Brazilian Journal of Probability and Statistics
2011, Vol. 25, No. 3, 362–391
DOI: 10.1214/11-BJPS148

Prediction-based estimating functions: review and new developments

Michael Sørensen

University of Copenhagen

Abstract. The general theory of prediction-based estimating functions for stochastic process models is reviewed and extended. Particular attention is given to optimal estimation, asymptotic theory and Gaussian processes. Several examples of applications are presented. In particular, partial observation of a system of stochastic differential equations is discussed. This includes diffusions observed with measurement errors, integrated diffusions, stochastic volatility models, and hypoelliptic stochastic differential equations. The Pearson diffusions, for which explicit optimal prediction-based estimating functions can be found, are briefly presented.

1 Introduction

Prediction-based estimating functions were proposed in Sørensen (2000) as a generalization of martingale estimating functions. While martingale estimating functions provide a simple and often quite efficient estimation method for Markovian models (see, e.g., Sørensen (2009, 2011)), they can usually not be applied to non-Markovian models such as stochastic volatility models, compartment models and other partially observed systems. The reason is that in most cases it is impossible to find tractable martingales. The prediction-based estimating functions provide a useful alternative to the martingale estimating functions for non-Markovian models.

A prediction-based estimating function is essentially a sum of weighted prediction errors. An estimator is given as the parameter value for which the prediction errors are small in a particular sense. The methodology is closely related to the method of prediction error estimation that is used in the stochastic control literature; see, for example, Ljung and Caines (1979). In the present paper we review the theory of prediction-based estimating functions developed over the last decade and extend the theory. In particular, the asymptotic theory is extended, and results for Gaussian processes are derived.

Key words and phrases. Asymptotic normality, consistency, diffusion with measurement errors, Gaussian process, integrated diffusion, linear predictors, non-Markovian models, optimal estimating function, partially observed system, Pearson diffusion, statistical inference for stochastic processes, stochastic differential equation, stochastic volatility model, superposition of diffusions.

Received January 2011; accepted March 2011.

In Section 2, general prediction-based estimating functions are presented with particular emphasis on finite-dimensional predictor-spaces, which is the most useful type in practice. The estimating functions considered in the present paper are slightly more general than those in the original paper in order to provide more flexibility in applications. Optimal prediction-based estimating functions are derived in Section 3, and Section 4 presents the asymptotic statistical theory for prediction-based estimating functions. The asymptotic results presented here are stronger than those in Sørensen (2000). The theory covers the more general estimating functions considered in this paper and includes a result on asymptotic uniqueness of the estimator. A general theory for Gaussian models is presented in Section 5. The results in this section are new. In Section 6 we briefly present the class of Pearson diffusions. This is a versatile class of stochastic differential equation models for which explicit optimal prediction-based estimating functions can be found. Finally, a number of applications of the methodology to partially observed systems of stochastic differential equations are discussed in Section 7. The examples include diffusion processes observed with measurement errors, sums of diffusion processes, integrated diffusions, and stochastic volatility models. It is shown how explicit prediction-based estimating functions can be obtained if Pearson diffusions are used as basic building blocks in these models.

2 Prediction-based estimating functions

Prediction-based estimating functions provide a versatile method for parametric inference that is applicable to observations $Y_1, Y_2, \ldots, Y_n$ from general d-dimensional stochastic processes. We assume that the data are observations from a class of stochastic process models parametrized by a p-dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, which we wish to estimate. Expectation under the model with parameter θ will be denoted by $E_\theta(\cdot)$.

First we give a couple of examples to illustrate the scope of the methodology.

Example 2.1. Let X be a D-dimensional diffusion process given as the solution to the stochastic differential equation

$$dX_t = b(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dW_t, \tag{2.1}$$

where σ is a $D \times D$-matrix and W a D-dimensional standard Wiener process. One type of data is partial observations of the system at discrete time points $t_1 < t_2 < \cdots < t_n$:

$$Y_i = k(X_{t_i}) + Z_i, \qquad i = 1, \ldots, n, \tag{2.2}$$

where k is a function with values in $\mathbb{R}^d$, $d \leq D$, and where the d-dimensional measurement errors Z_i are independent and identically distributed and independent of X. Another type of data is

$$Y_i = \int_{t_{i-1}}^{t_i} k(X_s)\,ds + Z_i, \qquad i = 1, \ldots, n, \tag{2.3}$$

with $t_0 = 0$. In both cases typical examples of the function k are $k(x) = x_1$ or $k(x) = x_1 + \cdots + x_D$, where x_i denotes the ith coordinate of x. For both types of data, the observed process is non-Markovian, which makes likelihood inference complicated and martingale estimating functions infeasible in practice.

An estimating function is a p-dimensional function $G_n(\theta)$ that depends on the parameter, θ, as well as on the observations. The dependence on the data is usually suppressed in the notation. An estimator is obtained by solving the equation $G_n(\theta) = 0$ with respect to θ, provided of course that a solution exists (0 denotes the p-dimensional zero-vector). In the statistics literature the theory of estimating functions dates back to the papers by Godambe (1960) and Durbin (1960). A modern survey of the statistical theory of estimating functions can be found in Heyde (1997). There has been a parallel development in the econometrics literature, where the foundation was laid in Hansen (1982) and Hansen (1985). A discussion of links between the econometrics and statistics literature can be found in Hansen (2000) and Sørensen (2011).

A prediction-based estimating function is essentially a sum of weighted prediction errors, and the idea is to choose as the estimator the parameter value that eliminates this sum of prediction errors. What is predicted are N real-valued functions of $s + 1$ consecutive observations ($s \geq 0$) and the parameter θ,

$$f_j(Y_i, \ldots, Y_{i-s}; \theta), \qquad j = 1, \ldots, N,$$

satisfying that

$$E_\theta(f_j(Y_i, \ldots, Y_{i-s}; \theta)^2) < \infty$$

for all $\theta \in \Theta$. These functions can be chosen freely. When possible, they will be chosen in such a way that the moments needed to find the best predictor and the optimal prediction-based estimating function can be calculated. In general, the functions are allowed to depend on several observations and on the parameter, but in many cases it is convenient to choose functions that are independent of θ, and often power functions of a single observation, $f_j(Y_i) = Y_i^{\nu_j}$, $\nu_j \in \mathbb{N}$, are sufficient.

The predictors of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ are functions of observations before time i. Let $\mathcal{H}_i^\theta$ denote the space of all real-valued functions of the first i observations, $h(Y_1, Y_2, \ldots, Y_i)$, for which $E_\theta(h(Y_1, Y_2, \ldots, Y_i)^2) < \infty$. This is a Hilbert-space with inner product given by

$$\langle h_1, h_2 \rangle_\theta = E_\theta(h_1(Y_1, \ldots, Y_i) h_2(Y_1, \ldots, Y_i)) \tag{2.4}$$

for $h_1, h_2 \in \mathcal{H}_i^\theta$. To construct our estimating function, we choose, for each i and j, a set of predictors $\mathcal{P}_{i-1,j}^\theta$, which is a closed linear subspace of $\mathcal{H}_{i-1}^\theta$. The predictor-spaces $\mathcal{P}_{i-1,j}^\theta$ can be chosen freely, but are usually chosen to be finite-dimensional in order to obtain tractable estimating functions. We shall later consider the case of finite-dimensional predictor-spaces in detail.

A general prediction-based estimating function has the form

$$G_n(\theta) = \sum_{i=s+1}^{n} \sum_{j=1}^{N} \Pi_j^{(i-1)}(\theta)\big\{f_j(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}_j^{(i-1)}(\theta)\big\}. \tag{2.5}$$

Here $\Pi_j^{(i-1)}(\theta) = (\pi_{1,j}^{(i-1)}(\theta), \ldots, \pi_{p,j}^{(i-1)}(\theta))^T$ is a p-dimensional data-dependent vector (T denotes transposition of matrices and vectors) with coordinates belonging to the predictor-space $\mathcal{P}_{i-1,j}^{\theta}$, and $\breve{\pi}_j^{(i-1)}(\theta)$ is the minimum mean square error predictor of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ in $\mathcal{P}_{i-1,j}^{\theta}$. As is well known, the predictor $\breve{\pi}_j^{(i-1)}(\theta)$ is the orthogonal projection of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ onto $\mathcal{P}_{i-1,j}^{\theta}$ with respect to the inner product (2.4) in $\mathcal{H}_i^{\theta}$. The projection exists and is uniquely determined by the normal equations

$$E_\theta\big(\pi\big\{f_j(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}_j^{(i-1)}(\theta)\big\}\big) = 0 \tag{2.6}$$

for all $\pi \in \mathcal{P}_{i-1,j}^{\theta}$; see, for example, Karlin and Taylor (1975). It follows from (2.6) that the prediction-based estimating function (2.5) is an unbiased estimating function, that is, that

$$E_\theta(G_n(\theta)) = 0 \tag{2.7}$$

for all $\theta \in \Theta$. This ensures, under additional regularity conditions given in Section 4, that a consistent estimator can be obtained by solving the estimating equation $G_n(\theta) = 0$.

Example 2.2. If we choose as our predictor-space the space of all functions $h(Y_1, Y_2, \ldots, Y_{i-1})$ satisfying that $E_\theta(h(Y_1, Y_2, \ldots, Y_{i-1})^2) < \infty$, that is, if

$$\mathcal{P}_{i-1,j}^{\theta} = \mathcal{H}_{i-1}^{\theta},$$

then the minimum mean square error predictor of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ in $\mathcal{P}_{i-1,j}^{\theta}$ is the conditional expectation

$$\breve{\pi}_j^{(i-1)}(\theta) = E_\theta(f_j(Y_i, \ldots, Y_{i-s}; \theta) | Y_1, Y_2, \ldots, Y_{i-1});$$

see, for example, Karlin and Taylor (1975). Hence $G_n(\theta)$ is a P_θ-martingale with respect to the filtration generated by the observed process, that is,$G_n(\theta)$ is a *martingale estimating function*; see Heyde (1997) or Sørensen (2011). Thus the martingale estimating functions form a subclass of the prediction-based estimating functions. Unfortunately it is, for most non-Markovian models, not practically feasible to calculate the expectations conditionally on the entire past. Therefore martingale estimating functions are mainly useful in the case of Markov processes (with $s = 1$), where the conditional expectations depend only on Y_{i-1}.

The idea behind the prediction-based estimating functions is to use a smaller and more tractable predictor-space than $\mathcal{H}_{i-1}^{\theta}$. We can interpret the minimum mean square error predictor in the smaller space as an approximation to the conditional expectation of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ given $X_1, \ldots, X_{i-1}$. Thus a prediction-based estimating function can be thought of as an approximation to a martingale estimating function.

Example 2.3. One possibility is that we choose the predictor-space $\mathcal{P}_{i-1,j}^{\theta}$ as the space of all functions $h(Y_{i-1}, \ldots, Y_{i-r})$ $(r \geq s)$ which satisfy that $E_\theta(h(Y_{i-1}, \ldots, Y_{i-r})^2) < \infty$. Then the minimum mean square error predictor of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ is

$$\breve{\pi}_j^{(i-1)}(\theta) = E_\theta(f_j(Y_i, \ldots, Y_{i-s}; \theta) | Y_{i-1}, \ldots, Y_{i-r}).$$

This makes good sense if the observed process Y is exponentially ρ-mixing (see Doukhan (1994) for a definition) because in this case the dependence on the past decreases quickly. However, except for Gaussian processes and the case $r = 1$, it is not practically feasible to calculate expectations conditional on $Y_{i-1}, \ldots, Y_{i-r}$ either.

Example 2.4. Suppose that the observations are one-dimensional and that $N = 1$ with $f(x) = x$ ($j = 1$ is suppressed in the notation when $N = 1$). We assume, moreover, that the observed process Y_i is stationary. We choose the space of predictors as

$$\mathcal{P}_{i-1}^{\theta} = \{a_0 + a_1 Y_{i-1} + \cdots + a_{q_i} Y_{i-q_i} | a_j \in \mathbb{R}, j = 0, 1, \ldots, q_i\},$$

where $q_i \leq i - 1$, and $i = 2, 3, \ldots$. Define $\mathcal{P}_0^{\theta} = \mathbb{R}$, the space of constant predictors.

Let $C^{(i-1)}(\theta)$ denote the covariance matrix of the stochastic vector $Z^{(i-1)} = (Y_{i-1}, \ldots, Y_{i-q_i})^T$, and define the vector of covariances

$$b^{(i-1)}(\theta) = (\mathrm{Cov}_\theta(Y_i, Y_{i-1}), \ldots, \mathrm{Cov}_\theta(Y_i, Y_{i-q_i}))^T.$$

Here and later Cov_θ denotes the covariance under the model with parameter value θ. By solving the normal equations (2.6) we find that the minimum mean square error predictor is given by

$$\breve{\pi}^{(i-1)}(\theta) = \breve{a}_0^{(i-1)}(\theta) + \breve{a}^{(i-1)}(\theta)^T Z^{(i-1)},$$

where $\breve{a}^{(i-1)}(\theta)$ is the q_i-dimensional vector given by

$$\breve{a}^{(i-1)}(\theta) = C^{(i-1)}(\theta)^{-1} b^{(i-1)}(\theta),$$

and where

$$\breve{a}_0^{(i-1)}(\theta) = E_\theta(Y_1)\{1 - (\breve{a}^{(i-1)}(\theta)_1 + \cdots + \breve{a}^{(i-1)}(\theta)_{q_i})\}.$$

Natural choices for the dimension of the predictor-spaces are $q_i = i - 1$ or $q_i = \min(i - 1, q)$ for some fixed $q \geq 1$. The latter choice is a natural simplification when the observed process Y is exponentially ρ-mixing, because in this case the coefficients $\breve{a}^{(i-1)}(\theta)_k$ will decrease exponentially to zero as k increases, that is, the dependence on observations in the far past is negligible. Therefore it is enough to use a bounded number of lagged values of the observed process.

2.1 Finite-dimensional predictor-spaces

To obtain estimators that can relatively easily be calculated in practice, we will now consider predictor-spaces, $\mathcal{P}^{\theta}_{i-1,j}$, that are finite-dimensional. A simple example of this was given in Example 2.4. In the rest of this section we assume that the observed process Y_i is stationary. Finite-dimensional predictor-spaces can also be used for nonstationary processes, but this is computationally more complicated because the coefficients of the minimum mean square error predictors will be time dependent.

Let h_{jk}, $j = 1, \ldots, N$, $k = 0, \ldots, q_j$, be functions from $\mathbb{R}^r$ into $\mathbb{R}$ ($r \geq s$), and define for $i \geq r + 1$

$$Z^{(i-1)}_{jk} = h_{jk}(Y_{i-1}, Y_{i-2}, \ldots, Y_{i-r}).$$

We assume that $E_\theta((Z^{(i-1)}_{jk})^2) < \infty$ for all $\theta \in \Theta$, and let $\mathcal{P}_{i-1,j}$ denote the subspace of $\mathcal{H}^{\theta}_{i-1}$ spanned by $Z^{(i-1)}_{j0}, \ldots, Z^{(i-1)}_{jq_j}$. Note that $\mathcal{P}_{i-1,j}$ does not depend on θ. We set $h_{j0} = 1$ and make the following natural assumption.

Condition 2.5. *The functions $h_{j0}, \ldots, h_{jq_j}$ are linearly independent.*

We write the elements of $\mathcal{P}_{i-1,j}$ in the form $a^T Z^{(i-1)}_j$, where $a^T = (a_0, \ldots, a_{q_j})$ and

$$Z^{(i-1)}_j = \big(Z^{(i-1)}_{j0}, \ldots, Z^{(i-1)}_{jq_j}\big)^T$$

are $(q_j + 1)$-dimensional vectors. With this specification of the predictor-spaces, the predictors are defined for $i \geq r + 1$ only, so the estimating function can only include terms with $i \geq r + 1$:

$$G_n(\theta) = \sum_{i=r+1}^{n} \sum_{j=1}^{N} \Pi^{(i-1)}_j(\theta)\big[f_j(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}_j(\theta)\big]. \tag{2.8}$$

The minimum mean square error predictor, $\breve{\pi}^{(i-1)}_j(\theta)$, is found by solving the normal equations (2.6). Define $C_j(\theta)$ as the covariance matrix of $(Z^{(r)}_{j1}, \ldots, Z^{(r)}_{jq_j})^T$ under P_θ, and $b_j(\theta)$ as the vector for which the kth coordinate is

$$b_j(\theta)_k = \mathrm{Cov}_\theta\big(Z^{(r)}_{jk}, f_j(Y_{r+1}, \ldots, Y_{r+1-s}; \theta)\big), \tag{2.9}$$

$k = 1, \ldots, q_j$. Then we have

$$\breve{\pi}_j^{(i-1)}(\theta) = \breve{a}_j(\theta)^T Z_j^{(i-1)}, \tag{2.10}$$

where $\breve{a}_j(\theta)^T = (\breve{a}_{j0}(\theta), \breve{a}_{j*}(\theta)^T)$ with

$$\breve{a}_{j*}(\theta) = C_j(\theta)^{-1} b_j(\theta) \tag{2.11}$$

and

$$\breve{a}_{j0}(\theta) = E_\theta(f_j(Y_{s+1}, \ldots, Y_1; \theta)) - \sum_{k=1}^{q_j} \breve{a}_{jk}(\theta) E_\theta\big(Z_{jk}^{(r)}\big). \tag{2.12}$$

That $C_j(\theta)$ is invertible follows from Condition 2.5.

Quite often the vector of coefficients $\breve{a}_j$ can be found by means of the N-dimensional Durbin–Levinson algorithm; see Brockwell and Davis (1991, p. 422). This is the case when the functions f_j do not depend on θ, and

$$Z_j^{(i-1)} = (1, F_{i-1}^T, \ldots, F_{i-u}^T)^T \tag{2.13}$$

for all j and for some fixed $u \in \mathbb{N}$. Here the stationary N-dimensional process $\{F_i\}$ is defined by

$$F_i^T = (f_1(Y_i, \ldots, Y_{i-s}), \ldots, f_N(Y_i, \ldots, Y_{i-s})), \tag{2.14}$$

$i = s+1, s+2, \ldots$. In this situation $r = s+u$ and $q_j = q = Nu$. The vector of coefficients $\breve{a}_{j*}(\theta)^T$ in the minimum mean square error predictor is equal to the jth row of the $N \times q$-matrix

$$(\Phi_{u,1}(\theta), \ldots, \Phi_{u,u}(\theta)), \tag{2.15}$$

where the $N \times N$-matrices $\Phi_{u,k}(\theta)$ can be found by running the Durbin–Levinson algorithm for $\ell = 1, \ldots, u$ as described below. The coefficients $\breve{a}_{j0}$ can be found from (2.12), which here simplifies to

$$\begin{pmatrix} \breve{a}_{10}(\theta) \\ \vdots \\ \breve{a}_{N0}(\theta) \end{pmatrix} = \left(I_N - \sum_{k=1}^{u} \Phi_{u,k}(\theta) \right) E_\theta(F_{s+1}), \tag{2.16}$$

where I_N denotes the $N \times N$ identity matrix.

Define the $N \times N$ matrices of autocovariances

$$\Gamma_i(\theta) = E_\theta(F_{s+1} F_{s+1+i}^T). \tag{2.17}$$

The stationary process $\{Y_t\}$ can be extended to be defined for time points $t \le 0$, so that F_i is defined for integers $i \le s$ and $\Gamma_i(\theta)$ for $i < 0$. Generally, $\Gamma_{-i}(\theta) = \Gamma_i(\theta)^T$. This can also be taken as the definition of $\Gamma_i(\theta)$ for $i < 0$. If the process $\{F_i\}$ is time-reversible, then $\Gamma_i(\theta)$ is symmetric, so $\Gamma_{-i}(\theta) = \Gamma_i(\theta)$ for all $i \in \mathbb{N}$.

The Durbin–Levinson algorithm is given by the following iteratively defined $N \times N$-matrices:

$$\Phi_{\ell,\ell}(\theta) = \Delta_{\ell-1}(\theta)\tilde{V}_{\ell-1}^{-1}(\theta), \tag{2.18}$$

$$\tilde{\Phi}_{\ell,\ell}(\theta) = \tilde{\Delta}_{\ell-1}(\theta)V_{\ell-1}^{-1}(\theta), \tag{2.19}$$

$$\Phi_{\ell,k}(\theta) = \Phi_{\ell-1,k}(\theta) - \Phi_{\ell,\ell}(\theta)\tilde{\Phi}_{\ell-1,\ell-k}(\theta), \qquad k = 1, \ldots, \ell-1, \tag{2.20}$$

$$\tilde{\Phi}_{\ell,k}(\theta) = \tilde{\Phi}_{\ell-1,k}(\theta) - \tilde{\Phi}_{\ell,\ell}(\theta)\Phi_{\ell-1,\ell-k}(\theta), \qquad k = 1, \ldots, \ell-1, \tag{2.21}$$

where $V_0 = \tilde{V}_0 = \Gamma_0(\theta)$ and $\Delta_0 = \tilde{\Delta}_0^T = \Gamma_1(\theta)$, and for $\ell \in \mathbb{N}$

$$V_\ell(\theta) = \Gamma_0(\theta) - \Phi_{\ell,1}(\theta)\Gamma_1(\theta)^T - \cdots - \Phi_{\ell,\ell}(\theta)\Gamma_\ell(\theta)^T, \tag{2.22}$$

$$\tilde{V}_\ell(\theta) = \Gamma_0(\theta) - \tilde{\Phi}_{\ell,1}(\theta)\Gamma_1(\theta) - \cdots - \tilde{\Phi}_{\ell,\ell}(\theta)\Gamma_\ell(\theta), \tag{2.23}$$

$$\Delta_\ell(\theta) = \Gamma_{\ell+1}(\theta) - \Phi_{\ell,1}(\theta)\Gamma_\ell(\theta) - \cdots - \Phi_{\ell,\ell}(\theta)\Gamma_1(\theta), \tag{2.24}$$

$$\tilde{\Delta}_\ell(\theta) = \Gamma_{\ell+1}(\theta)^T - \tilde{\Phi}_{\ell,1}(\theta)\Gamma_\ell(\theta)^T - \cdots - \tilde{\Phi}_{\ell,\ell}(\theta)\Gamma_1(\theta)^T. \tag{2.25}$$

The Durbin–Levinson algorithm requires that the autocovariances $\Gamma_i(\theta)$ are available. In general, these quantities must be determined by simulation. However, for the class of prediction-based estimating functions presented in the following example, the autocovariances can be calculated explicitly for a number of very useful models, including those presented in Section 6.

Example 2.6. An important particular case when $d = 1$ is the class of polynomial prediction-based estimating functions. For these

$$f_j(y) = y^{\nu_j}, \qquad j = 1, \ldots, N,$$

where $\nu_j \in \mathbb{N}$. For each $i = r+1, \ldots, n$ and $j = 1, \ldots, N$, we let $\{Z_{jk}^{(i-1)} | k = 0, \ldots, q_j\}$ be a subset of $\{Y_{i-\ell}^\kappa | \ell = 1, \ldots, r, \kappa = 0, \ldots, \nu_j\}$, where $Z_{j0}^{(i-1)}$ is always equal to 1. Here we need to assume that $E_\theta(Y_i^{2\bar{\nu}}) < \infty$ for all $\theta \in \Theta$, where $\bar{\nu} = \max\{\nu_1, \ldots, \nu_N\}$. To find $\breve{\pi}_j^{(i-1)}(\theta)$, $j = 1, \ldots, N$, by means of (2.11) and (2.12) (or by the Durbin–Levinson algorithm), we must calculate moments of the form

$$E_\theta(Y_1^\kappa Y_k^j), \qquad 0 \le \kappa \le j \le \bar{\nu}, k = 1, \ldots, r+1. \tag{2.26}$$

Suppose the observed process Y is exponentially ρ-mixing; see Doukhan (1994) for a definition. Then constants $K > 0$ and $\lambda > 0$ exist such that $|\operatorname{Cov}_\theta(Y_1^\kappa, Y_k^j)| \le Ke^{-\lambda k}$. Therefore a small value of r can usually be used.

In many situations it is reasonable to choose $N = 2$, $\nu_1 = 1$ and $\nu_2 = 2$ with the following simple predictor sets where $q_1 = q_2 = 2r$. For $j = 1, 2$, the predictor-spaces are spanned by $Z_{j0}^{(i-1)} = 1$, $Z_{jk}^{(i-1)} = Y_{i-k}$, $k = 1, \ldots, r$, and $Z_{jk}^{(i-1)} = Y_{i+r-k}^2$, $k = r+1, \ldots, 2r$. As explained above, the minimum mean

square error predictors of Y_i and Y_i^2 can in this case be found by applying the two-dimensional Durbin–Levinson algorithm to the process $F_i = (Y_i, Y_i^2)^T$. However, it might also be of relevance to include in the predictor terms of the form $Y_{i-k}Y_{i-k-\ell}$ for a number of lags ℓ.

3 Optimal estimating functions

A main issue in the theory of estimating functions is to find the optimal element in a class of estimating functions. A detailed exposition can be found in Heyde (1997), and a short review is given in Sørensen (2011). The optimal element in a class of estimating functions is the one that is closest to the score function (the vector of partial derivatives of the log-likelihood function) in a mean-square sense. If the estimators obtained from the estimating functions in the class are asymptotically normal, then the optimal estimating function is the one for which the corresponding estimator has the smallest asymptotic variance. Conditions ensuring asymptotic normality are given in the next section.

In this section we find the optimal estimating function in a class of prediction-based estimating functions with finite-dimensional predictor-spaces $\mathcal{P}_{ij}^{\theta}$. This is the type of estimating functions presented in Section 2.1. As there, we assume that the observed process Y_i is stationary.

First we introduce a more compact notation. The ℓth coordinate of the p-dimensional vector $\Pi_j^{(i-1)}(\theta)$ in (2.8) can be written as

$$\pi_{\ell,j}^{(i-1)}(\theta) = \sum_{k=0}^{q_j} a_{\ell jk}(\theta) Z_{jk}^{(i-1)}, \qquad \ell = 1, \ldots, p.$$

With this notation, (2.8) can be written in the form

$$G_n(\theta) = A(\theta) \sum_{i=r+1}^{n} H^{(i)}(\theta), \tag{3.1}$$

where

$$A(\theta) = \begin{pmatrix} a_{110}(\theta) & \cdots & a_{11q_1}(\theta) & \cdots & \cdots & a_{1N0}(\theta) & \cdots & a_{1Nq_N}(\theta) \\ \vdots & & \vdots & & & \vdots & & \vdots \\ a_{p10}(\theta) & \cdots & a_{p1q_1}(\theta) & \cdots & \cdots & a_{pN0}(\theta) & \cdots & a_{pNq_N}(\theta) \end{pmatrix}$$

and

$$H^{(i)}(\theta) = Z^{(i-1)}(F(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}(\theta)) \tag{3.2}$$

with $F = (f_1, \ldots, f_N)^T$, $\breve{\pi}^{(i-1)}(\theta) = (\breve{\pi}_1^{(i-1)}(\theta), \ldots, \breve{\pi}_N^{(i-1)}(\theta))^T$ and

$$Z^{(i-1)} = \begin{pmatrix} Z_1^{(i-1)} & 0_{q_1+1} & \cdots & 0_{q_1+1} \\ 0_{q_2+1} & Z_2^{(i-1)} & \cdots & 0_{q_2+1} \\ \vdots & \vdots & & \vdots \\ 0_{q_N+1} & 0_{q_N+1} & \cdots & Z_N^{(i-1)} \end{pmatrix}. \tag{3.3}$$

Here 0_{q_j+1} denotes the (q_j+1)-dimensional zero-vector. When we have chosen the functions f_j and the predictor-spaces, the quantities $H^{(i)}(\theta)$ are completely determined, whereas we are free to choose the matrix $A(\theta)$ in an optimal way, that is, such that the asymptotic covariance matrix of the estimators is minimized.

We can slightly more explicitly write

$$G_n(\theta) = A(\theta) \sum_{i=r+1}^{n} \big(Z^{(i-1)} F(Y_i, \ldots, Y_{i-s}; \theta) - Z^{(i-1)} \big(Z^{(i-1)}\big)^T \breve{a}(\theta)\big),$$

where

$$\breve{a}(\theta) = (\breve{a}_{10}(\theta), \ldots, \breve{a}_{1q_1}(\theta), \ldots, \breve{a}_{N0}(\theta), \ldots, \breve{a}_{Nq_N}(\theta))^T. \tag{3.4}$$

The quantities $\breve{a}_{jk}$ define the minimum mean square error predictors; cf. (2.10).

Condition 3.1. (1) *The coordinates of* $F(y_1, \ldots, y_{s+1}; \theta)$ *and* $\breve{a}(\theta)$ *are continuously differentiable functions of* θ.

(2) $p \le \bar{p} = N + q_1 + \cdots + q_N$.

(3) *The* $\bar{p} \times p$*-matrix* $\partial_{\theta^T} \breve{\alpha}(\theta)$ *has rank* p.

(4) *The functions* $1, f_1, \ldots, f_N$ *are linearly independent (for fixed* θ*) on the support of the conditional distribution of* $(Y_{r+1}, \ldots, Y_{r+1-s})$ *given* $(Y_r, \ldots, Y_1)$.

(5) *The* $\bar{p} \times p$*-matrix*

$$U(\theta)^T = E_\theta\big(Z^{(r)}\, \partial_{\theta^T} F(Y_{r+1}, \ldots, Y_{r+1-s}; \theta)\big) \tag{3.5}$$

exists.

Proposition 3.2. *Suppose Condition 3.1 is satisfied for all* $\theta \in \Theta$. *Then the Godambe optimal estimating function in the class of estimating functions of the form (3.1) is given by*

$$G_n^*(\theta) = A_n^*(\theta) \sum_{i=r+1}^{n} Z^{(i-1)} \big(F(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}(\theta)\big), \tag{3.6}$$

where

$$A_n^*(\theta) = S(\theta) \bar{M}_n(\theta)^{-1}. \tag{3.7}$$

Here

$$S(\theta) = U(\theta) - \partial_\theta \breve{a}(\theta)^T D(\theta) \tag{3.8}$$

with $D(\theta)$ *denoting the* $\bar{p} \times \bar{p}$*-matrix*

$$D(\theta) = \mathrm{E}_\theta\big(Z^{(r)} \big(Z^{(r)}\big)^T\big), \tag{3.9}$$

and

$$\bar{M}_n(\theta) = \mathrm{E}_\theta\big(H^{(r+1)}(\theta)H^{(r+1)}(\theta)^T\big) + \sum_{k=1}^{n-r-1} \frac{(n-r-k)}{(n-r)}\big\{\mathrm{E}_\theta\big(H^{(r+1)}(\theta)H^{(r+1+k)}(\theta)^T\big) + \mathrm{E}_\theta\big(H^{(r+1+k)}(\theta)H^{(r+1)}(\theta)^T\big)\big\}. \tag{3.10}$$

When the function F does not depend on θ, the expression for $A_n^*(\theta)$ simplifies slightly as in this case $U(\theta) = 0$.

Proof of Proposition 3.2. By Theorem 2.1 in Heyde (1997), G^* is optimal if and only if

$$E_\theta(\partial_{\theta^T} G_n(\theta))^{-1} E_\theta(G_n(\theta)G_n^*(\theta)^T) = E_\theta(\partial_{\theta^T} G_n^*(\theta))^{-1} E_\theta(G_n^*(\theta)G_n^*(\theta)^T)$$

for all G of the form (3.1), which is the case when

$$E_\theta(G_n(\theta)G_n^*(\theta)^T) = E_\theta(\partial_{\theta^T} G_n(\theta))$$

for all G of the form (3.1). This equation obviously holds when $A_n^*(\theta)$ is given by (3.7), because

$$\mathrm{E}_\theta(G_n(\theta)G_n^*(\theta)^T) = (n-r)A(\theta)\bar{M}_n(\theta)A_n^*(\theta)^T$$

and

$$E_\theta(\partial_{\theta^T} G_n(\theta)) = (n-r)A(\theta)[U(\theta)^T - D(\theta)\,\partial_{\theta^T}\breve{a}(\theta)].$$

The matrix $\bar{M}_n(\theta)$, which is the covariance matrix of $\sqrt{n-r}\,H_n(\theta)$ under the probability measure P_θ, where

$$H_n(\theta) = (n-r)^{-1} \sum_{i=r+1}^{n} H^{(i)}(\theta), \tag{3.11}$$

is invertible under Condition 3.1(4); see Sørensen (2000). □

If the process Y is sufficiently mixing, then the matrix $\bar{M}_n(\theta)$ converges by the ergodic theorem to a matrix $M(\theta)$ as $n \to \infty$; see Section 4. The matrix $M(\theta)$ is given by (4.2). The asymptotic covariance matrix of the prediction-based estimator does not depend on whether we use the weight matrix given by (3.7) or the one given by by

$$A^*(\theta) = S(\theta)M(\theta)^{-1}. \tag{3.12}$$

Both estimators are optimal, and they are usually almost identical. In practice the matrices $\bar{M}_n(\theta)$ and $M(\theta)$ can often most easily be calculated by simulating

$\sqrt{n-r}H_n(\theta)$ a large number of times under P_θ and then calculating the empirical covariance matrix. Alternatively, $M(\theta)$ can be calculated by truncating the series (4.2), and $\bar{M}_n(\theta)$ by including only the significant terms in the sum (3.10). If the observed process is geometrically ρ-mixing, the terms in both formulae will decrease rapidly to zero.

The matrix (3.7) or (3.12) is the computationally most demanding part of the optimal prediction-based estimating function. The time used to calculate the optimal estimator can therefore be reduced very considerably if $A_n^*(\theta)$ or $A^*(\theta)$ is calculated for one parameter value only. This can be achieved by replacing $A^*(\theta)$ by $A^*(\tilde{\theta}_n)$ (or $A_n^*(\theta)$ by $A_n^*(\tilde{\theta}_n)$), where $\tilde{\theta}_n$ is a consistent estimator. Under the Conditions 4.1 and 4.2, the estimating function obtained by this simplification gives an estimator with the same asymptotic variance as the original optimal estimating function. A consistent estimator can, for instance, be obtained from the estimating function that we get by using only p coordinates of $H_n(\theta)$ (without a weight matrix). This is possible because $\bar{p} \geq p$. Under the Conditions 4.1 and 4.2, this simple estimating function gives a consistent estimator. Note that in this case Condition 4.2(3) is automatically satisfied because here $A_n(\theta)$ equals the p-dimensional identity matrix.

Example 3.3. Consider again the polynomial prediction-based estimating functions discussed in Example 2.6. In order to calculate (3.10), we need mixed moments of the form

$$E_\theta[Y_{t_1}^{k_1} Y_{t_2}^{k_2} Y_{t_3}^{k_3} Y_{t_4}^{k_4}] \tag{3.13}$$

for $1 \leq t_1 \leq t_2 \leq t_3 \leq t_4 \leq r+1$ and $k_1+k_2+k_3+k_4 \leq 4\bar{\nu}$, where $k_i, i = 1, \ldots, 4$, are nonnegative integers.

For prediction-based estimating functions where the f_j's do not depend on θ and the predictor-space is given by (2.13) and (2.14), the derivatives $\partial_\theta \breve{a}_{jk}(\theta)$, $j = 1, \ldots, N$, $k = 1, \ldots, q$, in (3.4) can be found from the autocovariance matrices (2.17) and their derivatives with respect to θ by the following algorithm that is obtained by differentiating the Durbin–Levinson algorithm given by (2.18)–(2.25) with respect to θ_i for every $i = 1, \ldots, p$. The vector $\partial_{\theta_i} \breve{a}_{j*}(\theta)^T$ is the jth row of the matrix

$$(\partial_{\theta_i} \Phi_{u,1}(\theta), \ldots, \partial_{\theta_i} \Phi_{u,u}(\theta)),$$

where $\partial_{\theta_i} \Phi_{u,k}(\theta)$ is obtained by the following algorithm:

$$\partial_{\theta_i} \Phi_{\ell,\ell}(\theta) = \partial_{\theta_i} \Delta_{\ell-1}(\theta) \tilde{V}_{\ell-1}^{-1}(\theta) + \Delta_{\ell-1}(\theta) \tilde{W}_{\ell-1}(\theta),$$

$$\partial_{\theta_i} \tilde{\Phi}_{\ell,\ell}(\theta) = \partial_{\theta_i} \tilde{\Delta}_{\ell-1}(\theta) V_{\ell-1}^{-1}(\theta) + \tilde{\Delta}_{\ell-1}(\theta) W_{\ell-1}(\theta)$$

and for $k = 1, \ldots, \ell-1$

$$\partial_{\theta_i} \Phi_{\ell,k}(\theta) = \partial_{\theta_i} \Phi_{\ell-1,k}(\theta) - \partial_{\theta_i} \Phi_{\ell,\ell}(\theta) \tilde{\Phi}_{\ell-1,\ell-k}(\theta) - \Phi_{\ell,\ell}(\theta)\, \partial_{\theta_i} \tilde{\Phi}_{\ell-1,\ell-k}(\theta),$$

$$\partial_{\theta_i} \tilde{\Phi}_{\ell,k}(\theta) = \partial_{\theta_i} \tilde{\Phi}_{\ell-1,k}(\theta) - \partial_{\theta_i} \tilde{\Phi}_{\ell,\ell}(\theta) \Phi_{\ell-1,\ell-k}(\theta) - \tilde{\Phi}_{\ell,\ell}(\theta)\, \partial_{\theta_i} \Phi_{\ell-1,\ell-k}(\theta),$$

where

$$W_\ell(\theta) = -V_\ell(\theta)^{-1}\,\partial_{\theta_i} V_\ell(\theta) V_\ell(\theta)^{-1},$$
$$\tilde{W}_\ell(\theta) = -\tilde{V}_\ell(\theta)^{-1}\,\partial_{\theta_i} \tilde{V}_\ell(\theta) \tilde{V}_\ell(\theta)^{-1},$$

$V_0 = \tilde{V}_0 = \Gamma_0(\theta)$ and $\Delta_0 = \tilde{\Delta}_0^T = \Gamma_1(\theta)$. For $\ell \in \mathbb{N}$

$$\begin{aligned}
\partial_{\theta_i} V_\ell(\theta) &= \partial_{\theta_i}\Gamma_0(\theta) - \partial_{\theta_i}\Phi_{\ell,1}(\theta)\Gamma_1(\theta)^T - \cdots - \partial_{\theta_i}\Phi_{\ell,\ell}(\theta)\Gamma_\ell(\theta)^T \\
&\quad - \Phi_{\ell,1}(\theta)\,\partial_{\theta_i}\Gamma_1(\theta)^T - \cdots - \Phi_{\ell,\ell}(\theta)\,\partial_{\theta_i}\Gamma_\ell(\theta)^T, \\
\partial_{\theta_i} \tilde{V}_\ell(\theta) &= \partial_{\theta_i}\Gamma_0(\theta) - \partial_{\theta_i}\tilde{\Phi}_{\ell,1}(\theta)\Gamma_1(\theta) - \cdots - \partial_{\theta_i}\tilde{\Phi}_{\ell,\ell}(\theta)\Gamma_\ell(\theta) \\
&\quad - \tilde{\Phi}_{\ell,1}(\theta)\,\partial_{\theta_i}\Gamma_1(\theta) - \cdots - \tilde{\Phi}_{\ell,\ell}(\theta)\,\partial_{\theta_i}\Gamma_\ell(\theta), \\
\partial_{\theta_i} \Delta_\ell(\theta) &= \partial_{\theta_i}\Gamma_{\ell+1}(\theta) - \partial_{\theta_i}\Phi_{\ell,1}(\theta)\Gamma_\ell(\theta) - \cdots - \partial_{\theta_i}\Phi_{\ell,\ell}(\theta)\Gamma_1(\theta) \\
&\quad - \Phi_{\ell,1}(\theta)\,\partial_{\theta_i}\Gamma_\ell(\theta) - \cdots - \Phi_{\ell,\ell}(\theta)\,\partial_{\theta_i}\Gamma_1(\theta), \\
\partial_{\theta_i} \tilde{\Delta}_\ell(\theta) &= \partial_{\theta_i}\Gamma_{\ell+1}(\theta)^T - \partial_{\theta_i}\tilde{\Phi}_{\ell,1}(\theta)\Gamma_\ell(\theta)^T - \cdots - \partial_{\theta_i}\tilde{\Phi}_{\ell,\ell}(\theta)\Gamma_1(\theta)^T \\
&\quad - \tilde{\Phi}_{\ell,1}(\theta)\,\partial_{\theta_i}\Gamma_\ell(\theta)^T - \cdots - \tilde{\Phi}_{\ell,\ell}(\theta)\,\partial_{\theta_i}\Gamma_1(\theta)^T.
\end{aligned}$$

The matrices $\Phi_{\ell,k}(\theta)$ and $\tilde{\Phi}_{\ell,k}(\theta)$, $k = 1, \ldots, \ell$, are given by (2.18)–(2.21), and $V_\ell(\theta)$, $\tilde{V}_\ell(\theta)$, $\Delta_\ell(\theta)$ and $\tilde{\Delta}_\ell(\theta)$ by (2.22)–(2.25).

4 Asymptotic theory

In this section we give conditions ensuring that a prediction-based estimating function gives an estimator that is consistent, asymptotically normal, and ultimately unique. The result is based on general asymptotic statistical theory for stochastic processes, which is presented in a generality suitable for our purpose in Sørensen (1999) and Jacod and Sørensen (2011). We give asymptotic results only for estimating functions of the form

$$G_n(\theta) = A_n(\theta) \sum_{i=r+1}^{n} H^{(i)}(\theta), \tag{4.1}$$

which is the most useful case in practice. Here $A_n(\theta)$ is a (possibly data-dependent) $p \times \bar{p}$-matrix ($\bar{p} = q_1 + \cdots + q_N + N$), and $H^{(i)}(\theta)$ is given by (3.2). In this case the conditions for the asymptotic theory are particularly simple.

We assume the following conditions, where θ_0 is the true parameter value. We denote the state space of the observed process Y by $\mathcal{Y}$.

Condition 4.1. (1) *The observed process Y is stationary and geometrically α-mixing.*

(2) *There exists a* $\delta > 0$ *such that*

$$E_{\theta_0}(|h_{jk}(Y_r, \ldots, Y_1) f_j(Y_{r+1}, \ldots, Y_{r+1-s}; \theta_0)|^{2+\delta}) < \infty$$

and

$$E_{\theta_0}(|h_{jk}(Y_r, \ldots, Y_1) h_{j\ell}(Y_r, \ldots, Y_1)|^{2+\delta}) < \infty$$

for $j = 1, \ldots, N, k, \ell = 0, \ldots, q_j$.

For a definition of the concept of α-mixing, see Doukhan (1994). Condition 4.1 ensures that we can apply a central limit theorem to the estimating function.

Let Q denote the distribution of $(Y_1, \ldots, Y_{s+1})$. A function $f : \mathcal{Y}^{s+1} \times \Theta \mapsto \mathbb{R}$ is called locally dominated integrable with respect to Q if for each $\theta' \in \Theta$ there exists a neighbourhood $U_{\theta'}$ of θ' and a nonnegative Q-integrable function $h_{\theta'} : \mathcal{Y}^{s+1} \mapsto \mathbb{R}$ such that $|f(y_1, \ldots, y_{s+1}; \theta)| \leq h_{\theta'}(y_1, \ldots, y_{s+1})$ for all $(y_1, \ldots, y_{s+1}, \theta) \in \mathcal{Y}^{s+1} \times U_{\theta'}$.

Condition 4.2. (1) *The components of* $F(y_1, \ldots, y_{s+1}; \theta)$, $A_n(\theta)$ *and* $\breve{a}(\theta)$, *given by* (3.4), *are continuously differentiable functions of* θ.

(2) *The functions* $\|\partial_\theta f_j(y_1, \ldots, y_{s+1}; \theta)\|$, $j = 1, \ldots, N$, *are locally dominated integrable with respect to* Q.

(3) *There exists a nonrandom matrix* $A(\theta)$ *such that for any compact subset* $K \subseteq \Theta$

$$A_n(\theta) \xrightarrow{P_{\theta_0}} A(\theta), \qquad \partial_\theta A_n(\theta) \xrightarrow{P_{\theta_0}} \partial_\theta A(\theta)$$

uniformly for $\theta \in K$ *as* $n \to \infty$.

(4) *The matrix*

$$W = A(\theta_0) S(\theta_0)^T = A(\theta_0)\big(U(\theta_0)^T - D(\theta_0)\, \partial_{\theta^T} \breve{a}(\theta_0)\big)$$

has full rank p. *The matrices* $S(\theta)$, $U(\theta)$ *and* $D(\theta)$ *are given by* (3.8), (3.5) *and* (3.9).

(5)

$$A(\theta)\big(E_{\theta_0}\big(Z^{(r)} F(Y_{r+1}, \ldots, Y_{r+1-s}; \theta)\big) - D(\theta_0)\breve{a}(\theta)\big) \neq 0$$

for all $\theta \neq \theta_0$.

Theorem 4.3. *Assume that the true parameter value* θ_0 *belongs to the interior of the parameter space* Θ, *and that the Conditions* 4.1 *and* 4.2 *are satisfied. Then a consistent estimator* $\hat{\theta}_n$ *exists that, with a probability tending to one as* $n \to \infty$, *solves the estimating equation* $G_n(\hat{\theta}_n) = 0$ *and is unique in any compact subset* $K \subseteq \Theta$ *for which* $\theta_0 \in \operatorname{int} K$. *Moreover,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p(0, W^{-1} A(\theta_0) M(\theta_0) A(\theta_0)^T W^{T\,-1})$$

as $n \to \infty$, *where*

$$M(\theta) = \mathrm{E}_\theta\big(H^{(r+1)}(\theta)H^{(r+1)}(\theta)^T\big) + \sum_{k=1}^{\infty}\{\mathrm{E}_\theta\big(H^{(r+1)}(\theta)H^{(r+1+k)}(\theta)^T\big) + \mathrm{E}_\theta\big(H^{(r+1+k)}(\theta)H^{(r+1)}(\theta)^T\big)\}. \tag{4.2}$$

Proof. Consider $H_n(\theta)$ given by (3.11). Condition 4.2(1) and (2) implies that for any compact $K \subseteq \Theta$

$$\sup_{\theta\in K} \|H_n(\theta) - \tilde{W}(\theta)\| \xrightarrow{P_{\theta_0}} 0,$$

$$\sup_{\theta\in K} \|\partial_{\theta^T} H_n(\theta) - \tilde{W}'(\theta)\| \xrightarrow{P_{\theta_0}} 0,$$

where

$$\tilde{W}(\theta) = E_{\theta_0}\big(Z^{(r)}F(Y_{r+1},\ldots,Y_{r+1-s};\theta)\big) - D(\theta_0)\breve{a}(\theta),$$

$$\tilde{W}'(\theta) = E_{\theta_0}\big(Z^{(r)}\,\partial_{\theta^T}F(Y_{r+1},\ldots,Y_{r+1-s};\theta)\big) - D(\theta_0)\,\partial_{\theta^T}\breve{a}(\theta).$$

The components of $\tilde{W}(\theta)$ and $\tilde{W}'(\theta)$ are continuous functions of θ. Define

$$W(\theta) = \partial_{\theta^T}A(\theta)\tilde{W}(\theta) + A(\theta)\tilde{W}'(\theta).$$

From the unbiasedness of G_n (cf. (2.7)), we see that $\tilde{W}(\theta_0) = 0$, so $W(\theta_0) = A(\theta_0)\tilde{W}'(\theta_0) = W$, which is assumed to be an invertible matrix. By using that

$$\begin{aligned}
&\|(n-r)^{-1}\,\partial_{\theta^T}G_n(\theta) - W(\theta)\| \\
&\quad\le \big\|\big(\partial_{\theta^T}A_n(\theta) - \partial_{\theta^T}A(\theta)\big)\big(\big(H_n(\theta) - \tilde{W}(\theta)\big) + \tilde{W}(\theta)\big)\big\| \\
&\qquad + \big\|\big(A_n(\theta) - A(\theta)\big)\big(\big(\partial_{\theta^T}H_n(\theta) - \tilde{W}'(\theta)\big) + \tilde{W}'(\theta)\big)\big\| \\
&\qquad + \big\|\partial_{\theta^T}A(\theta)\big(H_n(\theta) - \tilde{W}(\theta)\big)\big\| + \big\|A(\theta)\big(\partial_{\theta^T}H_n(\theta) - \tilde{W}'(\theta)\big)\big\|,
\end{aligned}$$

it follows that there exists a constant $C > 0$ such that

$$\begin{aligned}
&\sup_{\theta\in K}\|(n-r)^{-1}\,\partial_{\theta^T}G_n(\theta) - W(\theta)\| \\
&\quad\le C\Big(\sup_{\theta\in K}\|\partial_{\theta^T}A_n(\theta) - \partial_{\theta^T}A(\theta)\| + \sup_{\theta\in K}\|A_n(\theta) - A(\theta)\| \\
&\qquad + \sup_{\theta\in K}\|H_n(\theta) - \tilde{W}(\theta)\| + \sup_{\theta\in K}\|\partial_{\theta^T}H_n(\theta) - \tilde{W}'(\theta)\|\Big),
\end{aligned}$$

where the right-hand side goes to zero in probability under P_{θ_0} as $n \to \infty$. This together with the observation that

$$n^{-1}G_n(\theta_0) \xrightarrow{P_{\theta_0}} A(\theta_0)\tilde{W}(\theta_0) = 0$$

imply the existence of a consistent estimator that ultimately solves the estimating equation. For details of this and the following arguments, see Jacod and Sørensen (2011). The uniqueness of the estimator follows from the fact that by Condition 4.2(5) the continuous function $A(\theta)\tilde{W}(\theta)$ (the limit of $n^{-1}G_n(\theta)$) is bounded away from zero on $K \setminus B$ for any compact $K \subseteq \Theta$ with $\theta_0 \in \operatorname{int} K$ and any open neighbourhood B of θ_0.

Condition 4.1 ensures that the central limit theorem for α-mixing processes can be applied to the estimating function $G_n(\theta_0)$ (see Doukhan (1994, Section 1.5)). Specifically, Condition 4.1 and Condition 4.2(3) imply that

$$\frac{1}{\sqrt{n}} G_n(\theta_0) \xrightarrow{\mathcal{D}} N(0, A(\theta_0)M(\theta_0)A(\theta_0)^T).$$

This implies the asymptotic normality of $\hat{\theta}_n$ by standard arguments. □

For the optimal estimator $A(\theta) = A^*(\theta) = S(\theta)M(\theta)^{-1}$, so

$$W = S(\theta_0)M(\theta_0)^{-1}S(\theta_0)^T$$

and the asymptotic variance simplifies to

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p(0, W^{-1}).$$

If the matrix $A_n(\theta)$ does not depend on n, then Condition 4.2(3) is trivially satisfied. This is, for instance, the case if the asymptotic optimal matrix $A^*(\theta)$ given by (3.12) is used. If $A_n(\theta) = A(\tilde{\theta}_n)$ for some matrix $A(\theta)$ independent of n, for example, $A^*(\theta)$, and some consistent estimator $\tilde{\theta}_n$, then Condition 4.2(3) is satisfied if $A(\theta)$ is a continuous function of θ.

In most applications the functions f_j do not depend on θ. If this is the case, Condition 4.2(2) is trivially satisfied, and $U(\theta) = 0$ in Condition 4.2(4).

Suppose the functions f_j do not depend on θ, and that the predictor-space is given by the natural specification (2.13) and (2.14). Suppose, moreover, that $A_n(\theta) = A(\tilde{\theta}_n)$ for some consistent estimator $\tilde{\theta}_n$ and some matrix $A(\theta)$ independent of n. Then Conditions 4.1 and 4.2 are implied by the following simpler condition.

Condition 4.4. (1) *The observed process Y is stationary and geometrically α-mixing.*

(2) *There exists a $\delta > 0$ such that*

$$E_{\theta_0}(|f_j(Y_{s+u+1}, \ldots, Y_{u+1}) f_k(Y_{s+u+1-v}, \ldots, Y_{u+1-v}; \theta_0)|^{2+\delta}) < \infty$$

for $j, k = 1, \ldots, N$, $v = 1, \ldots, u$.

(3) *The components of $\breve{a}(\theta)$, given by (3.4), are continuously differentiable functions of θ, and the components of $A(\theta)$ are continuous functions of θ.*

(4) *The matrix* $W = -A(\theta_0)D(\theta_0)\,\partial_{\theta^T}\breve{a}(\theta_0)$ *has full rank* p. *The matrix* $D(\theta)$ *is given by* (3.9).

(5)

$$A(\theta_0)\big(E_{\theta_0}\big(Z^{(s+u)}F(Y_{s+u+1},\ldots,Y_{u+1})\big) - D(\theta_0)\breve{a}(\theta)\big) \neq 0$$

for all $\theta \neq \theta_0$.

Similar asymptotic results can be given for general prediction-based estimating functions, provided that the predictor-spaces are subsets of the space of all functions $h(Y_{i-1},\ldots,Y_{i-r})$ (for a fixed $r \geq s$) satisfying that $E_\theta(h(Y_{i-1},\ldots,Y_{i-r})^2) < \infty$. If predictors can depend on all past observations, the situation is much more complicated, and it is an open question how to prove general asymptotic results. The situation is similar to that for hidden Markov models, which is a particular case.

5 Gaussian models

In this section we consider prediction-based estimating functions when the observed process Y is a one-dimensional stationary and geometrically and α-mixing Gaussian process. We simplify the exposition by assuming that the expectation of Y_i is zero. The following theory can easily be modified to cover the case of a nonzero mean.

The distribution of Y is determined by the autocovariances

$$K_i(\theta) = E_\theta(Y_1 Y_{1+i}), \qquad i \in \mathbb{N}_0, \tag{5.1}$$

which depend on the p-dimensional parameter $\theta \in \Theta$. We define $K_{-i}(\theta) = K_i(\theta)$ for all $i \in \mathbb{N}$. A natural estimator is obtained by maximizing the pseudo-likelihood function defined as the product of the conditional densities of Y_i given $Y_{i-1},\ldots,Y_{i-s}$ for $i = s+1,\ldots,n$. Here s will typically be relatively small. This pseudo-likelihood function was proposed by Sørensen (2003) in connection with stochastic volatility models, but the idea is more widely applicable. To calculate the pseudo-likelihood function, we define the s-dimensional vector

$$\kappa(\theta) = (K_1(\theta),\ldots,K_s(\theta))^T$$

and the $s \times s$-matrix

$$\mathcal{K}(\theta) = \{K_{i-j}(\theta)\}_{i,j=1,\ldots,s}.$$

The matrix $\mathcal{K}(\theta)$ is the covariance matrix of the vector of the s consecutive observations, for instance $(Y_1,\ldots,Y_s)$. We will make the very weak assumption that $\mathcal{K}(\theta)$ is invertible. The conditional distribution of the observation Y_i given the s previous observations $Y_{i-s},\ldots,Y_{i-1}$ is the normal distribution with expectation

$\phi(\theta)^T Y_{i-1:i-s}$ and variance $v(\theta)$, where $Y_{i:j} = (Y_i, \ldots, Y_j)^T$, $i > j \geq 1$, $\phi(\theta)$ is the s-dimensional vector given by

$$\phi(\theta) = \mathcal{K}(\theta)^{-1}\kappa(\theta),$$

and

$$v(\theta) = K_0(\theta) - \kappa(\theta)^T \mathcal{K}(\theta)^{-1}\kappa(\theta).$$

The vector $\phi(\theta)$ and the conditional variance $v(\theta)$ can be found by means of the Durbin–Levinson algorithm; see Section 2.

The pseudo-likelihood is given by

$$L_n(\theta) = \prod_{i=s+1}^{n} \left[\frac{1}{\sqrt{2\pi v(\theta)}} \exp\left(-\frac{1}{2v(\theta)} \left(Y_i - \phi(\theta)^T Y_{i-1:i-s}\right)^2 \right) \right]. \tag{5.2}$$

If we assume that the autocovariances $K_\theta(i)$, $i = 0, 1, \ldots,$ are continuously differentiable with respect to θ, we obtain the pseudo-score function as the vector of partial derivatives of $\log L_n(\theta)$ with respect to the coordinates of θ:

$$\begin{aligned} G_n^\circ(\theta) &= \partial_\theta \log L_n(\theta) \\ &= \sum_{i=s+1}^{n} \left\{ \frac{\partial_\theta \phi(\theta)^T Y_{i-1:i-s}}{v(\theta)} \left(Y_i - \phi(\theta)^T Y_{i-1:i-s}\right) \right. \\ &\qquad \left. + \frac{\partial_\theta v(\theta)}{2v(\theta)^2} \sum_{i=s+1}^{n} \left[\left(Y_i - \phi(\theta)^T Y_{i-1:i-s}\right)^2 - v(\theta) \right] \right\}. \end{aligned} \tag{5.3}$$

The derivatives $\partial_\theta \phi(\theta)$ and $\partial_\theta v(\theta)$ can be found from the autocovariances $K_i(\theta)$ and their derivatives with respect to θ by the algorithm that is obtained by differentiating the Durbin–Levinson algorithm; see Section 3.

The minimum mean square error linear predictors of Y_i and $(Y_i - \phi(\theta)^T \times Y_{i-1:i-s})^2$ given $Y_{i-1:i-s}$ are $\phi(\theta)^T Y_{i-1:i-s}$ and $v(\theta)$, respectively. This is because for Gaussian processes the two conditional expectations are linear in $Y_{i-1:i-s}$. Hence the pseudo-score function (5.3) is a prediction-based estimating function. Specifically, it is of the form

$$G_n(\theta) = A(\theta) \sum_{i=s+1}^{n} H^{(i)}(\theta),$$

where $A(\theta)$ is a $p \times (s+1)$-matrix of weights that can depend on the parameter, but not on the data, and

$$H^{(i)}(\theta) = Z^{(i)} \begin{pmatrix} Y_i - \phi(\theta)^T Y_{i-1:i-s} \\ \left(Y_i - \phi(\theta)^T Y_{i-1:i-s}\right)^2 - v(\theta) \end{pmatrix}$$

with

$$Z^{(i)} = \begin{pmatrix} Y_{i-1:i-s}^T & 0 \\ 0 \cdots 0 & 1 \end{pmatrix}^T,$$

$i = s+1, \ldots, n$. The pseudo-score, $G_n^\circ(\theta)$, is obtained if the weight matrix $A(\theta)$ is chosen as

$$\tilde{A}(\theta) = \left(\frac{\partial_\theta \phi(\theta)^T}{v(\theta)}, \frac{\partial_\theta v(\theta)}{2v(\theta)^2} \right).$$

The asymptotic optimal weight matrix is given by

$$A^*(\theta) = S(\theta) M(\theta)^{-1},$$

where the matrix $M(\theta)$ is given by (4.2) with $r = s$, and

$$S(\theta)^T = E_\theta\big(\partial_{\theta^T} H^{(i)}(\theta)\big) = -\begin{pmatrix} \mathcal{K}(\theta)\, \partial_{\theta^T} \phi(\theta) \\ \partial_{\theta^T} v(\theta) \end{pmatrix}.$$

In the expression for $M(\theta)$ the first term is given by

$$M^{(1)}(\theta) = E_\theta\big(H^{(s+1)}(\theta) H^{(s+1)}(\theta)^T\big) = \begin{pmatrix} v(\theta)\mathcal{K}(\theta) & O_{s,1} \\ O_{1,s} & 2v(\theta)^2 \end{pmatrix}$$

with O_{j_1,j_2} denoting the $j_1 \times j_2$-matrix of zeros. The optimal matrix $A_n^*(\theta)$ is given by a similar expression where $M(\theta)$ is replaced by the matrix (3.10). The pseudo-score function, $G_n^\circ(\theta)$, is not equal to the optimal prediction-based estimating function. In fact,

$$\tilde{A}(\theta) = -S(\theta) M^{(1)}(\theta)^{-1}.$$

The class of estimating functions considered here is not the full class of prediction-based estimating function to which the pseudo-score (5.3) belongs. The full class is obtained by replacing $A(\theta)$ by a $p \times 2(s+1)$-matrix and $H^{(i)}(\theta)$ by the $2(s+1)$-dimensional vectors $\breve{H}^{(i)}(\theta)$ obtained when $Z^{(i)}$ is replaced by the $2(s+1) \times 2$-matrix

$$\breve{Z}^{(i)} = \begin{pmatrix} Y_{i-1:i-s}^T & 0 & 1 & O_{1,s} \\ O_{1,s} & 1 & 0 & Y_{i-1:i-s}^T \end{pmatrix}^T$$

in the definition of $H^{(i)}(\theta)$. In this way $H^{(i)}(\theta)$ is extended by $s+1$ extra coordinates. Using that all moments of an odd order of a centered multivariate normal distribution equal zero, we see that the extra $s+1$ coordinates of $\breve{H}^{(i)}(\theta)$ have expectation zero under the true probability measure irrespectively of the value of the parameter θ. Therefore they cannot be expected to be a useful addition to $H^{(i)}(\theta)$. The extra coordinates might, however, be correlated with the coordinates of $H^{(i)}(\theta)$, and might thus be used to reduce the variance of the estimating function. To see that this is not the case, the optimal estimating function based on $\breve{H}^{(i)}(\theta)$ can be calculated. The covariance matrix of the random vector $\sum_{i=s+1}^n \breve{H}^{(i)}(\theta)/\sqrt{n-s}$ can be shown to be a block-diagonal matrix with two $(s+1) \times (s+1)$-blocks, the first of which equals $\bar{M}_n(\theta)$ given by (3.10). Here

we use again that moments of an odd order of a centered multivariate Gaussian distribution equal zero. Since

$$E_\theta\big(\partial_{\theta^T}\breve{H}^{(i)}(\theta)\big) = -\begin{pmatrix} \mathcal{K}(\theta)\,\partial_{\theta^T}\phi(\theta) \\ \partial_{\theta^T} v(\theta) \\ O_{s+1,p} \end{pmatrix},$$

it follows that the asymptotic optimal weight-matrix in the full class is

$$\breve{A}_n^*(\theta) = (\, A_n^*(\theta) \quad O_{p,s+1}\,).$$

Thus the optimal prediction-based estimating function obtained from $\breve{H}^{(i)}(\theta)$ equals the optimal estimating function obtained from $H^{(i)}(\theta)$. It is therefore sufficient to consider the smaller class of prediction-based estimating functions above.

We have generally assumed that the observed process is geometrically α-mixing, so the Conditions 4.1 and 4.2 ensuring the asymptotic results in Theorem 4.3 are implied by the following condition:

Condition 5.1. (a) *The functions $K_i(\theta)$ and $A(\theta)$ are twice continuously differentiable with respect to θ.*

(b) *The $p \times (s+1)$-matrix $(\partial_\theta \phi^T(\theta), \partial_\theta v(\theta))$ has rank p (in particular, $s+1 \geq p$).*

(c) *$A(\theta)\bar{\mathcal{K}}(\bar{\phi}(\theta_0) - \bar{\phi}(\theta)) = 0$ if and only if $\theta = \theta_0$, where*

$$\bar{\mathcal{K}} = \begin{pmatrix} \mathcal{K}(\theta_0) & O_{s,1} \\ O_{1,s} & 1 \end{pmatrix}$$

and

$$\bar{\phi}(\theta) = \begin{pmatrix} \phi(\theta) \\ v(\theta) + 2\phi(\theta)^T \kappa(\theta_0) - \phi(\theta)^T \mathcal{K}(\theta_0)\phi(\theta) \end{pmatrix}.$$

Example 5.2. Consider the stochastic delay differential equation

$$dY_t = -\beta\left(\int_{-c}^{0} Y_{t+s}\, ds\right) dt + \sigma\, dW_t,$$

where Y is one-dimensional, $c > 0$, $\sigma > 0$. According to Reiß (2002), a stationary solution exists exactly when $0 < \beta < \frac{1}{2}\pi^2/c^2$. The stationary solution is an exponentially β-mixing Gaussian process with expectation zero and auto-covariance function

$$K_t(\theta) = E_\theta(Y_1, Y_{1+t}) = \frac{\sigma^2 \sin(c\sqrt{2\beta}(1/2 - t))}{2c\sqrt{2\beta}\cos(c\sqrt{\beta/2})} + \frac{\sigma^2}{2\beta c^2}; \qquad 0 \leq t \leq c,$$

see Reiß (2002). Küchler and Sørensen (2009) studied prediction-based estimating functions for more general affine stochastic delay differential equations.

6 Pearson diffusions

The Pearson diffusions (see Wong (1964) and Forman and Sørensen (2008)) is a widely applicable class of diffusion models for which explicit expressions are available for the mixed moments (2.26) and (3.13) needed to calculate polynomial prediction-based estimating functions.

A Pearson diffusion is a stationary solution to a stochastic differential equation of the form

$$dX_t = -\beta(X_t - \alpha)\,dt + \sqrt{2\beta(aX_t^2 + bX_t + c)}\,dW_t, \tag{6.1}$$

where $\beta > 0$, and a, b and c are such that the square root is well defined when X_t is in the state space. A list of all possible cases is given below. The parameter $\beta > 0$ is a scaling of time that determines how fast the diffusion moves. The parameters α, a, b and c determine the state space of the diffusion as well as the shape of the invariant distribution. In particular, α is the expectation of the invariant distribution. The Pearson diffusions are ergodic and ρ-mixing with exponentially decaying mixing coefficients. This follows from Genon-Catalot, Jeantheau and Laredo (2000, Theorem 2.6).

The moments of the Pearson diffusions can, when they exist, be found explicitly. It can be shown that for $\kappa > 1$, $E(|X_t|^\kappa) < \infty$ if and only if $a < (\kappa - 1)^{-1}$. Thus if $a \leq 0$ all moments exist, while for $a > 0$ only the moments satisfying that $\kappa < a^{-1} + 1$ exist. In particular, the expectation always exists. The moments of the invariant distribution can be found by the recursion

$$E(X_t^n) = a_n^{-1}\{b_n \cdot E(X_t^{n-1}) + c_n \cdot E(X_t^{n-2})\}, \qquad n = 2, 3, \ldots, \tag{6.2}$$

where

$$a_n = n\{1 - (n-1)a\}\beta, \qquad b_n = n\{\alpha + (n-1)b\}\beta, \qquad c_n = n(n-1)c\beta$$

for $n = 0, 1, 2, \ldots$. The initial conditions are $E(X_t^0) = 1$ and $E(X_t) = \alpha$. To see this, note that by Ito's formula

$$\begin{aligned} dX_t^n = &-\beta n X_t^{n-1}(X_t - \alpha)\,dt + \beta n(n-1)X_t^{n-2}(aX_t^2 + bX_t + c)\,dt \\ &+ nX_t^{n-1}\sigma(X_t)\,dW_t, \end{aligned}$$

and use that if $E(X_t^{2n})$ is finite, that is, if $a < (2n-1)^{-1}$, then the last term is a martingale with expectation zero.

Explicit formulae for the conditional moments of a Pearson diffusion are given by

$$E(X_t^n | X_0 = x) = \sum_{k=0}^{n} \left(\sum_{\ell=0}^{n} q_{n,k,\ell} e^{-a_\ell t} \right) x^k, \tag{6.3}$$

where $q_{n,k,n} = p_{n,k}$, $q_{n,n,\ell} = 0$ for $\ell \leq n-1$, and

$$q_{n,k,\ell} = -\sum_{j=k\vee\ell}^{n-1} p_{n,j} q_{j,k,\ell}$$

for $k, \ell = 0, \ldots, n-1$. Here $p_{n,n} = 1$, $p_{n,n+1} = 0$ and $\{p_{n,j}\}_{j=0,\ldots,n-1}$, solve the linear system of equations

$$(a_j - a_n) p_{n,j} = b_{j+1} p_{n,j+1} + c_{j+2} p_{n,j+2}.$$

This equation defines a simple recursive formula if $a_n - a_j \neq 0$ for all $j = 0, 1, \ldots, n-1$. Note that $a_n - a_j = 0$ if and only if there exists an integer $n-1 \leq m < 2n-1$ such that $a = m^{-1}$ and $j = m-n+1$. In particular, $a_n - a_j = 0$ cannot occur if $a < (n-1)^{-1}$, that is, if the nth moment exists. Note also that a_n is positive if and only if $a < (n-1)^{-1}$. The formula (6.3) can be proved by using that explicit polynomial eigenfunctions are available for the Pearson diffusions; for details see Wong (1964) or Forman and Sørensen (2008).

From a modeling point of view, it is important that the class of stationary distributions equals the full Pearson system of distributions. Thus a very wide spectrum of standard distributions is available as marginal distributions ranging from distributions with compact support to heavy-tailed distributions with tails of the Pareto type. The density μ of the stationary distribution of the process given by (6.1) solves the differential equation

$$\mu'(x) = -\frac{(2a+1)x - \alpha + b}{ax^2 + bx + c}\mu(x),$$

and the Pearson system is defined as the class of probability densities obtained by solving a differential equation of this form; see Pearson (1895).

The following list of the possible Pearson diffusions shows that all distributions in the Pearson system can be obtained as invariant distributions for a model in the class of Pearson diffusions. Note that if X_t solves (6.1), then $\tilde{X}_t = \gamma X_t + \delta$ is also a Pearson diffusion with parameters $\tilde{a} = a$, $\tilde{b} = b\gamma - 2a\delta$, $\tilde{c} = c\gamma^2 - b\gamma\delta + a\delta^2$, $\tilde{\beta} = \beta$, and $\tilde{\alpha} = \gamma\alpha + \delta$. Up to affine transformations, the Pearson diffusions can take the following forms.

Case 1: $\sigma^2(x) = 2\beta$**.** This is the Ornstein–Uhlenbeck process with invariant distribution equal to the *normal distribution* with mean α and variance 1.

Case 2: $\sigma^2(x) = 2\beta x$**.** This is the square root process (CIR process) with state space $(0, \infty)$. For $\alpha > 0$ the invariant distribution is the *gamma distribution* with scale parameter 1 and shape parameter α.

Case 3: $a > 0$ **and** $\sigma^2(x) = 2\beta a(x^2 + 1)$**.** The state space is the real line. The solution is ergodic for all $a > 0$ and all $\alpha \in \mathbb{R}$. The invariant density is given by

$\mu(x) \propto (x^2+1)^{-1/(2a)-1} \exp(\frac{\alpha}{a} \tan^{-1} x)$. If $\alpha = 0$, the invariant distribution is a scaled *t-distribution* with $\nu = 1 + a^{-1}$ degrees of freedom and scale parameter $\nu^{-1/2}$. If $\alpha \neq 0$, the invariant distribution is skew and has tails decaying at the same rate as the t-distribution with $1 + a^{-1}$ degrees of freedom. This distribution is a *skew t-distribution* known as *Pearson's type IV distribution*. Because of its skew and heavy tailed marginal distribution, the class of diffusions with $\alpha \neq 0$ is potentially very useful in many applications, for example, finance. It was studied and fitted to financial data by Nagahara (1996).

Case 4: $a > 0$ **and** $\sigma^2(x) = 2\beta a x^2$**.** The state space is $(0, \infty)$ and the process is ergodic if and only if $\alpha > 0$. The invariant distribution is the *inverse gamma distribution* with shape parameter $1 + \frac{1}{a}$ and scale parameter $\frac{a}{\alpha}$. This process is sometimes referred to as the GARCH diffusion model.

Case 5: $a > 0$ **and** $\sigma^2(x) = 2\beta a x(x+1)$**.** The state space is $(0, \infty)$. For $a > 0$ and $\alpha > 0$, the invariant distribution is a scaled *F-distribution* with $\frac{2\alpha}{a}$ and $\frac{2}{a} + 2$ degrees of freedom and scale parameter $\frac{\alpha}{1+a}$.

Case 6: $a < 0$ **and** $\sigma^2(x) = 2\beta a x(x-1)$**.** This is a *Jacobi diffusion* with state space $(0, 1)$. For all $a < 0$ and all $\alpha \in (0, 1)$ the invariant distribution is the *Beta distribution* with shape parameters $\frac{\alpha}{-a}$ and $\frac{1-\alpha}{-a}$.

Let X be a Pearson diffusion. If we define a new diffusion by the transformation $Y_t = T(X_t)$, where T is an invertible and twice continuously differentiable real function, then we can find the moments and conditional moments of $T^{-1}(Y_t)$. Thus we can find estimating functions based on predictions of powers of $T^{-1}(Y_t)$. Thus by transformations we obtain a very broad class of diffusions for which we can calculate prediction-based estimating functions explicitly. We illustrate this idea by a single example.

Example 6.1. If the transformation, $F(x) = \log(x/(1-x))$, is applied to the general Jacobi diffusion (Case 6), then we obtain a process that, by Ito's formula, solves the equation

$$dY_t = -\beta\{1 - 2\alpha + (1-\alpha)e^{Y_t} - \alpha e^{-Y_t} - 16a\cosh^4(Y_t/2)\}\,dt + 2\sqrt{-2a\beta}\cosh(Y_t/2)\,dW_t.$$

This is a diffusion for which the invariant distribution is the generalized logistic distribution with density

$$f(x) = \frac{e^{\kappa_1 x}}{(1+e^x)^{\kappa_1+\kappa_2} B(\kappa_1, \kappa_2)}, \qquad x \in \mathbb{R},$$

where $\kappa_1 = -(1-\alpha)/a$, $\kappa_2 = -\alpha/a$ and B denotes the Beta-function. This distribution was introduced and studied in Barndorff-Nielsen, Kent and Sørensen (1982).

7 Partially observed systems of stochastic differential equations

Let the D-dimensional process X be the solution to the stochastic differential equation (2.1), where, as usual, the parameter θ varies in a subset Θ of $\mathbb{R}^p$. We assume that X is stationary. In this section we will consider a number of examples, where X is not observed directly, but where we have observations of the form (2.2) or (2.3).

7.1 Discrete time observations with measurement errors

First we consider observations of the type (2.2), where k is real valued, that is, $d = 1$. Let us find a polynomial prediction-based estimating function of the type considered in Example 2.6. To find the minimum mean square error predictor, we must find mixed moments of the form (2.26). By the binomial formula,

$$\begin{aligned} E_\theta(Y_1^{k_1} Y_\ell^{k_2}) &= E_\theta\big(\big(k(X_{t_1}) + Z_1\big)^{k_1}\big(k(X_{t_\ell}) + Z_\ell\big)^{k_2}\big) \\ &= \sum_{i_1=0}^{k_1}\sum_{i_2=0}^{k_2}\binom{k_1}{i_1}\binom{k_2}{i_2} E_\theta(k(X_{t_1})^{i_1} k(X_{t_\ell})^{i_2}) E_\theta(Z_1^{k_1-i_1}) E_\theta(Z_\ell^{k_2-i_2}). \end{aligned}$$

Note that the distribution of the measurement error Z_i can depend on components of the unknown parameter θ. We need to find the mixed moments $E_\theta(k(X_{t_1})^{i_1} k(X_{t_2})^{i_2})$ $(t_1 < t_2)$, which can easily be determined by simulation.

Sometimes these mixed moments can be found explicitly. As an example, consider the situation where a Pearson diffusion (see Section 6) has been observed with measurement errors. In this case $k(x) = x$, and by (6.3)

$$\begin{aligned} E_\theta(X_{t_1}^{i_1} X_{t_2}^{i_2}) &= E_\theta(X_{t_1}^{i_1} E_\theta(X_{t_2}^{i_2} | X_{t_1})) \\ &= \sum_{k=0}^{i_2}\left(\sum_{\ell=0}^{i_2} q_{i_2,k,\ell} e^{-\lambda_\ell(t_2-t_1)}\right) E_\theta(X_{t_1}^{i_1+k}), \end{aligned} \tag{7.1}$$

where $E_\theta(X_{t_1}^{i_1+k})$ can be found by (6.2), provided that it exists.

In order to find the optimal polynomial prediction-based estimating function, we must find the mixed moments of the form (3.13), which can be calculated in a similar way and for a Pearson diffusion can be found explicitly.

A more complex example is when the coordinates of X are D independent diffusions given by

$$dX_{i,t} = -\beta_i(X_{i,t} - \alpha_i)\,dt + \sigma_i(X_{i,t})\,dW_{i,t}, \qquad i = 1, \ldots, D, \tag{7.2}$$

and where

$$Y_i = X_{1,t_i} + \cdots + X_{D,t_i} + Z_i.$$

The sum

$$S_t = X_{1,t} + \cdots + X_{D,t}$$

is a useful model because its autocorrelation function has D time-scales. Specifically, the autocorrelation function is

$$r(t) = \phi_1 \exp(-\beta_1 t) + \cdots + \phi_D \exp(-\beta_D t),$$

where $\phi_i = \mathrm{Var}(X_{i,t})/(\mathrm{Var}(X_{1,t}) + \cdots + \mathrm{Var}(X_{D,t}))$. An autocorrelation of this form is often found in observed time series. Examples are financial time series (see Barndorff-Nielsen and Shephard (2001)) and turbulence (see Barndorff-Nielsen, Jensen and Sørensen (1990) and Bibby, Skovgaard and Sørensen (2005)).

Again we must find mixed moments of the form (2.26). The measurement errors can be taken care of as above, so we need to calculate mixed moments of the type $E_\theta(S_{t_1}^\kappa S_{t_\ell}^\nu)$. By the multinomial formula,

$$E(S_{t_1}^\kappa S_{t_\ell}^\nu) = \sum\sum \frac{\kappa!}{\kappa_1! \cdots \kappa_D!} \frac{\nu!}{\nu_1! \cdots \nu_D!} E(X_{1,t_1}^{\kappa_1} X_{1,t_\ell}^{\nu_1}) \cdots E(X_{D,t_1}^{\kappa_D} X_{D,t_\ell}^{\nu_D}),$$

where the first sum is over $0 \le \kappa_1, \ldots, \kappa_D$ such that $\kappa_1 + \cdots + \kappa_D = \kappa$, and the second sum is analogous for the ν_i's. The higher-order mixed moments of the form (3.13) can be found by using a similar formula with four sums and four multinomial coefficients. Such formulae may appear daunting, but are easy to program. For a Pearson diffusion, mixed moments of the form $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$ can be calculated as explained above.

Example 7.1 (Sum of two skew t-diffusions). Consider a sum of two independent diffusions of the form (7.2) with $\alpha_i = 0$ and

$$\sigma_i^2(x) = 2\beta_i(\nu - 1)^{-1}\big(x^2 + 2\rho\sqrt{\nu}x + (1 + \rho^2)\nu\big),$$

$i = 1, 2$, where $\nu > 3$. This is one of the Pearson diffusions. The stationary distribution of $X_{i,t}$ is a skew t-distribution, ρ is the skewness parameter, and for $\rho = 0$ the stationary distribution is a t-distribution with ν degrees of freedom. To simplify the exposition we consider equidistant observations at time points $t_i = \Delta i$, and assume that there are no measurement errors, and that the value, $r(\Delta)$, of the autocorrelation function at time Δ is known. Then the optimal estimating function based on predictions of Y_i^2 with predictors of the form $\pi^{(i-1)} = a_0 + a_1 Y_{i-1}$ is

$$\sum_{i=2}^{n} \begin{bmatrix} Y_i^2 - 2(1+\rho^2)\nu/(\nu-2) - 4\rho\sqrt{\nu}r(\Delta)Y_{i-1}/(\nu-3) \\ Y_{i-1}Y_i^2 - Y_{i-1}2(1+\rho^2)\nu/(\nu-2) - 4\rho\sqrt{\nu}r(\Delta)Y_{i-1}^2/(\nu-3) \end{bmatrix}.$$

From this we can obtain estimators of ρ and ν. We can estimate $r(\Delta)$ by the value at time Δ of the empirical autocorrelation function based on the observations Y_i and insert this value in the expressions for $\hat{\rho}$ and $\hat{\nu}$. The remaining parameters can be estimated by fitting the theoretical autocorrelation function to the empirical autocorrelation functions, or by using an estimating function where more power functions of the data are predicted.

7.2 Integrated diffusions

Next we consider observations of the form (2.3), where k is real valued. Again we will find polynomial prediction-based estimating functions. Measurement errors can be treated exactly as in the previous subsection, so to simplify the presentation we will here assume that there is no measurement error.

To find the minimum mean square error predictor, we must find mixed moments of the form

$$E(Y_1^{k_1} Y_\ell^{k_2}) = \int_A E(k(X_{v_1}) \cdots k(X_{v_{k_1}}) k(X_{u_1}) \cdots k(X_{u_{k_2}}))\, du_{k_2} \cdots du_1\, dv_{k_1} \cdots dv_1,$$

where $1 \leq \ell$ and $A = [0, t_1]^{k_1} \times [t_{\ell-1}, t_\ell]^{k_2}$. Thus we need to calculate mixed moments of the type $E(k(X_{t_1}) \cdots k(X_{t_m}))$. Such mixed moments can be determined by simulation. In order to find the optimal polynomial prediction-based estimating function, we must find the mixed moments of the form (3.13). By a similar argument such mixed moments can also be expressed as an integral of mixed moments of the type $E(k(X_{t_1}) \cdots k(X_{t_m}))$.

If X is a Pearson diffusion and $k(x) = x$, these mixed moments can be calculated by a simple iterative formula obtained from (6.3) and (6.2), as explained in the previous subsection. Moreover, for the Pearson diffusions, $E(X_{t_1} \cdots X_{t_m})$ depends on $t_1, \ldots, t_m$ through sums and products of exponential functions; cf. (6.3) and (7.1). Therefore the integral above can be explicitly calculated, and thus explicit optimal estimating functions of the type considered in Example 2.6 are available for observations of integrated Pearson diffusions.

Estimation based on observations that are integrals of a diffusion ($D = d = 1$, $k(x) = x$) with no measurement error was studied by Bollerslev and Wooldridge (1992), Ditlevsen and Sørensen (2004) and Gloter (2006), while maximum likelihood estimation in the case of measurement errors was studied by Baltazar-Larios and Sørensen (2010).

An interesting more general case is that of hypoelliptic stochastic differential equations, where one or more components are not directly affected by the Wiener process and hence are smooth. If the smooth components are observed at discrete time points, then we obtain data of the type (2.3). Hypoelliptic stochastic differential equations are, for instance, used to model molecular dynamics; see, for example, Pokern, Stuart and Wiberg (2009). A simple example is the stochastic harmonic oscillator

$$\begin{aligned} dX_{1,t} &= -(\beta_1 X_{1,t} + \beta_2 X_{2,t})\, dt + \gamma\, dW_t, \\ dX_{2,t} &= X_{1,t}\, dt, \end{aligned}$$

$\beta_1, \beta_2, \gamma > 0$, where the position of the oscillator, X_2, is observed at discrete time points.

Example 7.2. Consider observations where $D = d = 1$ and $k(x) = x$, and where the diffusion process X is the square root process

$$dX_t = -\beta(X_t - \alpha)\,dt + \tau\sqrt{X_t}\,dW_t, \qquad X_0 > 0.$$

We will find a prediction-based estimating function with $f_1(x) = x$ and $f_2(x) = x^2$ and with predictors given by $\pi_1^{(i-1)} = a_{1,0} + a_{1,1}Y_{i-1}$ and $\pi_2^{(i-1)} = a_{2,0}$. Then the minimum mean square error predictors are

$$\breve{\pi}_1^{(i-1)}(Y_{i-1};\theta) = \mu\big(1 - a(\beta)\big) + a(\beta)Y_{i-1},$$
$$\breve{\pi}_2^{(i-1)}(\theta) = \alpha^2 + \alpha\tau^2\beta^{-3}\Delta^{-2}(e^{-\beta\Delta} - 1 + \beta\Delta)$$

with

$$a(\beta) = \frac{(1 - e^{-\beta\Delta})^2}{2(\beta\Delta - 1 + e^{-\beta\Delta})}.$$

The optimal prediction-based estimating function is

$$\sum_{i=1}^{n}\begin{pmatrix}1\\ Y_{i-1}\\ 0\end{pmatrix}\big[Y_i - \breve{\pi}_1^{(i-1)}(Y_{i-1};\theta)\big] + \sum_{i=1}^{n}\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}\big[Y_i^2 - \breve{\pi}_2^{(i-1)}(\theta)\big],$$

from which we obtain the estimators

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{a(\hat{\beta})Y_n - Y_1}{(n-1)(1 - a(\hat{\beta}))},$$

$$\sum_{i=2}^{n} Y_{i-1}Y_i = \hat{\alpha}\big(1 - a(\hat{\beta})\big)\sum_{i=2}^{n} Y_{i-1} + a(\hat{\beta})\sum_{i=2}^{n} Y_{i-1}^2,$$

$$\hat{\tau}^2 = \frac{\hat{\beta}^3\Delta^2\sum_{i=2}^{n}(Y_i^2 - \hat{\alpha}^2)}{(n-1)\hat{\alpha}(e^{-\hat{\beta}\Delta} - 1 + \hat{\beta}\Delta)}.$$

The estimators are explicit apart from $\hat{\beta}$, which can easily be found numerically by solving a nonlinear equation in one variable. For details, see Ditlevsen and Sørensen (2004).

7.3 Stochastic volatility models

Consider a stochastic volatility model given by

$$dX_t = (\kappa + \beta v_t)\,dt + \sqrt{v_t}\,dW_t,$$

where the volatility, v_t, is a stochastic process that cannot be observed directly. If the data are observations of X at the time points Δi, $i = 0, 1, 2, \ldots, n$, then $Y_i = X_{i\Delta} - X_{(i-1)\Delta}$ can be written in the form

$$Y_i = \kappa\Delta + \beta S_i + \sqrt{S_i}A_i,$$

where the A_i's are independent, standard normal distributed random variables, and where

$$S_i = \int_{(i-1)\Delta}^{i\Delta} v_t \, dt.$$

In order to find a polynomial prediction-based estimating function of the type considered in Example 2.6, we must find mixed moments of the form (2.26). We assume that v and W are independent, so that the sequences $\{A_i\}$ and $\{S_i\}$ are independent. By the multinomial formula we find that

$$E(Y_1^{k_1} Y_{t_1}^{k_2}) = \sum K_{k_{11},\ldots,k_{23}} E(S_1^{k_{12}+k_{13}/2} S_{t_1}^{k_{22}+k_{23}/2}) E(A_1^{k_{13}}) E(A_{t_1}^{k_{23}}),$$

where the sum is over all nonnegative integers k_{ij}, $i = 1, 2$, $j = 1, 2, 3$, such that $k_{i1} + k_{i2} + k_{i3} = k_i$ $(i = 1, 2)$, and where

$$K_{k_{11},\ldots,k_{23}} = \frac{k_1!}{k_{11}!k_{12}!k_{13}!} \frac{k_2!}{k_{21}!k_{22}!k_{23}!} (\kappa\Delta)^{k_{\cdot 1}} \beta^{k_{\cdot 2}}$$

with $k_{\cdot j} = k_{1j} + k_{2j}$. The moments $E(A_i^{k_{i3}})$ are the well-known moments of the standard normal distribution. When k_{i3} is odd, these moments are zero. Thus we only need to calculate the mixed moments of the form $E(S_1^{\ell_1} S_{t_1}^{\ell_2})$, where ℓ_1 and ℓ_2 are integers. The moments (3.13), which are needed to find the optimal polynomial prediction-based estimating function, can be obtained in a similar way. To calculate these, we need mixed moments of the form $E(S_1^{\ell_1} S_{t_1}^{\ell_2} S_{t_2}^{\ell_3} S_{t_3}^{\ell_4})$, where $\ell_1, \ldots, \ell_4$ are integers.

If the volatility model is a diffusion process, then S_i is an integrated diffusion, so such mixed moments can be calculated by the methods in Section 7.2. In particular, they can be calculated explicitly if the volatility process is a Pearson diffusion.

Acknowledgments

The research was supported by the Danish Center for Accounting and Finance funded by the Danish Social Science Research Council, by the Center for Research in Econometric Analysis of Time Series funded by the Danish National Research Foundation, and by a grant from the University of Copenhagen Programme of Excellence. The author is grateful to Susanne Ditlevsen for her thorough reading of the paper and several helpful comments.

References

Baltazar-Larios, F. and Sørensen, M. (2010). Maximum likelihood estimation for integrated diffusion processes. In *Contemporary Quantitative Finance*: *Essays in Honour of Eckhard Platen* (C. Chiarella and A. Novikov, eds.) 407–423. Berlin: Springer. MR2732856

Barndorff-Nielsen, O. E., Jensen, J. L. and Sørensen, M. (1990). Parametric modelling of turbulence. *Philos. Trans. R. Soc. Lond. Ser. A* **332**, 439–455.

Barndorff-Nielsen, O. E., Kent, J. and Sørensen, M. (1982). Normal variance–mean mixtures and z-distributions. *International Statistical Review* **50**, 145–159. MR0678296

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial econometrics (with discussion). *J. R. Stat. Soc. Ser. B* **63**, 167–241. MR1841412

Bibby, B. M., Skovgaard, I. M. and Sørensen, M. (2005). Diffusion-type models with given marginals and autocorrelation function. *Bernoulli* **11**, 191–220. MR2132002

Bollerslev, T. and Wooldridge, J. (1992). Quasi-maximum likelihood estimators and inference in dynamic models with time-varying covariances. *Econometric Review* **11**, 143–172. MR1185178

Brockwell, P. J. and Davis, R. A. (1991). *Time Series*: *Theory and Methods*. New York: Springer-Verlag. MR1093459

Ditlevsen, S. and Sørensen, M. (2004). Inference for observations of integrated diffusion processes. *Scand. J. Stat.* **31**, 417–429. MR2087834

Doukhan, P. (1994). *Mixing*, *Properties and Examples*. *Lecture Notes in Statist.* **85**. New York: Springer-Verlag. MR1312160

Durbin, J. (1960). Estimation of parameters in time-series regression models. *J. R. Stat. Soc. Ser. B* **22**, 139–153. MR0121950

Forman, J. L. and Sørensen, M. (2008). The Pearson diffusions: A class of statistically tractable diffusion processes. *Scand. J. Stat.* **35**, 438–465. MR2446729

Genon-Catalot, V., Jeantheau, T. and Laredo, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* **6**, 1051–1079. MR1809735

Gloter, A. (2006). Parameter estimation for a discretely observed integrated diffusion process. *Scand. J. Stat.* **33**, 83–104. MR2255111

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1212. MR0123385

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054. MR0666123

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *J. Econometrics* **30**, 203–238. MR0835578

Hansen, L. P. (2000). Method of moments. In *International Encyclopedia of the Social and Behavioral Sciences* (N. J. Smelser and P. B. Bates, eds.). Oxford: Pergamon.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. New York: Springer-Verlag. MR1461808

Jacod, J. and Sørensen, M. (2011). Aspects of asymptotic statistical theory for stochastic processes. Preprint report, Dept. Mathematical Sciences, Univ. Copenhagen.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. New York: Academic Press. MR0356197

Küchler, U. and Sørensen, M. (2009). Statistical inference for discrete-time samples from affine stochastic delay differential equations. Preprint report, Dept. Mathematical Sciences, Univ. Copenhagen.

Ljung, L. and Caines, P. E. (1979). Asymptotic normality of predictor error estimators for approximate system models. *Stochastics* **3**, 29–46. MR0546698

Nagahara, Y. (1996). Non-Gaussian distribution for stock returns and related stochastic differential equation. *Financial Engineering and the Japanese Markets* **3**, 121–149.

Pearson, K. (1895). Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material. *Philos. Trans. R. Soc. Lond. Ser. A* **186**, 343–414.

Pokern, Y., Stuart, A. M. and Wiberg, P. (2009). Parameter estimation for partially observed hypoelliptic diffusions. *J. R. Stat. Soc. Ser. B* **71**, 49–73. MR2655523

Reiß, M. (2002). Nonparametric estimation for stochastic delay differential equations. Ph.D. thesis, Institut für Mathematik, Humboldt-Universität zu Berlin.

Sørensen, H. (2003). Simulated likelihood approximations for stochastic volatility models. *Scand. J. Stat*. **30**, 257–276. MR1983125

Sørensen, M. (1999). On asymptotics of estimating functions. *Braz. J. Probab. Stat*. **13**, 111–136. MR1803041

Sørensen, M. (2000). Prediction-based estimating functions. *Econom. J*. **3**, 123–147. MR1820411

Sørensen, M. (2009). Parametric inference for discretely sampled stochastic differential equations. In *Handbook of Financial Time Series* (T. G. Andersen, R. A. Davis, J. P. Kreiss and T. Mikosch, eds.) 531–553. Heidelberg: Springer.

Sørensen, M. (2011). Estimating functions for diffusion-type processes. In *Statistical Methods for Stochastic Differential Equations* (M. Kessler, A. Lindner and M. Sørensen, eds.). London: Chapmann & Hall.

Wong, E. (1964). The construction of a class of stationary Markoff processes. In *Stochastic Processes in Mathematical Physics and Engineering* (R. Bellman, ed.) 264–276. Rhode Island: Amer. Math. Soc. MR0161375

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark
E-mail: michael@math.ku.dk
URL: http://www.math.ku.dk/~michael