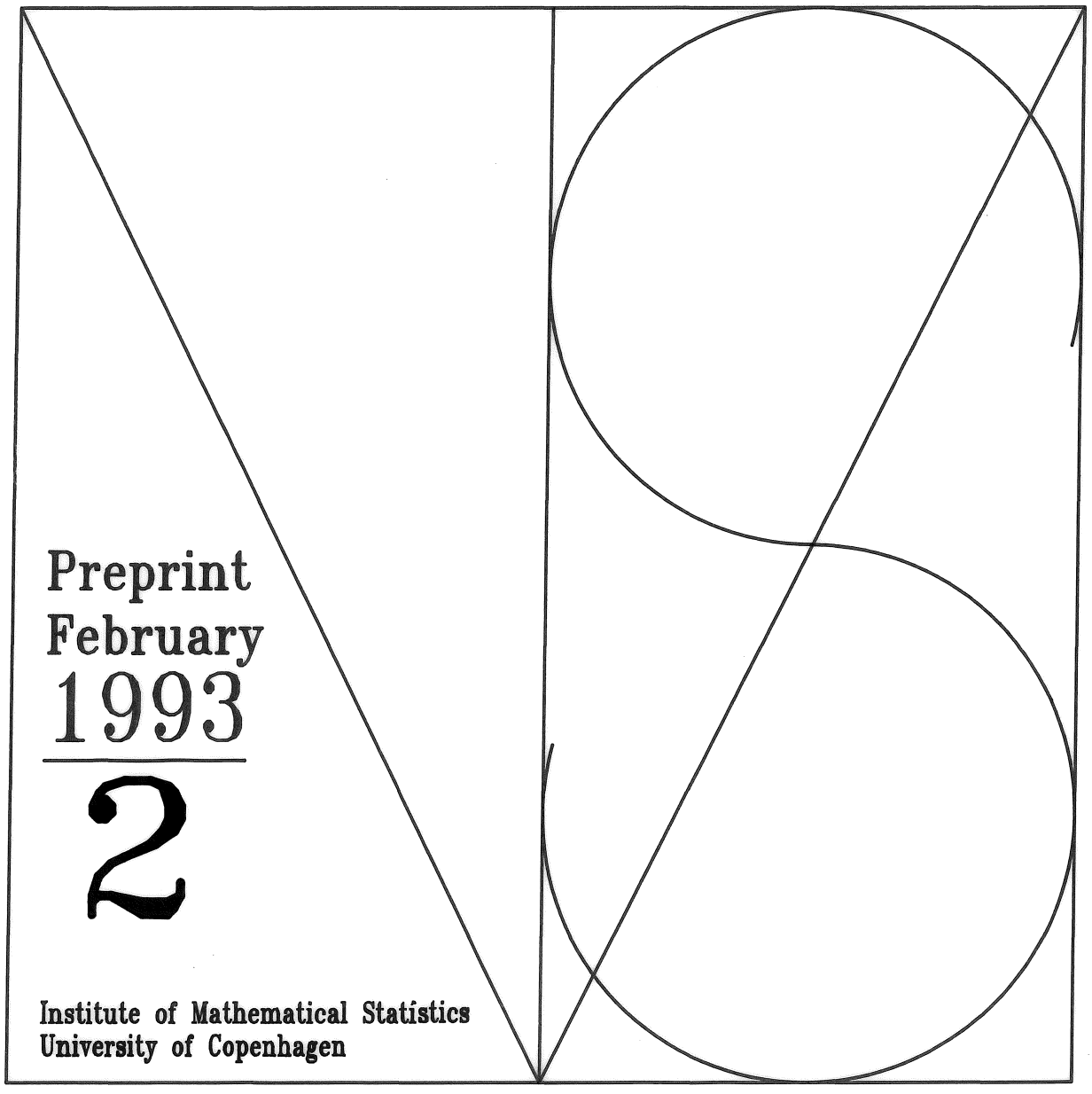
A. Stockmarr M. Jacobsen

Gaussian Diffusions and Autoregressive Processes: Weak Convergence and Statistical Inference



Preprint
February
1993
2

Institute of Mathematical Statistics
University of Copenhagen

ISSN 0902-8846

A. Stockmarr and M. Jacobsen

GAUSSIAN DIFFUSIONS AND AUTOREGRESSIVE PROCESSES: WEAK CONVERGENCE AND STATISTICAL INFERENCE

Preprint 1993 No. 2

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

February 1993

ABSTRACT. The statistical analysis of some hypothesis in models for d-dimensional homogeneous Gausian diffusions (HGD's) is discussed, and expressions for the MLE's and LR test statistics are derived. Regarded as distributions, the HGD's themselves are derived as weak limits of autoregressive processes, and the connection between the continous time (diffusion) and the discrete time (autoregressive) case is analysed. MLE's and LR test statistics in the two cases are connected by weak convergence.

Key words: Statistical models of Gaussian diffusions, approximation of autoregressive models, estimation, the hypothesis of r cointegrating vectors.

1. Introduction.

The processes constituting the main topic of the present paper are defined as solutions to the stochastic differential equation

$$dX(t) = (A + BX(t))\,dt + D\,dW(t), \qquad X(0) = x_0, \tag{1.1}$$

where W is a standard r-dimensional Brownian motion, defined on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ satisfying the "usual conditions", and where

- A is a $d \times 1$ coloumn vector.
- B, D are $d \times d$- respectively $d \times r$- matrices, $r \leq d$, $D \neq 0$.
- x_0 is the initial state of the proces, i.e. a point in $\mathbf{R}^d$.

It is well-known, that the solution to (1.1) is a timehomogeneous diffusion, which is Gaussian (an HGD), see e.g. Jacobsen (1991),and we shall regard (1.1) as a dominated statistical model, when A and B (not D, see section 3) varies.

Putting $\Lambda = DD^T$, with M^T denoting the transpose of M, the class of solutions to (1.1) is injectively parametrised by A,B and Λ. With $P^{x_0}_{A,B,\Lambda}$ the distribution of the corresponding solution, we thus consider the model

$$\left(P^{x_0}_{A,B,\Lambda} \in \mathcal{P}(C_{\mathbf{R}^d}[0;\infty)) \mid (A,B,\Lambda) \in \mathbf{R}^d \times \mathrm{Mat}_d(\mathbf{R}) \times H^+_d(\mathbf{R})\right), \tag{1.2}$$

with $\mathcal{P}(E)$ denoting the Borel probabilities on a topological space E. Here, the notation is

$C_{\mathbf{R}^d}[0;\infty)$: The space of continuous, $\mathbf{R}^d$-valued paths, defined on $[0;\infty)$;

$\mathrm{Mat}_d(\mathbf{R})$: The set of $d \times d$-matrices with real coefficients;

$\mathbf{H}^+_d(\mathbf{R})$: The set of positive semidefinite $d \times d$-matrices with real coefficients.

we are going to study the diffusions (1.1) not only in their own right, but also as approximations of first order autoregressive ($AR(1)$) processes. thus the model (1.2) is approximated by a sequence of autoregressive models, and the main purpose of the paper is the show how problems of statistical inference for the approximating sequence are related to inference in the diffusion model (1.2).

The approximations used are established by an invariance principle for weak convergence of $AR(1)$-processes to HGD's, which is presented in section 2. Section 3 is devoted to the problem of estmating the parameters in the diffusion model, and in section 4, having treated first the corresponding estimation problem for the approximating models, we then give one of our main results, Theorem 4.2, to the effect that the sequence of maximum likelihood estimators (MLE's) for the approximating $AR(1)$-models, converge in distribution to the MLE in the limiting diffusion model. (The result is formulated in terms of weak convergence of continuous time processes of estimators, obtained by considering MLE's when observing an approximating $AR(1)$-process or an HGD on an

interval $[0;t]$, where t varies in the open interval $(0;\infty)$. In the one-dimensional case, similar results has been obtained by Cox (1991).

In the final section 5 we study the hypothesis that the linear drift matrix B in (1.1) be of reduced rank, together with the corresponding hypothesis for the $AR(1)$-models. The latter is closely related to the concept of cointegration known from the econometrics and time series litterature. Our work here is much inspired by Johansen (1991), and apart from briefly discussing a cointegration theory for HGD's, we show the same type of consistency results as in section 4: The MLE's under the cointegration hypothesis for the $AR(1)$-models, together with the likelihood ratio test statistics, converge in distribution to the corresponding quantities for the diffusion model.

2. An Invariance Principle for Autoregressive Processes.

Consider a d-dimensional, discrete-time first order autoregressive process Y, satisfying

$$Y_0 \sim \nu \in \mathcal{P}(\mathbf{R}^d), \qquad \Delta Y_k = A + BY_{k-1} + \varepsilon_k, \quad k \in \mathbf{N}, \tag{2.1}$$

where $\varepsilon = (\varepsilon_k)_{k\in\mathbf{N}}$ is a noise-proces, consisting of independent, identically distributed random variables, with the 2.nd order moment representation

$$E(\varepsilon_1) = 0, \qquad V(\varepsilon_1) = \Lambda. \tag{2.2}$$

It is <u>not</u> assumed, that the distribution of ε_1 is Gaussian. Of course, Δ is the difference operator

$$(\Delta Y)_k = Y_k - Y_{k-1}, \quad k \in \mathbf{N}.$$

With (2.2) the only assumption about the error distribution, likelihood inference is out of the question. Yet, in some situations, it is possible to do some approximate likelihood inference.

The idea is to obtain the model (1.2) as the limit of models of the form (2.1), where under the limit the time points of observations get close to each other. To make this exact, we shall do the following :

Define processes $Y^{(n)}$ and $X^{(n)}$, $\quad n \in \mathbf{N}$, by

$$\begin{aligned} Y_0^{(n)} &:= \sqrt{n}Y_0, \\ \Delta Y_k^{(n)} &:= \tfrac{1}{\sqrt{n}}A + \tfrac{1}{n}BY_{k-1}^{(n)} + \varepsilon_k, \quad k \in \mathbf{N} \\ X^{(n)}(t) &:= \tfrac{1}{\sqrt{n}}Y_{[nt]}^{(n)}, \quad k \in \mathbf{N}, t \geq 0. \end{aligned} \tag{2.3}$$

Then $X^{(n)}$ jumps at the timepoints $(k/n)_{k\in\mathbf{N}}$ and is a right-continuous, piecewise constant counterpart to a discrete-time process $\widehat{X}$, satisfying

$$\begin{aligned} &\widehat{X}_0^{(n)} \sim \nu \in \mathcal{P}(\mathbf{R}^d), \\ &\Delta\widehat{X}_k^{(n)} = (A + B\widehat{X}_{k-1}^{(n)})\Delta t + \varepsilon_k^{(n)}, \quad k \in \mathbf{N}, \end{aligned} \tag{2.4}$$

where $(\varepsilon_k^{(n)})_{k\in\mathbf{N}}$ are i.i.d with $E(\varepsilon_1^{(n)}) = 0$, $V(\varepsilon_1^{(n)}) = 1/n\Lambda$, and where $\Delta t = 1/n$.

Now, if $\varepsilon_1 \sim N_d(0,\Lambda)$, $\varepsilon_k^{(n)}$ could be regarded as the increment over the time interval from $(k-1)/n$ to k/n for a process of the form $\Lambda^{1/2}W$, where W is a d-dimensional, standard Brownian motion, and $\Lambda^{1/2}$ is a square root of Λ. With an obvious notation (2.3) then translates into

$$X_0^{(n)} \sim \nu, \qquad \Delta X_k^{(n)} = (A + BX_{k-1}^{(n)})\Delta t + \Lambda^{\frac{1}{2}}\Delta W(t). \tag{2.5}$$

When n tends to infinity, Δt becomes small, and it is tempting to replace difference with differential and write

$$X(0) \sim \nu, \qquad dX(t) = (A + BX(t))\,dt + \Lambda^{\frac{1}{2}}\,dW(t), \tag{2.6}$$

suggesting that the proces $X^{(n)}$ converges in distribution on $D_{\mathbf{R}^d}[0,\infty)$, the space of Skorokhod- (cadlag-) paths, to the solution of (2.6).

This idea works, even in the more general setting with non-Gaussian errors, since the following holds :

Theorem 2.1.

As n tends to infinity, the sequence of processes $(X^{(n)})_{n\in\mathbf{N}}$ determined by (2.3), converges in distribution to the uniquely determined solution of the SDE (2.6), irrespectively of the distribution of ε_1. □

Proof.

We shall use Corollary 7.4.2 of Ethier & Kurtz (1986), which involves the probability transition function μ_n for the Markov Chain $\widehat{X}^{(n)}$ and its first and second order truncated moments. We have, that

$$\mu_n(x,\cdot) = \mathcal{L}(x + \tfrac{1}{n}(A + Bx) + \tfrac{1}{\sqrt{n}}\varepsilon_1).$$

With $\|\cdot\|$ the Euclidean norm on $\mathbf{R}^d$, for each $\delta > 0$ it holds that

$$\begin{aligned} &n\mu_n\big(x,\{y : \|x-y\| \geq \delta\}\big) \\ &= nP(\|\frac{1}{n}(A+Bx) + \tfrac{1}{\sqrt{n}}\varepsilon_1\| \geq \delta). \end{aligned} \tag{2.7}$$

Now if Z is a real-valued random variable with $E(|Z|) < \infty$, then $nP(|Z| > n) \to 0$ as $n \to \infty$. Using this and the fact that $E(\|\varepsilon_1\|^2) < \infty$, it follows that (2.7) converges to zero, uniformly on compacts (in x).

Putting

$$a_n(x) = n\int_{\{y:\|y-x\|\leq 1\}} (y-x)(y-x)^T\,\mu_n(x,dy) \tag{2.8}$$

$$b(x) = A + Bx$$

$$b_n(x) = n\int_{\{y:\|y-x\|\leq 1\}} (y-x)\,\mu_n(x,dy) \tag{2.9}$$

we have

$$a_n(x) \to \Lambda \qquad b_n(x) \to b(x)$$

uniformly on compacts:

Take $\|\cdot\|_\infty$ as the supremum norm on $\mathrm{Mat}_d(\mathbf{R})$. Then

$$\begin{aligned}
&\|a_n(x) - \Lambda\|_\infty \\
=&\|E\big(n(\tfrac{1}{n}(A+Bx) + \tfrac{1}{\sqrt{n}}\varepsilon_1)^{\otimes 2} 1_{\{\|\frac{1}{n}(A+Bx)+\frac{1}{\sqrt{n}}\varepsilon_1\| \le 1\}} - \Lambda\big)\|_\infty \\
\le& \tfrac{1}{n}\|A+Bx\|^2 \\
&+ \tfrac{2}{\sqrt{n}}\|A+Bx\|E(\|\varepsilon_1\|) \\
&+ \|E\big(\varepsilon_1^{\otimes 2} 1_{\|\frac{1}{n}(A+Bx)+\frac{1}{\sqrt{n}}\varepsilon_1\| \le 1\}} - \Lambda\big)\|_\infty \\
\le& \tfrac{1}{n}\|A+Bx\|^2 \qquad (2.10) \\
&+ \tfrac{2}{\sqrt{n}}\|A+Bx\|E(\|\varepsilon_1\|) \\
&+ \|E\big(\varepsilon_1^{\otimes 2} 1_{\|\frac{1}{n}(A+Bx)+\frac{1}{\sqrt{n}}\varepsilon_1\| > 1\}}\big)\|_\infty
\end{aligned}$$

because $E\varepsilon_1^{\otimes 2} = \Lambda$. In (2.10), the two first terms converges to 0, uniformly on compacts, and the third term converges to 0, using dominated convergence.The convergence of $b_n(x)$ follows similary.

Now consider $C_c^\infty(\mathbf{R}^d)$, the space of smooth functions with compact support, and the differential operator

$$\mathcal{L} : C_c^\infty(\mathbf{R}^d) \to C_c^\infty(\mathbf{R}^d)$$

given by

$$\mathcal{L}f = \tfrac{1}{2}\sum_{i,j=1}^{d} \Lambda_{ij}\partial_i\partial_j f + \sum_{i=1}^{d} b_i \partial_i f,$$

where $\partial_i = d/dx_i$, $\quad i = 1, \ldots, d$, and $b = (b_i)$ is the function

$$\begin{aligned} b :& \mathbf{R}^d \to \mathbf{R}^d \\ & x \mapsto A + Bx. \end{aligned}$$

The martingale problem for $\mathcal{L}$ is wellposed. This is a consequence of the fact, that a probability Q on the path-space $C_{\mathbf{R}^d}[0;\infty)$ of continous functions solves the martingale problem for $(\mathcal{L}, \nu)$, iff the coordinate process $(x_t)_{t\ge 0}$ solves the SDE (2.6) on the space $\big(C_{\mathbf{R}^d}[0;\infty), \mathbf{B}(C_{\mathbf{R}^d}[0;\infty)), Q\big)$, see Rogers & Williams (1987) V.19 – 22. Since (2.6) is an exact SDE, there exists exactly one probability defined on $\big(C_{\mathbf{R}^d}[0;\infty), \mathbf{B}(C_{\mathbf{R}^d}[0,\infty))\big)$ solving the martingale problem for $(\mathcal{L}, \nu)$.

Now, by Corollary 7.4.2 in Ethier & Kurtz (1986), the well-posedness and the three noted convergences of (2.7) – (2.9) are enough to ensure, that the processes $(X^{(n)})_{n\in\mathbf{N}}$ converge in distribution to the solution of the martingale problem for $(\mathcal{L}, \nu)$. As noted above this is exactly the unique solution of (2.6), and the theorem follows. □

Remark.

It is possible to generalize Theorem 2.1, assuming only that the error processes $\varepsilon^{(n)} = (\sum_{k=1}^{[nt]} \varepsilon_k)_{t\geq 0}$ converges in distribution to the Brownian motion $\Lambda^{1/2}W$. This is so, because the solutions to (2.3) may be written as $X^{(n)} = F_n(X^{(n)}(0), \varepsilon^{(n)})$, where the sequence of functions (F_n) converge pointwise on $\mathbf{R}^d \times D_{\mathbf{R}^d}[0;\infty)$ to the function F representing the strong solution to (2.6), $X = F(X(0), W)$. □

Theorem 2.1 makes it possible to use the model (1.2) as an approximative model for the phenomenon described by the process (2.3).

Limit theorems in sections 4 and 5 will justify, that it makes good sense to use the usual estimators (i.e. when the errors are Gaussian) and test statistics in the model for (2.3).

3. The Full Diffusion Model : The Set-up and Estimation.

Again, consider the space $C_{\mathbf{R}^d}[0;\infty)$, equipped with the σ-algebra generated by the coordinate projections $(\omega_t)_{t\geq 0}$. Letting $\mathcal{F}_t$ be the σ-algebra generated by $(\omega_s : 0 \leq s \leq t)$, we have the model

$$\left(P^{x_0,t}_{A,B,\Lambda} \in \mathcal{P}(C_{\mathbf{R}^d}[0;\infty), \mathcal{F}_t) \mid (A,B,\Lambda) \in \mathbf{R}^d \times \mathrm{Mat}_d(\mathbf{R}) \times H_d^+(\mathbf{R})\right)$$

for observing the process on [0;t] . In the sequel it is assumed, that t and the initial state x_0 are fixed.

Subsuming the initial condition, we put $P^t = P^{x_0,t}_{A,B,\Lambda}$, $\widetilde{P}^t = P^{x_0,t}_{\widetilde{A},\widetilde{B},\widetilde{\Lambda}}$ for some arbitrary parameters A, B, Λ, $\widetilde{A}$, $\widetilde{B}$ and $\widetilde{\Lambda}$, and write $\omega = (\omega_t)$ for the canonical process. From Liptser and Shiryayev (1977), see also Jacobsen (1991), Theorem 7.1, we have the following

Theorem 3.1.

(a) *Either* $P^t \sim \widetilde{P}^t$ *for all* $t > 0$ *or* $P^t \perp \widetilde{P}^t$ *for all* $t > 0$.

(b) *If* Λ *is nonsingular,* $P^t \sim \widetilde{P}^t$ *if and only if* $\Lambda = \widetilde{\Lambda}$, *and in that case we have*

$$\begin{aligned}\frac{d\widetilde{P}^t}{dP^t} =& \exp\Big(\int_0^t (\widetilde{Z}_s^T - \widetilde{Z}_s^T)\Lambda^{-1}\, dx_s \\ &- \frac{1}{2}\int_0^t (\widetilde{Z}_s^T\Lambda^{-1}\widetilde{Z}_s - Z_s^T\Lambda^{-1}Z_s)\, ds\Big)\end{aligned} \tag{3.1}$$

where $Z_s = A + B\omega_s$, $\widetilde{Z}_s = \widetilde{A} + \widetilde{B}\omega_s$. □

Whether Λ is regular or not is easily determined from a given observation : Subject to P, the quadratic variation on [0;t] of ω is $t\Lambda$, hence

$$\sum_{k=1}^{[2^n t]} (\omega_{k2^{-n}} - \omega_{(k-1)2^{-n}})^{\otimes 2} \to t\Lambda \tag{3.2}$$

$P^t_{A,B,\Lambda}$-a.s. , i.e. Λ can be read off from the observation. Therefore, in the remainder of this section, we make the following

Assumption.

Λ is known and non-singular. □

Thus, we arrive at the model

$$\left(P^{x_0,t}_{A,B,\Lambda} \in \mathcal{P}(C_{\mathbf{R}^d}[0;\infty), \mathcal{F}_t) \mid A \times B \in \mathbf{R}^d \times \mathrm{Mat}_d(\mathbf{R})\right) \tag{3.3}$$

with log-likelihood for observing a process X on $[0,t]$ given by

$$\begin{aligned} \ell_t(A,B) = &\int_0^t (A + BX(s))^T \Lambda^{-1}\, dX(s) \\ &- \tfrac{1}{2} \int_0^t (A + BX(s))^T \Lambda^{-1} (A + BX(s))\, ds \end{aligned} \tag{3.4}$$

based on the density with respect to the Brownian motion $P^{x_0,t}_{0,0,\Lambda}$ starting at x_0. (On the left side of (3.4) we have suppressed the initial state x_0, and we shall continue to do so).

Before finding the maximum-likelihood estimators, we introduce the following

Notation.

Let E_t and V_t denote the operators

$$\begin{aligned} E_t : &D_{\mathbf{R}^d}[0;\infty) \to \mathbf{R}^d \\ &\omega \mapsto \tfrac{1}{t} \int_0^t \omega(s)\, ds \end{aligned}$$

$$\begin{aligned} V_t : &D_{\mathbf{R}^d}[0;\infty) \to \mathrm{Mat}_d(\mathbf{R}) \\ &\omega \mapsto \int_0^t (\omega(s) - E_t(\omega))^{\otimes 2}\, ds \end{aligned}$$

Note, that E_t and V_t acts like expectation- and (except for a factor $1/t$) variance operators on the space of sample paths.

For any $k \in \mathbf{N}$ and any $\sigma \in H^+_k(\mathbf{R})$ we shall write $< \cdot, \cdot >_\sigma$ and $\| \cdot \|_\sigma$ for the inner product $(u,v) \mapsto u^T \sigma^{-1} v$, resp. the norm $u \mapsto < u, u >^{1/2}_\sigma$ on $\mathbf{R}^k$. Also, we shall denote the matrix norm $U \mapsto \mathrm{trace}(U^T U)$, $U \in \mathrm{Mat}_k(\mathbf{R})$, by $\| \cdot \|_*$. □

Maximizing ℓ_t partially in A gives us

$$\widehat{A}^{(t)}(B) = \tfrac{1}{t}(X(t) - x_0) - BE_t(X)$$

and the partially maximized likelihood

$$\begin{aligned}\ell_t(\widehat{A}^{(t)}(B), B) =& \tfrac{1}{2t}\|X(t) - x_0\|_\Lambda^2 + \int_0^t Y(s)^T B^T \Lambda^{-1}\, dX(s) \\ &- \tfrac{1}{2}\int_0^t \|BY(s)\|_\Lambda^2\, ds,\end{aligned} \tag{3.5}$$

where the process Y is defined as

$$Y(s) = X(s) - E_t(X).$$

Note, that although the process Y is not adapted to the filtration for X, the stochastic integral in (3.5) is well defined, treating $E_t(X)$ as a constant so that e. g.

$$\int_0^t Y^i(s)\, dX^j(s) := (X^i \bullet X^j)(t) - E_t^i(t)(X^j(t) - x_0^j),$$

using the "big dot"-notation for (Itô-) stochastic integrals.

Before we carry on with the estimation of B, note that the matrix $V_t(X)$ is positive definite a.s. : Let $\omega = X(w)$ be an observed path, for which the convergence (3.2) holds. Then

$$y^T V_t(\omega) y = 0$$

means, that $y^T\big(\omega(s) - E_t(\omega)\big)$ vanishes on $[0;t]$. Thus, y is perpendicular to $\omega(s_1) - \omega(s_2) \quad s_1, s_2 \leq t$, so (3.2) gives us $y^T \Lambda y = 0$, i.e. $y = 0$.

Next,note that $\ell_t(\widehat{A}^{(t)}(B), B)$ is a second order polynomial in B of the form

$$\ell_t(x) = \text{ const. } + < \rho, x >_\sigma - \tfrac{1}{2}\|x\|_\sigma^2,$$

where ρ is $(Y \bullet X^T)^T(t) V_t(X)^{-1} \in \mathrm{Mat}_d(\mathbf{R})$, intepreted as a vector, and the matrix σ is $\Lambda \otimes V_t(X)^{-1}$.

Having a σ, which is positive definite, ℓ_t attains its maximum at the unique point

$$\widehat{B}^{(t)} = \rho = (Y \bullet X^T)^T(t) V_t(X)^{-1}.$$

thus the ML-estimator for (A, B) is

$$(\widehat{A,B})^{(t)} = \big(\tfrac{1}{t}(X(t) - x_0) - \widehat{B}^{(t)} E_t(X),\ (Y \bullet X^T)^T(t) V_t(X)^{-1}\big) \tag{3.6}$$

(not depending on Λ), and the maximized log-likelihoodfunction becomes

$$\ell_t(\widehat{A}^{(t)}, \widehat{B}^{(t)}) = \tfrac{1}{2t}\|X(t) - x_0\|_\Lambda^2 + \frac{1}{2}\|(Y \bullet X^T)(t)\|_{\Lambda \otimes V_t(X)}^2. \tag{3.7}$$

(depending on Λ). We have proved

Theorem 3.2.
In the model (3.1), *the ML-estimator for* (A, B) *exists with probability 1, does not depend on* Λ *and is given by* (3.6). *The maximal value of the log-likelihood function is given by* (3.7). □

The model (3.3) has been studied by among others Le Breton (1977) and Le Breton & Musiela (1985) in the special case $A = 0$. In both papers the MLE $\widehat{B}^{(t)}$ is derived, and asymptotics for $t \mapsto \infty$ is studied.The estimator for B obtained by the two authors can be found by substituting X for Y in the formula (3.6).

4. Estimation in the autoregressive Processes. Consistency.

We first derive the estimator $\widehat{(A,B,\Lambda)}^{(n,t)}$ in the model for observing $X^{(n)}$ on [0;t], working conditionally on the initial state, and assuming the errors to be Gaussian. Then it is shown, that whether the errors are Gaussian or not, the process $(\widehat{A,B,\Lambda})^{(n)} := (\widehat{A,B,\Lambda})^{(n,t)}_{t\geq 0}$ converges in distribution to $(\widehat{A},\widehat{B},\Lambda) := (\widehat{A},\widehat{B},\Lambda)^{(t)}_{t\geq 0}$, the process of the estimators in the diffusion model found in section 3. This result justifies the use of the functionals $\widehat{(A,B,\Lambda)}^{(n,t)}$ as estimators in the general model (2.3), even though we do not know the distribution of the errors.

In preparation for this, we need a result about weak convergence. From (2.5) we obtain the representation

$$X^{(n)}(0) \sim \nu,\ X^{(n)}(t) = X^{(n)}(0) + \int_0^{[nt]/n} (A + BX^{(n)}(s))\,ds + \sum_{i=1}^{[nt]} \varepsilon_i^{(n)} \quad (4.1)$$

of the autoregressive processes with i.i.d., not necessarily Gaussian, errors, and similarly we write (2.6) as

$$X(0) \sim \nu, \qquad X(t) = X(0) + \int_0^t (A + BX(s))\,ds + \Lambda^{\frac{1}{2}} W(t). \quad (4.2)$$

With the representations (4.1) and (4.2) in mind, the following statement seems reasonable.

Lemma 4.1.
With $X^{(n)}$, X defined as in (4.1) , *respectively* (4.2), *the convergence*

$$(X^{(n)}, X^{(n)}_- \bullet X^{(n)^T}) \xrightarrow{\mathcal{D}} (X, X \bullet X^T)$$

holds, using the Skorokhod topology on $\mathbf{R}^d \times D_{\mathrm{Mat}_d(\mathbf{R})}[0;\infty)$. □

Warning.

The convergence fails, if instead of the piecewise constant $X^{(n)}$ used here, a continuous version with linear interpolation is used. See Kurtz & Protter (1991), Example 1.2. □

Proof of Lemma 4.1.
We use Theorem (2.2) of Kurtz & Protter (1991). At first we assume that ν is degenerate at x_0: $X^{(n)}(0) \equiv x_0 \in \mathbf{R}^d$, $n \in \mathbf{N}$. Decompose $X^{(n)}$ as

$$X^{(n)} = M^{(n)} + V^{(n)}$$

with $M^{(n)}(t) := \sum_{i=1}^{[nt]} \varepsilon_i^{(n)}$ being the martingale part of the process, and $V^{(n)}(t) := x_0 + \int_0^{[nt]/n}(A + BX^{(n)}(s))\,ds$. We shall prove, that for all $t \geq 0$

$$\sup_n \|E([M^{(n)}](t))\| < \infty \tag{4.3}$$

and

$$\sup_n E(T_t(V_j^{(n)})) < \infty \qquad j = 1, \ldots, d, \tag{4.4}$$

holds, with $[M^{(n)}]$ the quadratic variation process of $M^{(n)}$, $T(V_j^{(n)})$ the total variation of $V_j^{(n)}$, and as usual $\|\cdot\|$ the Euclidean norm. (In the notation of Kurtz & Protter, we have taken $\delta = \infty$, $\tau_n^\alpha = \infty$ for all α.)

(4.3) is trivial, whence we turn to (4.4). Let $T_t(V^{(n)})$ denote the d-vector with coordinates $T_t(V_j^{(n)})$.We shall show that

$$\sup_n E\|T_t(V^{(n)})\| < \infty. \tag{4.5}$$

To obtain this we need to express $X^{(n)}$ as a function of x_0 and $\varepsilon^{(n)}$: Solving the difference equation (2.4) we get, that

$$\begin{aligned} X^{(n)}(t) =& (I + \tfrac{1}{n}B)^{[nt]}x_0 + \tfrac{1}{n}\sum_{i=1}^{[nt]}(I + \tfrac{1}{n}B)^{[nt]-i}A \\ &+ \sum_{i=1}^{[nt]}(I + \tfrac{1}{n}B)^{[nt]-i}\varepsilon_i^{(n)} \quad . \end{aligned}$$

For $s = k/n$ it follows that

$$\begin{aligned} \Delta V^{(n)}(s) =& \Delta X^{(n)}(s) - \varepsilon_k^{(n)} \\ =& \tfrac{1}{n}(A + BX^{(n)}(\tfrac{k-1}{n})) \end{aligned}$$

and hence

$$\begin{aligned} \|T_t(V^{(n)})\| &\leq \sum_{0 \leq s \leq t} \|\Delta V^{(n)}(s)\| \\ &\leq \tfrac{1}{n}\sum_{k=1}^{[nt]}\big(\|A + B(I + \tfrac{1}{n}B)^{k-1}x_0 \\ &\qquad + \tfrac{1}{n}\sum_{i=1}^{k-1} B(I + \tfrac{1}{n}B)^{k-1-i}A\|\big) \qquad (4.6) \\ &\quad + \tfrac{1}{n}\sum_{k=1}^{[nt]}\|\sum_{i=1}^{k-1} B(I + \tfrac{1}{n}B)^{k-1-i}\varepsilon_i^{(n)}\|. \qquad (4.7) \end{aligned}$$

The non-random term (4.6) converges, and the mean of the square of the random term (4.7) is

$$
\begin{aligned}
&E\big(\tfrac{1}{n}\sum_{k=1}^{[nt]}\|\sum_{i=1}^{k-1}B(I+\tfrac{1}{n}B)^{k-i-1}\varepsilon_i^{(n)}\|\big)^2\\
\leq&\tfrac{[nt]}{n^2}\sum_{i=1}^{[nt]}E\|\sum_{i=1}^{k-1}B(I+\tfrac{1}{n}B)^{k-i-1}\varepsilon_i^{(n)}\|^2\\
=&\tfrac{[nt]}{n^2}\sum_{k=1}^{[nt]}\sum_{i=1}^{k-1}E\|B(I+\tfrac{1}{n}B)^{k-i-1}\varepsilon_i^{(n)}\|^2 \qquad (4.8)\\
\leq&\tfrac{[nt]}{n}\|B\|^2e^{2t\|B\|}\mathrm{trace}(\Lambda)\tfrac{1}{n^2}\sum_{k=1}^{[nt]-1}(k-1), \qquad (4.9)
\end{aligned}
$$

where we use the independence of the errors to obtain (4.8).

Since (4.9) converges, the expectation of (4.7) is bounded in n, and (4.5) follows. Thus we have proved the lemma in the case of a degenerate initial distribution.

For general ν, take $f \in C_b(D_{\mathbf{R}^d\times Mat_d(\mathbf{R})}[0;\infty))$ (f is continous and bounded). Then

$$
\begin{aligned}
&E(f(X^{(n)},X_-^{(n)}\bullet X^{(n)^T}))\\
=&\int E(f(X^{(n)},X_-^{(n)}\bullet X^{(n)^T})\mid X^{(n)}(0)=x)\,d\nu(x)\\
\to&\int E(f(X,X\bullet X^T)\mid X(0)=x)\,d\nu(x)\\
=&E(f(X,X\bullet X^T))
\end{aligned}
$$

by the above and dominated convergence. The lemma follows. □

We return to the estimation problem in the n'th autoregressive model (2.3). *Assuming Gaussian errors*, the likelihood function is

$$
\begin{aligned}
&L_i^{(n)}(A,B,\Lambda)=\\
&\det(\tfrac{1}{n}\Lambda)^{-\frac{[nt]}{2}}\exp\big(-\tfrac{1}{2}\sum_{i=1}^{[nt]}\|X^{(n)}(\tfrac{i}{n})-(\tfrac{1}{n}A+(I+\tfrac{1}{n}B))X^{(n)}(\tfrac{i-1}{n})\|^2_{\frac{1}{n}\Lambda}\big),
\end{aligned}
$$

with a log-likelihood that is easily seen to have the form

$$\ell_t^{(n)}(A,B,\Lambda) = -\tfrac{[nt]}{2}\log\det(\tfrac{1}{n}\Lambda) - \tfrac{1}{2}\sum_{i=1}^{[nt]} \|X^{(n)}(\tfrac{i}{n}) - X^{(n)}(\tfrac{i-1}{n})\|^2_{\frac{1}{n}\Lambda}$$
$$+ \int_0^{\frac{[nt]}{n}} (A + BX^{(n)}(s-)\, dX^{(n)}(s)$$
$$- \tfrac{1}{2}\int_0^{\frac{[nt]}{n}} \|(A + BX^{(n)}(s)\|^2_\Lambda \, ds,$$

an expression which, as a function af A and B, has the same structure as the log-likelihood (3.4) for the diffusion model. It follows that the maximum likelihood estimator for (A,B) exists (provided $[nt] \geq d$ so that $V_{[nt]/n}(X^{(n)})$ is non-singular), does not depend on Λ and is given by an expression similar to (3.6). Standard methods from multivariate analysis then gives the MLE for Λ. Defining the processes

$$(\widehat{A},\widehat{B},\widehat{\Lambda})^{(n)}(t) := (\widehat{A}^{(n,t)},\widehat{B}^{(n,t)},\widehat{\Lambda}^{(n,t)}),$$
$$(\widehat{A},\widehat{B},\widehat{\Lambda})(t) := (\widehat{A}^{(t)},\widehat{B}^{(t)},\Lambda),$$

we end up with the following

Theorem 4.2.

(a) *In the model (2.3) with i.i.d. Gaussian errors, the MLE $(\widehat{A},\widehat{B},\widehat{\Lambda})^{(n,t)}$ exist with probability 1, provided $[nt] \geq d$, and coincides with the functional $(\widehat{A}^{(n,t)},\widehat{B}^{(n,t)},\widehat{\Lambda}^{(n,t)})$ given by*

$$\widehat{A}^{(n,t)} = \tfrac{n}{[nt]}(X^{(n)}(\tfrac{[nt]}{n}) - x_0^{(n)}) - \widehat{B}^{(n,t)}E_{\frac{[nt]}{n}}(X^{(n)})$$
$$\widehat{B}^{(n,t)} = (Y_-^{(n)} \bullet X^{(n)^T})^T(t) V_{\frac{[nt]}{n}}(X^{(n)})^{-1}$$
$$\widehat{\Lambda}^{(n,t)} = \tfrac{n}{[nt]}\sum_{i=1}^{[nt]}(X^{(n)}(\tfrac{i}{n}) - X^{(n)}(\tfrac{i-1}{n}))^{\otimes 2}$$
$$- \tfrac{1}{[nt]}(X^{(n)}(t) - x_0)^{\otimes 2}$$
$$- \tfrac{1}{[nt]}\widehat{B}^{(n,t)}Y_-^{(n)} \bullet X^{(n)^T}(t),$$

where
- *$Y^{(n)}$ is the process $X^{(n)} - E_{\frac{[nt]}{n}}(X^{(n)})$;*
- *$Y_-^{(n)}$ is the process $Y_-^n(s) = Y^{(n)}(s-)$.*

(b) *The sequence* $(\widehat{A},\widehat{B},\widehat{\Lambda})^{(n)}$ *is consistent in the sense that*

$$(\widehat{A},\widehat{B},\widehat{\Lambda})^{(n)} \xrightarrow{\mathcal{D}} (\widehat{A},\widehat{B},\Lambda), \tag{4.10}$$

in the Skorokhod topology on $D_{\mathbf{R}^d\times \mathrm{Mat}_d(\mathbf{R})\times \mathbf{H}_d^+(\mathbf{R})}(0;\infty)$, *irrespectively of the distribution of* ε_1. □

Remark.

In part (b) itis essential to consider the *open* time interval $(0;\infty)$. With a continuous limit process, to show convergence on the Skorokhod space $D_E(0;\infty)$, where E is any Polish space, amounts to showing convergence on $D_E[s;t]$ for the processes restricted to $[s;t]$ for any $0<s<t<\infty$. □

Proof.

The convergence of $(\widehat{A}^{(n)},\widehat{B}^{(n)})$ follows directly from Lemma 4.1, since $(\widehat{A}^{(n)},\widehat{B}^{(n)})$ is a function of $(\frac{[nt]}{n}, X^{(n)}, X_-^{(n)} \bullet X^{(n)^T})$, which is a.s. continous in the limit $(t, X, X\bullet X^T)$.

The essential part of $\widehat{\Lambda}^{(n)}$ is

$$\begin{aligned}&\sum_{i=1}^{[nt]}(X^{(n)}(\tfrac{i}{n}) - X^{(n)}(\tfrac{i-1}{n}))^{\otimes 2}\\ &= X^{(n)}(t)^{\otimes 2} - X^{(n)}(0)^{\otimes 2} - (X_-^{(n)}\bullet X^{(n)^T})^T(t) - (X_-^{(n)}\bullet X^{(n)^T})(t),\end{aligned}$$

which by Lemma 4.1 tends to the process

$$X(t)^{\otimes 2} - X(0)^{\otimes 2} - (X\bullet X^T)^T(t) - (X\bullet X^T)(t) = t\Lambda$$

in distribution, so that $\Lambda^{(n)}$ converges in distribution to the process $t\mapsto\Lambda$ on $D_{\mathbf{H}_d^+(\mathbf{R})}(0;\infty)$ with the Skorokhod topology. Since this is a continuous, non-stochastic process, the simultaneous convergence (4.10) follows. □

Theorem 4.2 generalizes Theorem 3.2 in Cox (1991). Cox deals with weak convergence in dimension 1 of variables similar to ours, but for t kept fixed.

5. The Hypothesis of Reduced Rank of B. Consistency and relations to cointegration theory.

Again, we consider the model

$$X(0)=x_0, \qquad dX(t) = (A+BX(t))dt + \Lambda^{\frac{1}{2}}\,dW(t), \tag{5.1}$$

still assuming Λ to be nonsingular.

5.1. Intepretation.

We are interested in the hypothesis

$$H_r : \text{rank}(B) \leq r, \tag{5.2}$$

where $r \leq d$, or, in parametric terms:

$$H_r : B = \alpha\beta^T, \quad \alpha, \beta \in \text{Mat}_{d\times r}(\mathbf{R}). \tag{5.3}$$

What makes H_r interesting is the following fact :

Proposition 5.1.
Suppose that Λ is regular, and that X is given by (5.1). *Then the following are equivalent :*

(i) *rank$(B) \leq r$.*

(ii) *There exists $v_1, \ldots, v_{d-r} \in \mathbf{R}^d$ linearly independent, so that $V^T(X - x_0)$ is a $(d-r)$-dimensional Brownian motion with a constant drift, where V denotes the matrix $V = (v_1 : \cdots : v_{d-r})$ with columns $v_1, \ldots, v_{d-r}$.*

□

Proof.

$(i) \Rightarrow (ii)$: There exists $v_1, \ldots, v_{d-r} \in \mathbf{R}^d$ lineary independent so that

$$v_i^T B = 0, \quad i = 1, \ldots, d-r.$$

We claim that $V^T(X - x_0)$ is a Brownian motion with constant drift $V^T A$: According to Jacobsen (1991), Theorem 2.7, we must show, that the process U, where

$$U(t) = V^T X(t) - tV^T A$$

is a local martingale. But, since $V^T B = 0$,

$$dU(t) = V^T \Lambda^{\frac{1}{2}}\, dW(t),$$

this is evident.

$(ii) \Rightarrow (i)$: Let $\xi \in \mathbf{R}^{d-r}$ be the constant drift vector for $V^T(X - x_0)$. Then the process

$$U(t) := V^T(X(t) - x_0) - t\xi$$

is a martingale. Since

$$dU = (V^T A - \xi + V^T BX)dt + V^T \Lambda^{\frac{1}{2}} dW,$$

it follows that

$$V^T A - \xi + V^T BX(t) \equiv 0 \qquad a.s.,$$

but since $V_t(X)$ is *a.s.* positive definite for all $t > 0$, X must leave the affine subspace $(V^T B)^{-1}(V^T A - \xi)$, if this differs from $\mathbf{R}^d$. Consequently, $V^T B = 0$ and hence rank$(B) \leq r$. □

Theorem 5.1 shows, that rank$(B) \leq r$ iff there exists a $(d-r)$-dimensional component of the process, which acts like a "driving force" in the sense of a Brownian motion with a constant drift. Another interesting consequence of the hypothesis H_r relies on the notion of cointegration:

5.2. Ouverture to Cointegration theory.

Consider once again the autoregressive process (cf. (2.4))

$$\widehat{X}_k^{(n)} = \tfrac{1}{n}A + (I + \tfrac{1}{n}B)\widehat{X}_{k-1} + \varepsilon_k^{(n)}, \tag{5.4}$$

with i.i.d. errors,possibly non-Gaussian, use the parametrisation (5.3) and define $\alpha_\perp, \beta_\perp$ to be injective $d \times (d-r)$-matrices, with columns perpendicular to the columns of α, resp. β.

Take $\phi_n(\lambda) := I - \lambda(I + 1/nB)$ for any complex number λ, and make the following three assumptions, which are referred to as the hypothesis of the existence of r cointegration vectors :

- $(a1)$ rank$(B) = r$, so that α and β are injective ;
- $(a2)$ $\det(\phi_n(\lambda)) = 0$ implies $\lambda = 1$ or $|\lambda| > 1$, so that the only allowed root not outside the closed unit disc is 1;
- $(a3)$ $\alpha_\perp^T \beta_\perp$ is of full rank $d - r$.

Then the so-called Granger representation (Engle & Granger (1987)) implies, that there exists two set of "directions", the span of the columns of respectively $\alpha_\perp$ and β, so that in the first direction, the process $\alpha_\perp^T X^{(n)}$ is a random walk with a constant drift, and in the second direction, the process $\beta^T X^{(n)}$ is stationary (that is, admits a stationary initial distribution).

The r columns of β are called the cointegrating vectors for B, and they are of course not uniquely determined. But the span of them is uniquely defined, and is called the cointegration space.

The assumptions $(a1) - (a3)$ ensures, that the process (5.4) is integrated of order 1 (denoted I(1)), that is, the differenced process $\Delta\widehat{X}^{(n)}$ is stationary, and $\widehat{X}^{(n)}$ is non-stationary. We shall not go into details with these concepts apart from noting, that by substituting $(a3)$ with other assumptions, one obtains I(k)-processes for $k > 1$.

Note, that $(a1)$ implies

$$\begin{aligned} \text{span}(B) &= \text{span}(\alpha), \\ \text{span}(B^T) &= \text{span}(\beta), \end{aligned} \tag{5.5}$$

and that $(a2)$ is equivalent to

$(a2')$ $\text{spec}(B) \subset \{0\} \cup K(-n, n)$,

where spec(M) is the spectrum in the complex plane, defined as

$$\text{spec}(M) = \{\lambda \in \mathbf{C} : Mz = \lambda z \,\text{for some}\, z \in \mathbf{C}^d \setminus \{0\}\},$$

and where $K(-n,n) = \{z \in \mathbf{C} \mid |z+n| < n\}$. Also, note that $(a3)$ is equivalent to

$(a3')$ $(\alpha_\perp, \beta)$ form a base for $\mathbf{R}^d$

or

$(a3'')$ $(\alpha, \beta_\perp)$ form a base for $\mathbf{R}^d$,

so that the "driving force"-part and the stationary part of the autoregressive describes it completely: To see this, note that $(a3)$ is equivalent to $\beta_\perp^T \alpha_\perp$ being injective, so that $a \in \mathbf{R}^{d-r}$, $a \neq 0$ implies that $\beta_\perp^T \alpha_\perp a \neq 0$, i.e. that, writing $< \cdot, \cdot >$ for the usual inner product on $\mathbf{R}^d$,

$$\sum_{i=1}^{d-r} a_i < \alpha_{\perp,i}\,,\, \beta_{\perp,j} > = < \sum_{i=1}^{d-r} a_i \alpha_{\perp,i}\,,\, \beta_{\perp,j} > \neq 0 \tag{5.6}$$

for some $j \in \{1, \ldots, r\}$. Therefore

$$\text{span}(\alpha_\perp) \cap \text{span}(\beta_\perp)^\perp = \text{span}(\alpha_\perp) \cap \text{span}(\beta) = 0\,, \tag{5.7}$$

so that the columns in $\alpha_\perp$ and β are linearly independent. Since both $\alpha_\perp$ and β are injective, $(a3')$ follows.

Conversely, $(a3')$ implies (5.7), so that (5.6) holds for every $a \in \mathbf{R}^{d-r} \setminus \{0\}$ and every choice of β, which implies $(a3)$. The equivalence of $(a3')$ and $(a3'')$ is immediate.

Theorem 5.2.

There exists α,$\beta \in Mat_{d\times r}(\mathbf{R})$, so that $B = \alpha\beta^T$ and the assumptions $(a1) - (a3)$ holds for some $n \in \mathbf{N}$, (and then automatically for all larger n), if and only if there exists a base $(\tilde{\alpha}, \tilde{\beta})$ for $\mathbf{R}^d$, so that with $\tilde{\alpha}$,$\tilde{\beta}$ the matrices corresponding to the first $d-r$, respectively last r vectors of the base, the following two conditions hold:

- $(b1)$ *$\tilde{\alpha}^T X$ is a $(d-r)$-dimensional Brownian motion with a constant drift;*
- $(b2)$ *$\tilde{Y} := \tilde{\beta}^T X$ is a homogeneous Gaussian diffusion, that admits an initial distribution that makes it stationary.*

If this is the case, $(\tilde{\alpha}, \tilde{\beta})$ may be taken as $(\alpha_\perp, \beta)$. The space span$(\tilde{\beta})$ equals the cointegration space for B. □

In Theorem 5.2 and the preceding we have identified matrices with the bases consisting of the columns of the corresponding matrices.

Proof.

Suppose that $(a1)-(a3)$ holds for some (fixed) n. Then, as in the proof of Theorem 5.1, the process $\alpha_\perp^T X$ is Brownian with a constant drift. Next, note that the process $Y := \beta^T X$ satisfies

$$dY(t) = (\beta^T A + \beta^T \alpha Y(t))dt + \beta^T \Lambda^{\frac{1}{2}} dW(t).$$

Thus Y is itself a homogeneous Gaussian diffusion, and it is well known, that Y may be given an initial distribution that makes it stationary iff

$$\text{spec}(\beta^T \alpha) \subset \{\lambda \in \mathbf{C} : \text{re}\lambda < 0\}, \tag{5.8}$$

see e.g. Theorem 6.2. in Jacobsen (1991). We shall verify (5.8).

Suppose that $\lambda \in \text{spec}(\beta^T \alpha)$, with a corresponding eigenvector $z \in \mathbf{C}^r \setminus \{0\}$. Then $B(\alpha z) = \alpha(\beta^T \alpha z) = \lambda \alpha z$, so that $\lambda \in \{0\} \cup K(-n, n)$. And $\lambda = 0$ is impossible: If so, $\alpha z \in \ker(B)$, the kernel space for B, which by (5.5) equals $\text{span}(\beta_\perp)$. But also $\alpha z \in \text{span}(\alpha)$ holds, and since $\alpha z \neq 0$ by $(a1)$, a contradiction to $(a3'')$ is reached. Taking $(\tilde{\alpha}, \tilde{\beta})$ as $(\alpha_\perp, \beta)$ completes the first part of the proof.

Conversely, suppose that $(b1)-(b2)$ holds. Then $(b1)$ and Theorem 5.2 forces $\text{rank}(B) \leq r$. And $\text{rank}(B) < r$ cannot occur:

Consider the process $\widetilde{Y}$. Since $\tilde{\beta}$ is of full rank r, $\widetilde{Y}$ is Markov with Gaussian transition probabilities of full rank r (since those of X are of full rank d). Therefore, $\widetilde{Y}$ must solve an SDE of the form

$$d\widetilde{Y}(t) = (\widetilde{A} + \widetilde{B}\widetilde{Y}(t))dt + \widetilde{\Lambda}^{\frac{1}{2}} d\widetilde{W}(t)$$

(where $\widetilde{W}$ is standard r-dimensional Brownian motion) with respect to a suitable filtration, for some $\widetilde{A}, \widetilde{B}, \widetilde{\Lambda}$, where $\widetilde{B}$ according to $(b2)$ is nonsingular. Furthermore, Corollary 2.11 & Proposition 5.15 in Jacobsen (1991) together implies, that

$$\widetilde{B}\tilde{\beta} = \tilde{\beta}^T B \tag{5.9}$$

so that $\text{rank}(B) \geq \text{rank}(\widetilde{B}\tilde{\beta}) = r$ by the injectivity of $\tilde{\beta}$. $(a1)$ follows, and there exists $\alpha, \beta \in \text{Mat}_{d \times r}(\mathbf{R})$ injective, so that $B = \alpha\beta^T$. Inserting this in (5.9) gives us

$$\tilde{\beta}^T = T\beta^T,$$

where $T = \widetilde{B}^{-1}\tilde{\beta}\alpha \in \text{Mat}_r(\mathbf{R})$ must have full rank. Thus $\text{span}(\beta) = \text{span}(\tilde{\beta})$, and since $\tilde{\alpha}^T$ is Brownian, we have $\tilde{\alpha}^T B = 0$ (cf. the proof of Theorem 5.1). We may therefore take $\alpha_\perp := \tilde{\alpha}$, so that $(a3')$ and hence $(a3)$ follows.

Finally, $(a2)$: Since $\text{span}(\tilde{\beta}) = \text{span}(\beta)$, we may take $\tilde{\beta}$ as β. The stationarity of Y forces $\beta^T \alpha$ to be regular, whence $\text{spec}(\beta^T \alpha)$ lies in the circle $K(-n, n)$

for some $n \in \mathbf{N}$ (and hence for all larger n), and we may assume, that $-n \notin \text{spec}(\beta^T\alpha)$. Then the roots of $\det(\phi_n(\lambda))$ are all either 1 or outside the the unit disc, so that $(a2)$ is fullfilled.

The last assertions are proved in the above. □

The idea with the autoregressive model (5.4) is that it describes the data for some large n. The preceding Theorem shows, that the hypothesis of r cointegration vectors is equivalent to the existence of a decomposition of the process into a stationary part and a "driving force"-part, acting like a Brownian motion with a constant drift.The discrete-time study of the hypothesis of r cointegration vectors is carried out in details in Johansen (1991), and relies on the notion of I(1)-processes. Even though it is difficult to find a notion similar to that of I(1) in continous time, there is no trouble in detecting the stationary component of the process, and hence we can define cointegrating vectors in continuous time by copying the discrete-time definition.

5.3. Estimation, tests and consistency.

We shall now turn to the problem of estimating A and B subject to H_r. The result here is the following

Theorem 5.3.

If $\lambda_{(d)} \geq \cdots \geq \lambda_{(1)}$ are the ordered eigenvalues of the matrix

$$M = V_t(X)^{-\frac{1}{2}}(Y \bullet X^T)(t)\Lambda^{-1}(Y \bullet X^T)^T(t)V_t(X)^{-\frac{1}{2}}, \tag{5.10}$$

the ML-estimator $(\widehat{A, B})^{(t)}$ in the model (5.1) under the hypothesis H_r exists a.s. iff $\lambda_{(d-r+1)} \neq \lambda_{(d-r)}$. If this is the case, the estimator is given by

$$\widehat{A}^{(t)} = \tfrac{1}{t}(X(t) - x_0) - \widehat{B}^{(t)}E_t(X) \tag{5.11}$$

$$\widehat{B}^{(t)} = (Y \bullet X^T)^T(t)V_t(X)^{-\frac{1}{2}}P_rV_t(X)^{-\frac{1}{2}} \tag{5.12}$$

where P_r is the matrix representing the usual orthogonal projection on the eigenspace corresponding to the eigenvalues $\lambda_{(d-r+1)}, \ldots, \lambda_{(d)}$. The value of the maximized log-likelihood is

$$\ell_t(\widehat{A}^{(t)}, \widehat{B}^{(t)}) = \tfrac{1}{2t}\|X(t) - x_0\|_\Lambda^2 + \tfrac{1}{2}\sum_{i=d-r+1}^{d} \lambda_{(i)}. \ \square \tag{5.13}$$

Corollary 5.4.

The likelihood ratio test statistic for testing H_r against H_{r+k} is given by

$$-2logQ^{(t)}_{r+k,r} = \sum_{i=d-r-k+1}^{d-r} \lambda_{(i)}. \ \square$$

Remarks.

It seems reasonable, that the condition $\lambda_{(d-r+1)} \neq \lambda_{(d-r)}$ should hold almost surely, but we don't have a proof at present.

The condition does not affect the existence of the maximum of the likelihood function, so the expression for the test statistics remains valid in all cases.

Besides this, one should note the interesting fact, that if $0 < r < d$, the estimator $(\widehat{A,B})^{(t)}$ now depends on Λ, c.f. Theorem (3.2).

Also, note that the expressions (5.11) – (5.13) for $r = d$ coincides with the corresponding expressions (3.6) – (3.7) found under the full model, and finally, that under H_r, $\text{rank}(\widehat{B}) = r$. $\square$

Proof of theorem 5.3.

The theorem holds trivially for $r = 0$. Therefore, we assume $r \geq 0$. From section 3 it follows, that for a given $t > 0$

$$\widehat{A}(B) = \tfrac{1}{t}(X(t) - x_0) - BE_t(X), \tag{5.15}$$

and the partially maximized likelihood is, cf. (3.4),

$$\begin{aligned} \ell_t(\widehat{A}(B), B) =& \tfrac{1}{2t}\|X(t) - x_0\|^2_\Lambda \\ &+ \int_0^t Y(s)^T B^T \Lambda^{-1}\, dX(s) - \tfrac{1}{2}\int_0^t \|BY(s)\|^2_\Lambda\, ds, \end{aligned} \tag{5.16}$$

with, as before,

$$Y(s) := X(s) - E_t(X).$$

We are going to maximize (5.16) subject to the condition (5.2). However,since B does not vary in a linear space, if $r < d$, we cannot use standard methods. Instead we do the following:

The rank of B is at most r iff the rows of B lie in some r-dimensional subspace N of $\mathbf{R}^d$, or equivalently, $B \in \mathbf{R}^d \otimes N$. Letting B vary under H_r is therefore exactly the same as letting B vary in $\mathbf{R}^d \otimes N$ with N fixed, and then afterwards letting N vary in the set of r-dimensional subspaces.

Therefore, let $N \subseteq \mathbf{R}^d$ be any r-dimensional subspace, and choose a base $(\beta_1, \ldots, \beta_r)$ for N, orthonormal w.r.t. $< \cdot, \cdot >_{V_t(X)^{-1}}$. Take $\beta_N \in \text{Mat}_{d\times r}(\mathbf{R})$ as the matrix $(\beta_1 : \cdots : \beta_r)$, and observe that $B \in \mathbf{R}^d \otimes N$ iff there exists $\alpha \in \text{Mat}_{d\times r}(\mathbf{R})$, such that

$$B = \alpha\beta_N^T. \tag{5.17}$$

Thus we have parametrised $\mathbf{R}^d \otimes N$ by $\mathrm{Mat}_{d\times r}(\mathbf{R})$, and we can maximize ℓ_t subject to $B \in \mathbf{R}^d \otimes N$ by using the parametrisation (5.17), letting α vary freely in $\mathrm{Mat}_{d\times r}(\mathbf{R})$. Regarded as a function of α, ℓ_t given by (5.16) is, by this choice of β_N, a second order polynomial of the form

$$\ell_t(\alpha) = \tfrac{1}{2t}\|X(t)\|^2_\Lambda + < \rho_N, \alpha >_\gamma - \tfrac{1}{2}\|\alpha\|^2_\gamma$$

where the vector $\rho_N \in \mathrm{Mat}_{d\times r}(\mathbf{R})$ is given by

$$\rho_N = (Y \bullet X^T)^T(t)\beta_N,$$

and with γ the matrix

$$\gamma = \Lambda \otimes I_r.$$

Consequently, ℓ_t attains its maximum on $\mathbf{R}^d \otimes N$ at the unique point

$$\widehat{B}_N = \rho_N \beta_N^T = (Y \bullet X^T)^T(t)\beta_N\beta_N^T, \tag{5.18}$$

and the maximal value is

$$\begin{aligned}\ell_t(\widehat{A}(\widehat{B}_N), \widehat{B}_N) =& \tfrac{1}{2t}\|X(t) - x_0\|^2_\Lambda + \tfrac{1}{2}\|\rho_N\|^2_\gamma \\ =& \tfrac{1}{2t}\|X(t) - x_0\|^2_\Lambda \\ &+ \tfrac{1}{2}\mathrm{trace}\big(\beta_N^T(Y \bullet X^T)(t)\Lambda^{-1}(Y \bullet X^T)^T(t)\beta_N\big).\end{aligned}$$

That the columns of β_N are orthonormal w.r.t. $< \cdot, \cdot >_{V_t(X)^{-1}}$ is the same as saying that the coloumns of $\Gamma = V_t(X)^{1/2}\beta_N$ are orthonormal in the usual sense. Hence, we must maximize the expression

$$\Gamma \mapsto \tfrac{1}{2}\mathrm{trace}(\Gamma^T M \Gamma), \tag{5.19}$$

subject to the condition, that the columns of Γ are orthonormal vectors.

But clearly, the maximal value of (5.19) is exactly $\frac{1}{2}\sum_{i=d-r+1}^{d}\lambda_{(i)}$, where $\lambda_{(1)} \le \cdots \le \lambda_{(d)}$ are the ordered eigenvalues for M, and the maximal value of ℓ_t is attained at a unique point $\widehat{B}$ iff the corresponding eigenspace $\mathrm{span}(\widehat{\Gamma})$ is well-defined, i.e. iff $\lambda_{(d-r+1)} \neq \lambda_{(d-r)}$. In that case the columns of $\widehat{\Gamma}$ must form an orthonormal base for the eigenspace determined by $\lambda_{(d-r+1)}, \ldots, \lambda_{(d)}$. The theorem now follows by inserting in (5.15) and (5.18), since $\widehat{\Gamma}\widehat{\Gamma}^T = P_r$. □

Finally, we shall deal with the autoregressive processes $X^{(n)}$ given by (2.3). Here, the inference depends crucially on the following unproven fact :

Assumption 5.5.

With M given by (5.10), the $(d-r+1)$'st largest eigenvalue $\lambda_{(d-r+1)}$ and the $(d-r)$'th differs, simultaneously for all t, with probability 1. □

Assumption 5.5. is necessary to ensure the consistency of the estimators in the model (2.3) under H_r, which we are about to define.

As in section 4, take

$$Y^{(n)} = X^{(n)} - E_{\frac{[nt]}{n}}(X^{(n)}),$$

and define M_n as the matrix

$$\begin{aligned} M_n = V_{\frac{[nt]}{n}}(X^{(n)})^{-\frac{1}{2}}(Y_-^{(n)} \bullet X^{(n)^T})(t) \\ \times \left(\tfrac{n}{[nt]} \sum_{i=1}^{[nt]} (X^{(n)}(\tfrac{i}{n}) - X^{(n)}(\tfrac{i-1}{n}))^{\otimes 2} \right. \\ \left. - \tfrac{1}{[nt]}(X^{(n)}(t) - x_0)^{\otimes 2} \right)^{-1} \\ \times (Y_-^{(n)} \bullet X^{(n)^T})^T(t) V_{\frac{[nt]}{n}}(X^{(n)})^{-\frac{1}{2}}. \end{aligned} \tag{5.20}$$

Define in the model (2.3), under H_r, the estimator

$$(A, \widehat{B}, \Lambda)^{(n,t)} = (\widehat{A}^{(n,t)}, \widehat{B}^{(n,t)}, \widehat{\Lambda}^{(n,t)})$$

by

$$\widehat{A}^{(n,t)} = \tfrac{n}{[nt]}(X^{(n)}(t) - x_0) - \widehat{B}^{(n,t)} E_{\frac{[nt]}{n}}(X^{(n)}), \tag{5.21}$$

$$\widehat{B}^{(n,t)} = (Y_-^{(n)} \bullet X^{(n)^T})^T(t) V_{\frac{[nt]}{n}}(X^{(n)})^{-\frac{1}{2}} P_r^{(n)} V_{\frac{[nt]}{n}}(X^{(n)})^{-\frac{1}{2}} \tag{5.22}$$

$$\begin{aligned} \widehat{\Lambda}^{(n,t)} = \tfrac{n}{[nt]} \sum_{i=1}^{[nt]} (X^{(n)}(\tfrac{i}{n}) - X^{(n)}(\tfrac{i-1}{n}))^{\otimes 2} \\ - \tfrac{1}{[nt]}(X^{(n)}(t) - x_0)^{\otimes 2} \\ - \tfrac{1}{n}\widehat{B}^{(n,t)}(Y_-^{(n)} \bullet X^{(n)^T})(t), \end{aligned} \tag{5.23}$$

where $P_r^{(n)}$ is the matrix representing the (usual orthogonal-) projection on the eigenspace corresponding to the r largest eigenvalues for M_n (if $r = 0$, we take $B = 0$, of course).

Under Assumption 5.5 we see, using Lemma 4.1, that $M_n \xrightarrow{\mathcal{D}} M$, so that (5.21) – (5.23) are well defined with a probability tending to 1.

If the errors in the model (2.3) are Gaussian, it follows from the analysis in Johansen (1991), that the estimators given by (5.21) – (5.23) are the ML-estimators for A, B and Λ under H_r. But $(A, \widehat{B}, \Lambda)^{(n)}$ is defined for arbitrary errors with the second-order moment representation (2.2).

The reason for using these estimators in general is, apart from the coincidence with the Gaussian error-estimators, the following

Theorem 5.6.

Under H_r, if Assumption 5.5 is valid, the convergence

$$(\widehat{A,B},\Lambda)^{(n)} \xrightarrow{\mathcal{D}} (\widehat{A,B},\Lambda) \tag{5.24}$$

holds in the Skorokhod topology on $D_{\mathbf{R}^d\times \mathrm{Mat}_d(\mathbf{R})\times \mathbf{H}_d^+(\mathbf{R})}(0;\infty)$, where $(\widehat{A,B}) = ((\widehat{A}^{(t)},\widehat{B}^{(t)}))_{t\geq 0}$ given by (5.11) – (5.12) is the diffusion estimate process for (A,B) under H_r, and Λ is regarded as the constant process $t\mapsto\Lambda$. □

Proof.

If $C\in \mathrm{Mat}_d(\mathbf{R})$ is some positive semidefinite matrix, with the r'th largest eigenvalue different from the $(r+1)$'st, and C_n is positive semidefinite, $n\in\mathbf{N}$, such that $C_n\to C$, then the eigenvalues of C_n converge to those of C, and if $P_r^{(n)}, P_r\in \mathrm{Mat}_d(\mathbf{R})$ are matrices representing the projections on the eigenspace corresponding to the r largest eigenvalues of C_n, resp. C, then $P_r^{(n)}\to P_r$.

From this argument it follows, that the process $(\widehat{A,B})^{(n)}$ is a function of $\big(\frac{[nt]}{n},(Y_-^{(n)}\bullet X^{(n)^T}),X^{(n)}\big)$, that is a.s. continuous in the limit $\big(t,(Y\bullet X^T),X\big)$. Since (A,B) is the same function of $\big(t,(Y\bullet X^T),X\big)$, the convergence

$$(\widehat{A,B})^{(n)} \xrightarrow{\mathcal{D}} (A,B)$$

follows, because

$$\big((Y_-^{(n)}\bullet X^{(n)^T}),X^{(n)}\big) \xrightarrow{\mathcal{D}} \big((Y\bullet X^T),X\big).$$

Hence (5.24) follows, since

$$\tfrac{n}{[nt]}\sum_{i=1}^{[nt]}(X^{(n)}(\tfrac{i}{n})-X^{(n)}(\tfrac{i-1}{n}))^{\otimes 2}$$

is earlier shown to converge to Λ in probability, uniformly on compacts. □

Remark.

We use Assumption 5.5 to deduce that the above mentioned function is a.s. continuous in $\big((Y\bullet X^T),X\big)$ with respect to the Skorokhod topology on $D_{\mathrm{Mat}_d(\mathbf{R})\times\mathbf{R}^d}(0;\infty)$. □

The convergence result from Theorem 5.6 makes it natural to propose the following test statistics for testing H_r against H_{r+k} in the model (2.3) :

Define $Q_{r+k,r}^{(n,t)}$ by

$$-2\log Q_{r+k,r}^{(n,t)} := \sum_{i=d-r-k+1}^{d-r}\lambda_{(i)}^{(n)},$$

where $\lambda_{(1)}^{(n)}\leq\cdots\leq\lambda_{(d)}^{(n)}$ are the (ordered) eigenvalues of the matrix M_n given by (5.20). With Gaussian errors, $Q_{r+k,r}^{(n,t)}$ coincides with the likelihood ratio test statistic for testing H_r against H_{r+k} (see Johansen (1991)), and with Theorem 5.6 in mind the following is evident :

Theorem 5.7.

If assumption 5.5 is valid, then under H_r the convergence of the processes

$$Q^{(n)}_{r+k,r} \xrightarrow{\mathcal{D}} Q_{r+k,r}$$

holds, simultaneously with the convergence (5.24). □

References.

Cox, D. D. (1991). Gaussian Likelihood Estimation for Nearly Nonstationary AR(1)-processes. *Annals of Statistics* **19**, 1129-1142.

Engle, R. F. & Granger, C. W. J. (1987). Co-integration and Error Correction: Representation, Estimation and Testing. *Econometrica* **55**, 251-276.

Ethier, S. N. , & Kurtz, T. G. (1986). *Markov Processes: Characterisation and Convergence.* Wiley, New York.

Jacobsen, M. (1991). Homogeneous Gaussian Diffusions in Finite Dimensions. Preprint 3. Institute of Mathematical Statistics, University of Copenhagen.

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica* **59**, 1551-1580.

Kurtz, T. G. & Protter, P. (1991). Weak Limit Theorems for Stochastic Integrals and Stochastic Differential Equations. *Annals of Probability* **19**, 1035-1070.

Le Breton, A. (1977). Parameter Estimation in a Linear Stochastic Differential Equation. *Transactions of the 7'th Prague Conference and of the European Meeting of Statisticians 1974* **A**, 353-366. Academia, Prague.

Le Breton, A. & Musiela, M. (1985). Some Parameter Estimation Problems for Hypoelliptic Homogeneous Gaussian Diffusions. *Sequential Methods in Statistics, Banach Center Publications* **19**, 337-356. PWN - Polish Scientific Publishers, Warzaw.

Liptser, A. S. & Shiryaev, A. N. (1977). *Statistics of Random Processes I, General Theory.* Springer, New York.

Rogers, L. C. G. & Williams, D. (1987). *Diffusions, Markov Processes and Martingales. Volume 2: Itô Calculus.* Wiley, Chichester.

A. Stockmarr, Institute of Mathematical Statistics, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark.

PREPRINTS 1992

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK. TELEPHONE + 45 35 32 08 99.

No. 1 Johansen, Søren: The Role of the Constant Term in Cointegration Analysis of Nonstationary Variables.

No. 2 Paruolo, Paolo: Asymptotic Inference on the Moving Average Impact Matrix in Cointegrated I(1) VAR Systems.

No. 3 Johansen, Søren and Juselius, Katarina: Identification of the Long-Run and the Short-Run Structure. An Application to the ISLM Model.

No. 4 Johansen, Søren: Identifying Restrictions of Linear Equations.

PREPRINTS 1993

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK. TELEPHONE + 45 35 32 08 99.

No. 1 Hansen, Henrik and Johansen, Søren: Recursive Estimation in Cointegration VAR-Models.

No. 2 Stockmarr, A. and Jacobsen, M.: Gaussian Diffusions and Autoregressive Processes: Weak Convergence and Statistical Inference.