In the case of a finite set $\varUpsilon$, a natural candidate is $\operatorname{pen}(\boldsymbol{\upsilon}) = \log(\#\varUpsilon)$. Below we show how this simple choice can be extended to the case of a general set $\varUpsilon$. There exists a number of results about a supremum of a centered random field which are heavily based on the theory of empirical processes. See e.g. the monographes van der Vaart and Wellner (1996), Van de Geer (2000), Massart (2007), and references therein. Our approach is a bit different. First the process $\mathcal{Y}(\boldsymbol{\upsilon})$ does not need to be centered, instead we use the normalization $\boldsymbol{E} \exp\{\mathcal{Y}(\boldsymbol{\upsilon})\} = 1$. Secondly we do not assume any particular structure of this process like independence of observations, so the methods of the empirical processes do not apply here. Finally, our analysis is focuses on the penalty function $\operatorname{pen}(\cdot)$ rather then on the deviation probability of $\max_{\boldsymbol{\upsilon}} \mathcal{Y}(\boldsymbol{\upsilon})$.

5.1 A local bound

Define $\mathcal{M}(\boldsymbol{\upsilon}) = \boldsymbol{E}\mathcal{Y}(\boldsymbol{\upsilon})$, $\zeta(\boldsymbol{\upsilon}) = \mathcal{Y}(\boldsymbol{\upsilon}) - \boldsymbol{E}\mathcal{Y}(\boldsymbol{\upsilon})$, and denote $\zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') = \zeta(\boldsymbol{\upsilon}) - \zeta(\boldsymbol{\upsilon}')$ for $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \varUpsilon$. We assume a nonnegative symmetric function $\mathfrak{D}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}')$ is given such that the following condition is fulfilled:

$(\mathcal{E}\epsilon)$ *There exist numbers* $\epsilon > 0$ *and* $\lambda^* > 0$, *such that for any* $\lambda \le \lambda^*$

$$\sup_{\boldsymbol{\upsilon},\boldsymbol{\upsilon}' \in \varUpsilon \,:\, \mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}') \le \epsilon} \log \boldsymbol{E} \exp\Bigl\{ 2\lambda \frac{\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}')}{\mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}')} \Bigr\} \le 2\lambda^2.$$

Let $\epsilon > 0$ be shown in condition $(\mathcal{E}\epsilon)$. Define for any point $\boldsymbol{\upsilon}^\circ \in \varUpsilon$ the "ball"

$$\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ) = \bigl\{ \boldsymbol{\upsilon} : \mathfrak{D}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\circ) \le \epsilon \bigr\}.$$

To state the result, we have to introduce the notion of *local entropy.* We say that a discrete set $\mathcal{D}(\epsilon, \mathcal{C})$ is an ϵ-net in $\mathcal{C} \subseteq \varUpsilon$, if

$$\mathcal{C} \subset \bigcup_{\boldsymbol{\upsilon}^\circ \in \mathcal{D}(\epsilon,\mathcal{C})} \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ). \tag{5.1}$$

By $\mathbb{N}(\epsilon_0, \epsilon, \boldsymbol{\upsilon}^\circ)$ for $\epsilon_0 \le \epsilon$ we denote the local covering number defined as the minimal number of sets $\mathcal{B}(\epsilon_0, \cdot)$ required to cover $\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$. With this covering number we associate the local entropy

$$\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^\circ) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} 2^{-k} \log \mathbb{N}(2^{-k}\epsilon, \epsilon, \boldsymbol{\upsilon}^\circ).$$

Assume that $\boldsymbol{\upsilon}^\circ \in \varUpsilon$ is fixed. The following result controls the supremum in $\boldsymbol{\upsilon}$ of the penalized process $\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})$ over the ball $\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$.

Theorem 5.1. *Assume* $(\mathcal{E})$ *and* $(\mathcal{E}\epsilon)$. *For any* $\varrho \in (0,1)$ *with* $\varrho\epsilon/(1-\varrho) \le \lambda^*$, *any* $\boldsymbol{\upsilon}^\circ \in \varUpsilon$

$$\log \boldsymbol{E} \exp\Bigl\{ \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon,\boldsymbol{\upsilon}^\circ)} \varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr] \Bigr\} \le \frac{2\epsilon^2 \varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^\circ) - \varrho \operatorname{pen}_\epsilon(\boldsymbol{\upsilon}^\circ)$$

with

$$\operatorname{pen}_\epsilon(\boldsymbol{\upsilon}^\circ) = \inf_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon,\boldsymbol{\upsilon}^\circ)} \operatorname{pen}(\boldsymbol{\upsilon}).$$

Proof. We begin with some result which bounds the stochastic component of the process $\mathcal{Y}(\boldsymbol{\upsilon})$ within the local ball $\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$.

Lemma 5.2. *Assume that* $\zeta(\boldsymbol{\upsilon})$ *is a separable process satisfying condition* $(\mathcal{E}\epsilon)$. *Then for any given* $\boldsymbol{\upsilon}^\circ \in \varUpsilon$, $\boldsymbol{\upsilon}^\sharp \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$, *and* $\lambda \le \lambda^*$

$$\log \boldsymbol{E} \exp\Bigl\{ \frac{\lambda}{\epsilon} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon,\boldsymbol{\upsilon}^\circ)} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^\sharp) \Bigr\} \le \mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^\circ) + 2\lambda^2.$$

Proof. The proof is based on the standard chaining argument; see e.g. van der Vaart and Wellner (1996). Without loss of generality, we assume that $\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^{\circ}) < \infty$. Then for any integer $k \geq 0$, there exists a $2^{-k}\epsilon$-net $\mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})$ in the local ball $\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})$ having the cardinality $\mathbb{N}(2^{-k}\epsilon, \epsilon, \boldsymbol{\upsilon}^{\circ})$. Using the nets $\mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})$ with $k = 1, \ldots, K-1$, one can construct a chain connecting an arbitrary point $\boldsymbol{\upsilon}$ in $\mathcal{D}_K(\epsilon, \boldsymbol{\upsilon}^{\circ})$ and $\boldsymbol{\upsilon}^{\sharp}$. It means that one can find points $\boldsymbol{\upsilon}_k \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ}),\ k = 1, \ldots, K-1$, such that $\mathfrak{D}(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}) \leq 2^{-k+1}\epsilon$ for $k = 1, \ldots, K$. Here $\boldsymbol{\upsilon}_K$ means $\boldsymbol{\upsilon}$ and $\boldsymbol{\upsilon}_0$ means $\boldsymbol{\upsilon}^{\sharp}$. Notice that $\boldsymbol{\upsilon}_k$ can be constructed recurrently: $\boldsymbol{\upsilon}_{k-1} = \tau_{k-1}(\boldsymbol{\upsilon}_k),\ k = K, \ldots, 1$, where

$$\tau_{k-1}(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}' \in \mathcal{D}_{k-1}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \mathfrak{D}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}').$$

It obviously holds

$$\zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp}) = \sum_{k=1}^{K} \zeta(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}).$$

It holds for $\xi(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}) = \zeta(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}) / \mathfrak{D}(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1})$ that

$$\zeta(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}) = \mathfrak{D}(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}) \xi(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1}) = 2\epsilon\, c_k\, \xi(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1})$$

with $c_k = \mathfrak{D}(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1})/(2\epsilon) \leq 2^{-k}$. By condition $(\mathcal{E}\epsilon)$ $\log \boldsymbol{E} \exp\bigl\{2\lambda \xi(\boldsymbol{\upsilon}_k, \boldsymbol{\upsilon}_{k-1})\bigr\} \leq 2\lambda^2$. Next,

$$\begin{aligned}
\sup_{\boldsymbol{\upsilon} \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp}) &\leq \sum_{k=1}^{K} \sup_{\boldsymbol{\upsilon}' \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}', \tau_{k-1}(\boldsymbol{\upsilon}')) \\
&\leq 2\epsilon \sum_{k=1}^{K} \sup_{\boldsymbol{\upsilon}' \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})} c_k \xi(\boldsymbol{\upsilon}', \tau_{k-1}(\boldsymbol{\upsilon}')). \qquad (5.2)
\end{aligned}$$

Since $c_k \leq 2^{-k}$, the Hölder inequality and condition $(\mathcal{E}\epsilon)$ imply

$$\begin{aligned}
&\log \boldsymbol{E} \exp\Bigl\{\frac{\lambda}{\epsilon} \sup_{\boldsymbol{\upsilon} \in \mathcal{D}_K(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp})\Bigr\} \leq \log \boldsymbol{E} \exp\Bigl\{2\lambda \sum_{k=1}^{K} \sup_{\boldsymbol{\upsilon}' \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})} c_k \xi(\boldsymbol{\upsilon}', \tau_{k-1}(\boldsymbol{\upsilon}'))\Bigr\} \\
&\leq \sum_{k=1}^{K} 2^{-k} \log\Bigl[\boldsymbol{E} \exp\Bigl\{\sup_{\boldsymbol{\upsilon}' \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})} 2^k c_k \times 2\lambda \xi(\boldsymbol{\upsilon}', \tau_{k-1}(\boldsymbol{\upsilon}'))\Bigr\}\Bigr] \\
&\leq \sum_{k=1}^{K} 2^{-k} \log\Bigl[\sum_{\boldsymbol{\upsilon}' \in \mathcal{D}_k(\epsilon, \boldsymbol{\upsilon}^{\circ})} \boldsymbol{E} \exp\bigl\{2^k c_k \times 2\lambda \xi(\boldsymbol{\upsilon}', \tau_{k-1}(\boldsymbol{\upsilon}'))\bigr\}\Bigr] \\
&\leq \sum_{k=1}^{K} 2^{-k} \bigl\{\log \mathbb{N}(2^{-k}\epsilon, \epsilon, \boldsymbol{\upsilon}^{\circ}) + 2\lambda^2\bigr\}.
\end{aligned}$$

These inequalities and the separability of $\zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp})$ yield

$$\begin{aligned}
\log \boldsymbol{E} \exp\Bigl\{\frac{\lambda}{\epsilon} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp})\Bigr\} &= \lim_{K \to \infty} \log \boldsymbol{E} \exp\Bigl\{\frac{\lambda}{\epsilon} \sup_{\boldsymbol{\upsilon} \in \mathcal{D}_K(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp})\Bigr\} \\
&\leq \sum_{k=1}^{\infty} 2^{-k} \bigl\{2\lambda^2 + \log \mathbb{N}(2^{-k}\epsilon, \epsilon, \boldsymbol{\upsilon}^{\circ})\bigr\} \leq 2\lambda^2 + \mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^{\circ})
\end{aligned}$$

which completes the proof of the lemma. □

Now define for a fixed a point $\boldsymbol{\upsilon}^{\circ}$

$$\boldsymbol{\upsilon}^{\sharp} = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \{\mathcal{M}(\boldsymbol{\upsilon}) + \operatorname{pen}(\boldsymbol{\upsilon})\},$$

where $\mathcal{M}(\boldsymbol{\upsilon}) = -\boldsymbol{E}\mathcal{Y}(\boldsymbol{\upsilon})$. If there are many such points, then take any of them as $\boldsymbol{\upsilon}^{\sharp}$. Obviously

$$\sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \bigl\{\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr\} \le \mathcal{Y}(\boldsymbol{\upsilon}^{\sharp}) - \operatorname{pen}(\boldsymbol{\upsilon}^{\sharp}) + \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp}).$$

Therefore, by the Hölder inequality and Lemma 5.2 with $\lambda = \epsilon\varrho/(1-\varrho)$

$$\begin{aligned}
&\log \boldsymbol{E} \exp\Bigl\{ \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr] \Bigr\} \\
&\quad \le \varrho \log \boldsymbol{E} \exp\bigl\{\mathcal{Y}(\boldsymbol{\upsilon}^{\sharp}) - \operatorname{pen}(\boldsymbol{\upsilon}^{\sharp})\bigr\} + \ (1-\varrho) \log \boldsymbol{E} \exp\Bigl\{ \frac{\varrho}{1-\varrho} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})} \zeta(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\sharp}) \Bigr\} \\
&\quad \le 2\epsilon^2 \varrho^2/(1-\varrho) + (1-\varrho)\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^{\circ}) - \varrho \operatorname{pen}(\boldsymbol{\upsilon}^{\sharp}) \\
&\quad \le 2\epsilon^2 \varrho^2/(1-\varrho) + (1-\varrho)\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^{\circ}) - \varrho \operatorname{pen}_{\epsilon}(\boldsymbol{\upsilon}^{\circ}).
\end{aligned}$$

which is the assertion of the theorem. □

5.2 A global exponential bound for the penalized process

This section presents some sufficient conditions on the penalty function $\operatorname{pen}(\boldsymbol{\upsilon})$ which ensure the general exponential bound for the penalized process $\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})$. For simplicity we assume that the local entropy numbers $\mathbb{Q}(\epsilon, \boldsymbol{\upsilon})$ are uniformly bounded by a constant $\mathbb{Q}^*(\Upsilon)$. Let also π be a σ-finite measure on the space Υ and $\pi(A)$ stand for the π-measure of a set $A \subset \Upsilon$. The standard proposal for π is the usual Lebesgue measure.

Theorem 5.3. *Assume* $(\mathcal{E})$ *and* $(\mathcal{E}\epsilon)$ *with some fixed* ϵ *and* λ^*. *Let* $\varrho < 1$ *be such that* $\varrho\epsilon/(1-\varrho) \le \lambda^*$. *Let also* $\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}) \le \mathbb{Q}^*(\Upsilon)$ *for all* $\boldsymbol{\upsilon} \in \Upsilon$. *Let a* σ*-finite measure* π *on* Υ *be such that for some* $\nu \ge 1$

$$\sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \colon \mathfrak{D}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \le \epsilon} \frac{\pi(\mathcal{B}(\epsilon, \boldsymbol{\upsilon}))}{\pi(\mathcal{B}(\epsilon, \boldsymbol{\upsilon}'))} \le \nu. \tag{5.3}$$

Finally, let a function $\operatorname{pen}(\boldsymbol{\upsilon})$ *satisfy*

$$\mathfrak{H}_{\epsilon}(\varrho) \overset{\mathrm{def}}{=} \log \int_{\Upsilon} \frac{1}{\pi(\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ}))} \exp\bigl\{-\varrho \operatorname{pen}_{\epsilon}(\boldsymbol{\upsilon}^{\circ})\bigr\} d\pi(\boldsymbol{\upsilon}^{\circ}) < \infty$$

with $\operatorname{pen}_{\epsilon}(\boldsymbol{\upsilon}^{\circ}) = \inf_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon,\boldsymbol{\upsilon}^{\circ})} \operatorname{pen}(\boldsymbol{\upsilon})$ *. Then*

$$\boldsymbol{E} \exp\Bigl\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon} \varrho \bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon}) \bigr] \Bigr\} \le \mathfrak{Q}(\varrho, \epsilon), \tag{5.4}$$

where

$$\log \mathfrak{Q}(\varrho, \epsilon) = \frac{2\epsilon^2 \varrho^2}{1 - \varrho} + (1 - \varrho) \mathbb{Q}^*(\varUpsilon) + \log \nu + \mathfrak{H}_{\epsilon}(\varrho). \tag{5.5}$$

Proof. We begin with a simple technical result which bounds the maximum of a given function via the weighted integral of the local maxima.

Lemma 5.4. *Let* $f(\boldsymbol{\upsilon})$ *be a nonnegative function on* $\varUpsilon \subset I\!\!R^p$ *and let for every point* $\boldsymbol{\upsilon} \in \varUpsilon$ *a vicinity* $A(\boldsymbol{\upsilon})$ *be fixed such that* $\boldsymbol{\upsilon}' \in A(\boldsymbol{\upsilon})$ *implies* $\boldsymbol{\upsilon} \in A(\boldsymbol{\upsilon}')$ *. Let also the measure* $\pi\bigl(A(\boldsymbol{\upsilon})\bigr)$ *of the set* $A(\boldsymbol{\upsilon})$ *fulfill for every* $\boldsymbol{\upsilon}^{\circ} \in \varUpsilon$

$$\sup_{\boldsymbol{\upsilon} \in A(\boldsymbol{\upsilon}^{\circ})} \frac{\pi\bigl(A(\boldsymbol{\upsilon})\bigr)}{\pi\bigl(A(\boldsymbol{\upsilon}^{\circ})\bigr)} \le \nu. \tag{5.6}$$

Then

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon} f(\boldsymbol{\upsilon}) \le \nu \int_{\varUpsilon} f^*(\boldsymbol{\upsilon}) \frac{1}{\pi\bigl(A(\boldsymbol{\upsilon})\bigr)} d\pi(\boldsymbol{\upsilon})$$

with

$$f^*(\boldsymbol{\upsilon}) \overset{\text{def}}{=} \sup_{\boldsymbol{\upsilon}' \in A(\boldsymbol{\upsilon})} f(\boldsymbol{\upsilon}').$$

Proof. For every $\boldsymbol{\upsilon}^{\circ} \in \varUpsilon$

$$\begin{aligned} \int_{\varUpsilon} f^*(\boldsymbol{\upsilon}) \frac{1}{\pi\bigl(A(\boldsymbol{\upsilon})\bigr)} d\pi(\boldsymbol{\upsilon}) &\ge \int_{A(\boldsymbol{\upsilon}^{\circ})} f^*(\boldsymbol{\upsilon}) \frac{1}{\pi\bigl(A(\boldsymbol{\upsilon})\bigr)} d\pi(\boldsymbol{\upsilon}) \\ &\ge f(\boldsymbol{\upsilon}^{\circ}) \int_{A(\boldsymbol{\upsilon}^{\circ})} \frac{1}{\pi\bigl(A(\boldsymbol{\upsilon})\bigr)} d\pi(\boldsymbol{\upsilon}) \end{aligned}$$

because $\boldsymbol{\upsilon} \in A(\boldsymbol{\upsilon}^{\circ})$ implies $\boldsymbol{\upsilon}^{\circ} \in A(\boldsymbol{\upsilon})$ and hence, $f(\boldsymbol{\upsilon}^{\circ}) \le f^*(\boldsymbol{\upsilon})$. Now by (5.6)

$$\int_{\varUpsilon} f^*(\boldsymbol{\upsilon}) \frac{1}{\pi\bigl(A(\boldsymbol{\upsilon})\bigr)} d\pi(\boldsymbol{\upsilon}) \ge \frac{f(\boldsymbol{\upsilon}^{\circ})}{\nu} \int_{A(\boldsymbol{\upsilon}^{\circ})} \frac{1}{\pi\bigl(A(\boldsymbol{\upsilon}^{\circ})\bigr)} d\pi(\boldsymbol{\upsilon}) = f(\boldsymbol{\upsilon}^{\circ})/\nu$$

as required. □

This result applied to $f(\boldsymbol{\upsilon}) = \exp\bigl\{\varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr]\bigr\}$ and $A(\boldsymbol{\upsilon}) = \mathcal{B}(\epsilon, \boldsymbol{\upsilon})$ implies

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon} \exp\Bigl\{\varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr]\Bigr\} \le \nu \int_{\varUpsilon} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon,\boldsymbol{\upsilon}^{\circ})} \exp\Bigl\{\varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr]\Bigr\} \frac{d\pi(\boldsymbol{\upsilon}^{\circ})}{\pi\bigl(\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ})\bigr)}.$$

This implies by Theorem 5.1

$$\begin{aligned}
&\log \boldsymbol{E} \sup_{\boldsymbol{\upsilon} \in \varUpsilon} \exp\Bigl\{ \varrho \bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon}) \bigr] \Bigr\} \\
&\quad \le \frac{2\epsilon^2 \varrho^2}{1-\varrho} + (1-\varrho) \mathbb{Q}^*(\varUpsilon) + \log \Bigl\{ \nu \int_{\varUpsilon} \exp\bigl\{ -\varrho \operatorname{pen}_{\epsilon}(\boldsymbol{\upsilon}^{\circ}) \bigr\} \frac{d\pi(\boldsymbol{\upsilon}^{\circ})}{\pi\bigl(\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ}) \bigr)} \Bigr\} \\
&\quad \le \frac{2\epsilon^2 \varrho^2}{1-\varrho} + (1-\varrho) \mathbb{Q}^*(\varUpsilon) + \log(\nu) + \mathfrak{H}_{\epsilon}(\varrho)
\end{aligned}$$

and the assertion follows. □

5.3 Smooth case

Here we discuss the special case when $\varUpsilon \subset I\!R^p$, the process $\mathcal{Y}(\boldsymbol{\upsilon})$ and its stochastic component $\zeta(\boldsymbol{\upsilon})$ are absolutely continuous and the gradient $\nabla \zeta(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} d\zeta(\boldsymbol{\upsilon})/d\boldsymbol{\upsilon}$ has bounded exponential moments. We also assume that π is the Lebesgue measure on $\varUpsilon$. Suppose the following condition is fulfilled:

$(\mathcal{E}D)$ *There exist $\lambda^* > 0$ and for each $\boldsymbol{\upsilon} \in \varUpsilon$, a symmetric non-negative matrix $H(\boldsymbol{\upsilon})$ such that for any $\lambda \le \lambda^*$*

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon} \sup_{\boldsymbol{\gamma} \in S^p} \log \boldsymbol{E} \exp\Bigl\{ 2\lambda \frac{\boldsymbol{\gamma}^{\top} \nabla \zeta(\boldsymbol{\upsilon})}{\| H(\boldsymbol{\upsilon}) \boldsymbol{\gamma} \|} \Bigr\} \le 2\lambda^2 .$$

The matrix function $H(\boldsymbol{\upsilon})$ can be used for defining a natural topology in $\varUpsilon$. Namely, for any $\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' \in \varUpsilon$ define $\mathfrak{d} = \| \boldsymbol{\upsilon} - \boldsymbol{\upsilon}' \|$, $\boldsymbol{\gamma} = (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}')/\mathfrak{d}$ and

$$\mathfrak{D}^2(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \stackrel{\mathrm{def}}{=} \| \boldsymbol{\upsilon} - \boldsymbol{\upsilon}' \|^2 \int_0^1 \boldsymbol{\gamma}^{\top} H^2(\boldsymbol{\upsilon} + t \mathfrak{d} \boldsymbol{\gamma}) \boldsymbol{\gamma} \, dt .$$

Next, introduce for each $\boldsymbol{\upsilon}^{\circ} \in \varUpsilon$ and $\epsilon > 0$ the set

$$\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^{\circ}) \stackrel{\mathrm{def}}{=} \{ \boldsymbol{\upsilon} : \mathfrak{D}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^{\circ}) \le \epsilon \}$$

To state the result, we need one more condition on the uniform continuity of the matrix $H(\boldsymbol{\upsilon})$ in $\boldsymbol{\upsilon}$.

$(\boldsymbol{H})$ *There exist constants $\epsilon > 0$ and $\nu_1 \ge 1$ such that*

$$\sup_{\boldsymbol{\upsilon}, \boldsymbol{\upsilon}' : \mathfrak{D}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}') \le \epsilon} \ \sup_{\boldsymbol{\gamma} \in S^p} \frac{\boldsymbol{\gamma}^{\top} H^2(\boldsymbol{\upsilon}) \boldsymbol{\gamma}}{\boldsymbol{\gamma}^{\top} H^2(\boldsymbol{\upsilon}') \boldsymbol{\gamma}} \le \nu_1 .$$

Theorem 5.5. *Let $(\mathcal{E})$ be satisfied. Suppose that $(\mathcal{E}D)$ holds with some λ^* and a matrix function $H(\boldsymbol{\upsilon})$ which fulfills (H). If for some $\varrho \in (0,1)$ and $\epsilon > 0$ with $\varrho\epsilon/(1-\varrho) \le \lambda^*$, the penalty function $\operatorname{pen}(\boldsymbol{\upsilon})$ fulfills*

$$\mathfrak{H}_{\epsilon}(\varrho) \stackrel{\mathrm{def}}{=} \log \Bigl\{ \omega_p^{-1} \epsilon^{-p} \int_{\varUpsilon} \det(H(\boldsymbol{\upsilon}^{\circ})) \exp\bigl\{ -\varrho \operatorname{pen}_{\epsilon}(\boldsymbol{\upsilon}^{\circ}) \bigr\} d\boldsymbol{\upsilon}^{\circ} \Bigr\} < \infty$$

with $\operatorname{pen}_{\epsilon}(\boldsymbol{\upsilon}^{\circ}) = \inf_{\boldsymbol{\upsilon}\in\mathcal{B}(\epsilon,\boldsymbol{\upsilon}^{\circ})} \operatorname{pen}(\boldsymbol{\upsilon})$ *, then*

$$\boldsymbol{E}\exp\Bigl\{\sup_{\boldsymbol{\upsilon}\in\varUpsilon} \varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \operatorname{pen}(\boldsymbol{\upsilon})\bigr]\Bigr\} \le \mathfrak{Q}(\varrho,\epsilon) \tag{5.7}$$

where

$$\log \mathfrak{Q}(\varrho,\epsilon) \quad = \quad \frac{2\epsilon^2\varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}_p + \mathfrak{H}_{\epsilon}(\varrho) + p\log(\nu_1)$$

with $\mathbb{Q}_p$ *being the usual entropy number for the Euclidean ball in* $I\!\!R^p$ *.*

Proof. First we show that the differentiability condition $(\mathcal{E}D)$ implies the local moment condition $(\mathcal{E}\epsilon)$.

Lemma 5.6. *Assume that* $(\mathcal{E}D)$ *holds with some* λ^* *. Then for any* $\boldsymbol{\upsilon}^{\circ} \in \varUpsilon$ *and any* λ *with* $|\lambda| \le \lambda^*/\nu_1^{1/2}$ *, it holds*

$$\sup_{\boldsymbol{\upsilon}\in\mathcal{B}(\epsilon,\boldsymbol{\upsilon}^{\circ})} \log \boldsymbol{E}\exp\biggl\{2\lambda\frac{\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}{\mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}\biggr\} \le 2\lambda^2. \tag{5.8}$$

Proof. For $\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon,\boldsymbol{\upsilon}^{\circ})$, denote $\mathfrak{d} = \|\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ}\|$, $\boldsymbol{\gamma} = (\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ})/\mathfrak{d}$. With this notation

$$\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ}) = \mathfrak{d}\boldsymbol{\gamma}^{\top}\int_0^1 \nabla\zeta(\boldsymbol{\upsilon}^{\circ}+t\mathfrak{d}\boldsymbol{\gamma})dt.$$

The condition (H) implies for every $t \in [0,1]$ that

$$\lambda\frac{u\|H(\boldsymbol{\upsilon}^{\circ}+t\mathfrak{d}\boldsymbol{\gamma})\boldsymbol{\gamma}\|}{\mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})} \le \lambda\nu_1^{1/2} \le \lambda^*.$$

Now the Hölder inequality and $(\mathcal{E}D)$ yield

$$\begin{aligned}
&\log \boldsymbol{E}\exp\biggl\{2\lambda\frac{\zeta(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}{\mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})} - \lambda^2\biggr\} \\
&\quad = \log \boldsymbol{E}\exp\biggl\{\int_0^1 \boldsymbol{\gamma}^{\top}\biggl[\frac{2\lambda\mathfrak{d}}{\mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}\nabla\zeta(\boldsymbol{\upsilon}^{\circ}+t\mathfrak{d}\boldsymbol{\gamma}) - \frac{2\lambda^2\mathfrak{d}^2}{\mathfrak{D}^2(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}H^2(\boldsymbol{\upsilon}^{\circ}+t\mathfrak{d}\boldsymbol{\gamma})\boldsymbol{\gamma}\biggr]dt\biggr\} \\
&\quad \le \int_0^1 \log \boldsymbol{E}\exp\biggl\{\boldsymbol{\gamma}^{\top}\biggl[\frac{2\lambda\mathfrak{d}}{\mathfrak{D}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}\nabla\zeta(\boldsymbol{\upsilon}^{\circ}+t\mathfrak{d}\boldsymbol{\gamma}) - \frac{2\lambda^2\mathfrak{d}^2}{\mathfrak{D}^2(\boldsymbol{\upsilon},\boldsymbol{\upsilon}^{\circ})}H^2(\boldsymbol{\upsilon}^{\circ}+t\mathfrak{d}\boldsymbol{\gamma})\boldsymbol{\gamma}\biggr]\biggr\}\,dt \\
&\quad \le 0
\end{aligned}$$

as required. □

Next we show that condition (H) implies (5.3). Consider for every $\boldsymbol{\upsilon}^{\circ} \in \varUpsilon$ an elliptic neighborhood $\mathcal{B}'(\epsilon,\boldsymbol{\upsilon}^{\circ}) = \{\boldsymbol{\upsilon} : \|H(\boldsymbol{\upsilon}^{\circ})(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}^{\circ})\| \le \epsilon\}$.

Lemma 5.7. *Assume* (H)*. Then*

1. *for any $\epsilon > 0$ and any $\boldsymbol{\upsilon} \in \varUpsilon$*

$$\begin{array}{ccccc} \mathcal{B}'(\nu_1^{-1/2}\epsilon, \boldsymbol{\upsilon}) & \subset & \mathcal{B}(\epsilon, \boldsymbol{\upsilon}) & \subset & \mathcal{B}'(\nu_1^{1/2}\epsilon, \boldsymbol{\upsilon}), \\ \mathcal{B}(\nu_1^{-1/2}\epsilon, \boldsymbol{\upsilon}) & \subset & \mathcal{B}'(\epsilon, \boldsymbol{\upsilon}) & \subset & \mathcal{B}(\nu_1^{1/2}\epsilon, \boldsymbol{\upsilon}). \end{array} \tag{5.9}$$

2. *For every $\boldsymbol{\upsilon} \in \varUpsilon$,*

$$\nu_1^{-p/2} \le \epsilon^{-p}\pi(\mathcal{B}(\epsilon, \boldsymbol{\upsilon}))\det(H(\boldsymbol{\upsilon}))/\omega_p \le \nu_1^{p/2}, \tag{5.10}$$

where ω_p is the Lebesgue measure of the unit ball in $I\!\!R^p$.

3. *condition (5.3) holds with $\nu = \nu_1^p$.*

Proof. Condition (H) implies that for any $\boldsymbol{\upsilon}^\circ \in \varUpsilon$ and $\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$ that

$$\nu_1^{-1}\boldsymbol{\gamma}^\top H^2(\boldsymbol{\upsilon}^\circ)\boldsymbol{\gamma} \le \int_0^1 \boldsymbol{\gamma}^\top H^2(\boldsymbol{\upsilon}^\circ + t\mathfrak{d}\boldsymbol{\gamma})\boldsymbol{\gamma}\, dt \le \nu_1\boldsymbol{\gamma}^\top H^2(\boldsymbol{\upsilon}^\circ)\boldsymbol{\gamma}$$

with $\mathfrak{d} = \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ\|$ and $\boldsymbol{\gamma} = (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^\circ)/\mathfrak{d}$, which yields the first assertion of the lemma. The Lebesgue measure of the ellipsoid $\mathcal{B}'(\epsilon, \boldsymbol{\upsilon})$ is equal to $\omega_p\epsilon^p/\det(H(\boldsymbol{\upsilon}))$. This and (5.9) imply the second assertion. This, in turns, implies (5.3) in view of (H). □

The next result claims that in the smooth case the local entropy number $\mathbb{Q}(\epsilon, \boldsymbol{\upsilon}^\circ)$ is similar to the usual Euclidean situation.

Lemma 5.8. *Assume (H). Then* $\sup_{\boldsymbol{\upsilon}\in\varTheta} \mathbb{Q}(\epsilon, \boldsymbol{\upsilon}) \le \mathbb{Q}_p + p\log(\nu_1)$.

Proof. Fix any $\boldsymbol{\upsilon}^\circ \in \varUpsilon$. Linear transformation with the matrix $H^{-1}(\boldsymbol{\upsilon}^\circ)$ reduces the situation to the case when $H(\boldsymbol{\upsilon}^\circ) \equiv I$ and $\mathcal{B}'(\epsilon_0, \boldsymbol{\upsilon}^\circ)$ is a usual Euclidean ball for any $\epsilon_0 \le \epsilon$. Moreover, by (H), each elliptic set $\mathcal{B}'(\epsilon_0, \boldsymbol{\upsilon})$ for $\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$ is nearly an Euclidean ball in the sense that the ratio of its largest and smallest axes (which is the ratio of the largest and smallest eigenvalues of $H^{-1}(\boldsymbol{\upsilon}^\circ)H^2(\boldsymbol{\upsilon})H^{-1}(\boldsymbol{\upsilon}^\circ)$) is bounded by ν_1. Therefore, for any $\epsilon_0 \le \epsilon$, a Euclidean net $\mathcal{D}^e(\epsilon_0/\nu_1)$ with the step ϵ_0/ν_1 ensures a covering of $\mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$ by the sets $\mathcal{B}(\epsilon_0, \boldsymbol{\upsilon}^\circ)$, $\boldsymbol{\upsilon}^\circ \in \mathcal{D}^e(\epsilon_0/\nu_1)$. Therefore, the corresponding covering number is bounded by $(\nu_1\epsilon/\epsilon_0)^p$ yielding the claimed bound for the local entropy. □

Now the result of theorem 5.5 is reduced to the statement of Theorem 5.3. □

Computing of the penalty simplifies a lot when the matrix $H(\boldsymbol{\upsilon})$ is uniformly bounded by a matrix H^*, or, equivalently, condition (H) is fulfilled for $H(\boldsymbol{\upsilon}) \equiv H^*$. Then one can define $\operatorname{pen}(\boldsymbol{\upsilon})$ as a function of the norm $\|H^*(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\|$ for a fixed $\boldsymbol{\upsilon}_0$.

Theorem 5.9. *Assume additionally to the conditions of Theorem 5.5 that* $H(\boldsymbol{\upsilon}) \le H^*$ *for a symmetric matrix* H^* *. Suppose that* $\varkappa(t)$ *is a monotonously decreasing positive function on* $[0,+\infty)$ *satisfying*

$$\mathfrak{P}^* \stackrel{\mathrm{def}}{=} \omega_p^{-1} \int_{I\!\!R^p} \varkappa(\|\boldsymbol{u}\|) d\boldsymbol{u} = p \int_0^\infty \varkappa(t) t^{p-1} dt < \infty. \tag{5.11}$$

Define

$$\mathrm{pen}(\boldsymbol{\upsilon}) = -\varrho^{-1} \log \varkappa\bigl(\epsilon^{-1}\|H^*(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_0)\| + 1\bigr)$$

Then

$$\boldsymbol{E} \exp\Bigl\{ \sup_{\boldsymbol{\upsilon}\in\Upsilon} \varrho\bigl[\mathcal{Y}(\boldsymbol{\upsilon}) - \mathrm{pen}(\boldsymbol{\upsilon})\bigr] \Bigr\} \le \mathfrak{Q}(\varrho,\epsilon) \tag{5.12}$$

with

$$\log \mathfrak{Q}(\varrho,\epsilon) \quad = \quad \frac{2\epsilon^2\varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}_p + \log(\mathfrak{P}^*),$$

where ω_p *is the volume of the unit ball in* $I\!\!R^p$ *.*

Proof. Let us fix $\boldsymbol{\upsilon}^\circ \in \Upsilon$. Definition of the semi-metric $\mathfrak{D}$ and condition (H) imply for every $\boldsymbol{\upsilon} \in \mathcal{B}(\epsilon, \boldsymbol{\upsilon}^\circ)$ that

$$\|H^*(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon})\| \le \epsilon.$$

The triangle inequality and (H) now imply for this $\boldsymbol{\upsilon}$ that

$$\epsilon^{-1}\|H^*(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_0)\| + 1 \ge \epsilon^{-1}\|H^*(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}_0)\|$$

and $\mathrm{pen}_\epsilon(\boldsymbol{\upsilon}^\circ) \ge -\varrho^{-1}\log\varkappa\bigl(\epsilon^{-1}\|H^*(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}_0)\|\bigr)$. Therefore, it follows by change of variables $\boldsymbol{u} = \epsilon H^*(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)$ that

$$\begin{aligned} \omega_p^{-1}\epsilon^{-p} \int_{\Upsilon} \det(H^*) \exp\bigl\{-\varrho\, \mathrm{pen}_\epsilon(\boldsymbol{\upsilon})\bigr\} d\boldsymbol{\upsilon} \quad &\le \quad \omega_p^{-1} \int_{I\!\!R^p} \varkappa(\|\boldsymbol{u}\|) d\boldsymbol{u} \\ &\le \quad p \int_0^\infty \varkappa(t) t^{p-1} dt = \mathfrak{P}^*, \end{aligned}$$

and the result follows from Theorem 5.5. □

Natural candidates for the function $\varkappa(\cdot)$ and the corresponding $\mathfrak{P}^*$-values are:

$$\begin{aligned} \varkappa_1(t) \;&=\; e^{-\delta_1(t-1)_+^2}, & \mathfrak{P}_1^* \;&=\; 1 + \omega_p^{-1}(\pi/\delta_1)^{p/2}, \\ \varkappa_2(t) \;&=\; \|1+t\|^{-p-\delta_2}, & \mathfrak{P}_2^* \;&=\; p/\delta_2\,, \end{aligned}$$

where $\delta_1, \delta_2 > 0$ are some constants. The result of Theorem 5.9 yields

Corollary 5.10. *Under conditions of Theorem 5.9, the bound (5.12) holds with*

$$\begin{aligned}
\operatorname{pen}_1(\boldsymbol{\upsilon}) &= \varrho^{-1}\delta_1\,\epsilon^{-2}\|H^*(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_0)\|^2, \\
\log\mathfrak{Q}_2(\varrho,\epsilon) &= \frac{2\epsilon^2\varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}_p + \log(1+\omega_p^{-1}|\pi/\delta_1|^{p/2}). \\
\operatorname{pen}_1(\boldsymbol{\upsilon}) &= -\varrho^{-1}(p+\delta_2)\log\bigl(\epsilon^{-1}\|H^*(\boldsymbol{\upsilon}-\boldsymbol{\upsilon}_0)\|+2\bigr), \\
\log\mathfrak{Q}_1(\varrho,\epsilon) &= \frac{2\epsilon^2\varrho^2}{1-\varrho} + (1-\varrho)\mathbb{Q}_p + \log(p/\delta_2),
\end{aligned}$$

Sometimes it is useful to combine the functions $\varkappa_1(\cdot)$ and $\varkappa_2(\cdot)$ in the form

$$\varkappa(t) = \varkappa_1(t)\mathbf{1}(t \ge r) + \varkappa_2(t)\mathbf{1}(t \le r) \tag{5.13}$$

for a properly selected r which still ensures (5.11) with

$$\mathfrak{P}^* \le \omega_p^{-1}|\pi/\delta_1|^{p/2} + pr^{-\delta_2}/\delta_2.$$

References

Anderson, T. (1994). *The statistical analysis of time series.* Wiley Classics Library. Chichester: John Wiley &amp; Sons Ltd. xiv, 704 p. .

Basawa, I. V. and Brockwell, P. J. (1984). Asymptotic conditional inference for regular nonergodic models with an application to autoregressive processes. *Ann. Statist.*, 12(1):161–171.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, 31:307–327.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods. 2nd ed.* Springer Series in Statistics. Berlin etc.: Springer-Verlag.

Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.*, 16(1):367–401.

Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Ann. Statist.*, 23(4):1102–1129.

Chen, K., Hu, I., and Ying, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Ann. Statist.*, 27(4):1155–1163.

Chen, Y. and Spokoiny, V. (2007). Modeling and estimation for nonstationary time series with applications to robust risk management. *Econometrics Journal.*

Cox, D. D. and Llatas, I. (1991). Maximum likelihood type estimation for nearly nonstationary autoregressive time series. *Ann. Statist.*, 19(3):1109–1128.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrika*, 49(4):1057–1072.

Fan, J. and Yao, Q. (2003). *Nonlinear time series. Nonparametric and parametric methods.* Springer Series in Statistics. New York, Springer .

Fokianos, K. and Kedem, B. (2003). Regression theory for categorical time series. *Stat. Sci.*, 18(3):357–376.

Fountis, N. G. and Dickey, D. A. (1989). Testing for a unit root nonstationarity in multivariate autoregressive time series. *Ann. Statist.*, 17(1):419–428.

Francq, C. and Zakoian, J.-M. (2007). Quasi-maximum likelihood estimation in GARCH processes when some coefficients are equal to zero. *Stochastic Process. Appl.*, 117(9):1265–1284.

Giacomini, E., Härdle, W., and Spokoiny, V. (2007). Inhomogeneous dependency modelling with time varying copulae. *Journal of Business and Economic Statistics.*

Golubev, Y. and Spokoiny, V. (2009). Preprint, WIAS, Berlin.

Green, P. J. and Silverman, B. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* London: Chapman & Hall. .

Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models.* Oxford: Oxford Univ. Press.

Johansen, S. (2002). A small sample correction for the test of cointegrating rank in the vector autoregressive model. *Econometrica*, 70(5):1929–1961.

Kedem, B. and Fokianos, K. (2002). *Regression models for time series analysis.* Hoboken, Wiley. .

Koenker, R. (2005). *Quantile regression.* Cambridge University Press.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *J. Am. Stat. Assoc.*, 101(475):980–990.

Konev, V. and Pergamenshchikov, S. (1997). On guaranteed estimation of the mean of an autoregressive process. *Ann. Statist.*, 25(5):2127–2163.

Koul, H. L. and Saleh, A. K. M. E. (1993). R-estimation of the parameters of autoregressive [AR(p)] models. *Ann. Statist.*, 21(1):534–551.

Lai, T. L. and Siegmund, D. (1983). Fixed accuracy estimation of an autoregressive parameter. *Ann. Statist.*, 11(2):478–485.

Lee, S.-W. and Hansen, B. E. (1994). Asymptotic theory for the GARCH$(1,1)$ quasi-maximum likelihood estimator. *Econometric Theory*, 10(1):29–52.

Massart, P. (2007). *Concentration inequalities and model selection. Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003.* Lecture Notes in Mathematics. Springer.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models. 2nd ed.* Monographs on Statistics and Applied Probability. 37. London etc.: Chapman and Hall. xix, 511 p. .

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370.

Phillips, P. C. and Xu, K.-L. (2006). Inference in autoregression under heteroscedasticity. *J. Time Ser. Anal.*, 27(2):289–308.

Sentana, E. (1995). Quadratic ARCH models. *Rev. Econ. Stud.*, 62(4):639–661.

Shiryaev, A. N. and Spokoiny, V. G. (1997). On sequential estimation of an autoregressive parameter. *Stochastics Stochastics Rep.*, 60(3-4):219–240.

Spokoiny, V. (2007). Multiscale local change point detection with with applications to value-at-risk. *Ann. Statist.*

Sriram, T. N. (1987). Sequential estimation of the mean of a first-order stationary autoregressive process. *Ann. Statist.*, 15(3):1079–1090.

Sun, J., Loader, C., and McCormick, W. P. (2000). Confidence bands in generalized linear models. *Ann. Statist.*, 28(2):429–460.

Sun, Y. and Stengos, T. (2006). Semiparametric efficient adaptive estimation of asymmetric GARCH models. *J. Econometrics*, 133(1):373–386.

Van de Geer, S. A. (2000). *Applications of empirical process theory.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

van der Vaart, A. and Wellner, J. A. (1996). *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. New York, Springer.

Čížek, P., Härdle, W., and Spokoiny, V. (2007). Adaptive pointwise estimation in time-inhomogeneous time-series models. *Econometrics Journal.*

Wang, C. W. H. (1986). A minimum distance estimator for first-order autoregressive processes. *Ann. Statist.*, 14(3):1180–1193.