LUND UNIVERSITY

On Consistency for the Method of Least Squares Using Martingale Theory

Sternby, Jan

1976

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Sternby, J. (1976). *On Consistency for the Method of Least Squares Using Martingale Theory*. (Technical Reports TFRT-7104). Department of Automatic Control, Lund Institute of Technology (LTH).

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

TFRT-7104

ON CONSISTENCY FOR THE METHOD OF LEAST SQUARES USING MARTINGALE THEORY

J. STERNBY

Report 7634 (C) July 1976
Department of Automatic Control
Lund Institute of Technology

TILLHÖR REFERENSBIBLIOTEKET
UTLANAS EJ

ON CONSISTENCY FOR THE METHOD OF LEAST SQUARES USING MARTINGALE THEORY

Jan Sternby

Department of Automatic Control
Lund Institute of Technology
Box 725
S-220 07 Lund 7, Sweden

ABSTRACT

Least Squares Identification is considered from the Bayesian point of view. Necessary and sufficient condition for consistency almost everywhere is given under the assumption that the data is generated by a regression model with white and Gaussian noise.

This work was supported by the Swedish Board for Technical Development under contract No 74-3476.

1. INTRODUCTION

The method of Least Squares (LS) has been treated by many authors, starting with Gauss. Mann and Wald (1943) were the first ones to apply it to time-series modelling, and also to prove its consistency for this case. Åström (1968) extended the consistency result to systems with an input, and Ljung (1976) has shown convergence and consistency under very mild conditions that e.g. include general feedback situations.

In the present paper a Bayesian approach to the identification procedure is used which gives a different type of result. The true system parameters are thus regarded random variables and not constants, which is usually the case.

The main result is necessary and sufficient condition for consistency a.e. for the LS-method in the Gaussian white noise case under a week condition. This is proved in a series of theorems, where the basic idea is that the LS-estimate is in the Gaussian case a conditional mean, and therefore converges a.e. according to martingale theory.

2. THE LEAST SQUARES IDENTIFICATION METHOD. NOTATIONS

Let a system with p outputs $y(\cdot)$ and r inputs $u(\cdot)$ be governed by the vector difference equation

$$y(t) + A_1 y(t-1) + \ldots + A_n y(t-n) = B_1 u(t-1) + \ldots + B_m u(t-m) + v(t) \tag{1}$$

where $v(t)$ is some (vector) disturbance. Introduce

$$\varphi(t)^T = [-y(t-1)^T \ldots -y(t-n)^T \; u(t-1)^T \ldots u(t-m)^T]$$

and

$$\theta^T = [\; A_1 \ldots A_n \; B_1 \ldots B_m \;]$$

To get a formal similarity with the filtering problem it is convenient to represent the unknown parameters as a vector. Therefore, introduce

$$x = \text{col}\ \theta$$

which is obtained by writing the columns of θ under each other. This is of course not crucial, but it simplifies comparison with the well-known Kalman filter.

Then with

$$\Phi(t)^T = \begin{pmatrix} \varphi(t)^T & & 0 \\ & \ddots & \\ 0 & & \varphi(t)^T \end{pmatrix}$$

(1) can be written

$$y(t) = \Phi(t)^T x + v(t) \tag{2}$$

The weighted LS-estimate of x at time t, $\hat{\zeta}_t$, is obtained by minimizing

$$V_t(\zeta) = E\left(\frac{1}{t-t_0} \sum_{s=t_0+1}^{t} \| y(s) - \Phi(s)^T \cdot \zeta \|_W^2 \right) \tag{3}$$

with respect to ζ.

Notice that for every diagonal W the estimates will be the same as if W is the identity matrix.

In treating consistency, the true system parameters are usually considered as constants, given once for all (but not known). Then each point ω in the sample space gives a certain realization of the noise sequence $\{v(t)\}$ (and of the input sequence $\{u(t)\}$ in case of random input), whereas the true system parameters are the same for every ω. The concept "almost everywhere", a.e., then means "for almost every realization" for the particular system given. Ljung (1976) has given very general conditions for the LS-estimates to be consistent a.e. in the above meaning.

In the present paper, however, the true system parameters are considered as random variables. The choice of system is regarded as part of the experiment, and each realization starts by picking a system. Then, of course, the true system will probably not be the same in two different realizations. Each point ω in the sample space will thus give 1) the true system parameters and 2) a noise sequence. The sample space may be regarded as a product space, so that $\omega = (\omega_1 \quad \omega_2)$, where ω_1 determines the true system and ω_2 determines the noise sequence. In order to get consistency, the sets in the σ-algebra generated by all the measurements should tend to be parallel to the ω_2-axes as time tends to infinity, i.e. the variations caused by the noise should be averaged out.

With this point of view, the concept a.e. means "for almost every realization for almost every system". This must be remembered when comparing the results of this paper with other results.

The following additional notations will be used:

$\mathcal{F}_t$ - the σ-algebra generated by all measurements of y and u up to and including time t

$\mathcal{F}_\infty$ - the smallest σ-algebra containing $\mathcal{F}_t$ for every t

$\hat{x}_t = \mathcal{E}(x|\mathcal{F}_t)$ - the conditional expectation of x given $\mathcal{F}_t$

$P_t = \mathcal{E}\left((x-\hat{x}_t)^2|\mathcal{F}_t\right)$ - the conditional covariance of x given $\mathcal{F}_t$

$1_B(\omega)$ - the indicator function for the set B $\left(1_B(\omega) = 1 \text{ if } \omega \in B \text{ otherwise } 1_B(\omega) = 0\right)$

$\mathcal{P}(B|\mathcal{F}_t)$ - the conditional probability for B given $\mathcal{F}_t$

3. GENERAL RESULTS

It is well-known (see e.g. Kalman (1960) or Jazwinski (1970)) that under very general circumstances the conditional mean is also the LS-estimate. This fact makes the following theorem interesting.

Theorem 1: Suppose that the distribution of the true parameters x has finite second moments. Then $\hat{x}_t$ and P_t converge a.e. The limits are denoted by $\hat{x}_\infty$ and P_∞.

Proof: According to theorem 9.4.5 in Chung (1968) the conditional mean of an integrable variable is a martingale that converges a.e. Now x has finite second moments and each component of the vector $\hat{x}_t$ is a conditional mean. Moreover

$$(P_t)_{ij} = E\left((x-\hat{x}_t)_i \cdot (x-\hat{x}_t)_j \,|\, F_t\right) = E(x_i x_j \,|\, F_t) - (\hat{x}_t)_i \cdot (\hat{x}_t)_j$$

where the first term is a conditional mean and the second one has already been shown to converge. □

In fact, Chung (1968) also shows that the limit $\hat{x}_\infty = E(x \,|\, F_\infty)$ a.e. In the next theorem this limit is examined.

Theorem 2: With the assumptions of theorem 1, if M is the set $\{\omega \,|\, P_\infty = 0\}$ then $1_M \cdot \hat{x}_\infty = 1_M \cdot x$ a.e.
If $P(M) = 1$ then also $\hat{x}_t \to x$ in L^2.

Proof: It is sufficient to consider the scalar case, since $P_t \to 0$ implies that all its diagonal elements tend to zero. Since $M \in F_\infty$, $E(1_M \,|\, F_t) \to 1_M$ a.e. according to Lévy's zero-or-one law. Then

$$E(1_M \,|\, F_t)^2 \cdot P_t \to 0 \quad \text{a.e.}$$

But

$$0 \leq E(1_M \,|\, F_t)^2 \cdot P_t = E(1_M \,|\, F_t)^2 \cdot E(x^2 \,|\, F_t) - E(1_M \,|\, F_t)^2 \cdot \hat{x}_t^{\,2}$$

Now both terms of the right member are uniformly integrable since they are less than $E(x^2|F_t)$ (a.e.), which is a conditional mean and thus uniformly integrable by martingale theory [Chung (1968), theorem 9.4.3]. Also both terms converge a.e. and so they must converge in L^1 [Chung (1968), theorem 4.5.4]. Then the left member converges in L^1 and a.e., and the limits must be equal, i.e. zero. This means that

$$E\left\{E\left(1_M\middle|F_t\right)^2\cdot E\left((\hat{x}_t-x)^2\middle|F_t\right)\right\} = E\left\{E\left(E(1_M|F_t)^2(\hat{x}_t-x)^2\middle|F_t\right)\right\} =$$

$$= E\left(E(1_M|F_t)(\hat{x}_t-x)\right)^2 \to 0$$

so that

$$E(1_M|F_t)\cdot(\hat{x}_t-x) \to 0 \quad \text{in } L^2$$

and the last part of the theorem is proven. But Lévy's zero-or-one law and theorem 1 together imply that

$$E(1_M|F_t)\cdot(\hat{x}_t-x)$$

converges a.e. The limit must be zero because it is in L^2. Then also

$$1_M\cdot(\hat{x}_t-x) \to 0 \quad \text{a.e.}$$

which proves the theorem. □

<u>Remark</u>: From the proof it is evident that the theorem can be applied component-wise.

These two theorems might also be used in connection with other identification schemes than the LS-method. Then it must be shown that the difference between the estimate and the conditional mean tends to zero. The conditional mean is unfortunately difficult to calculate in general. For the Gaussian case, however, it is equal to the <u>linear</u> LS-estimate, which is given by the Kalman filter equations.

4. MAIN RESULTS. THE GAUSSIAN CASE

The following theorem, given in Åström, Wittenmark (1971), couples the Kalman filter equations to the conditional mean. The weighted LS-estimate also satisfies these equations. Thus the theorem makes theorems 1 and 2 applicable to LS-estimation in the Gaussian case.

Theorem 3: (Åström, Wittenmark)
Suppose that the true parameter vector x is Gaussian with *a priori* mean m and *a priori* covariance P_0, $\{v(t)\}$ is a sequence of independent, equally distributed normal vectors with zero mean value and positively definite covariance R, and x and v(t) are independent for all t. Let the output vector of the system be generated by (1). Then the conditional distribution of x given F_t is normal with mean $\hat{x}_t$ and covariance P_t, where $\hat{x}_t$ and P_t satisfy the difference equations

$$\hat{x}_t = \hat{x}_{t-1} + K(t)[y(t)-\Phi(t)^T\hat{x}_{t-1}] \tag{4}$$

$$P_t = P_{t-1} - P_{t-1}\Phi(t)[R+\Phi(t)^T P_{t-1}\Phi(t)]^{-1}\Phi(t)^T P_{t-1} \tag{5}$$

where

$$K(t) = P_{t-1}\Phi(t)[R+\Phi(t)^T P_{t-1}\Phi(t)]^{-1} = P_t\Phi(t)R^{-1} \tag{6}$$

and the initial conditions are $\hat{x}_{t_0} = m$, $P_{t_0} = P_0$.

Proof: For the single-input single-output case the proof is indicated in Åström, Wittenmark (1971). The extension to the multivariable case is straightforward.

It is well-known (see e.g. Åström (1968)) that the weighted LS-estimate is also given by equations (4)-(6) with the weighting matrix $W = R^{-1}$. Thus $\hat{x}_t$ minimizes $V_t(\cdot)$, so that $\hat{x}_t = \hat{\zeta}_t$ = the weighted LS-estimate.

Moreover, if R is diagonal then $\hat{x}_t$ is also the ordinary LS-estimate, i.e. it minimizes $V_t(\cdot)$ for W = I. Since R

must be known it can also be made diagonal by a transformation of variables, and so it is no restriction to assume R diagonal.

Corollary 1: Under the assumptions of the theorem and R diagonal it follows from theorem 2 that the estimate $\hat{\zeta}_t = \hat{x}_t$ is consistent a.e. and in L^2 provided $P_t \to 0$ a.e.

Corollary 2: Under the assumptions of the theorem

$$\sum_{s=t_0+1}^{\infty} K(s)K(s)^T < \infty \quad \text{a.e.}$$

which gives a lower bound to the convergence rate of $K(t)$.

Proof: By theorem 1 P_t converges a.e. and

$$P_t = P_0 - \sum_{s=t_0+1}^{t} K(s)[R + \Phi(s)^T P_s \Phi(s)] K(s)^T$$

Now a condition is needed to guarantee that $P_t \to 0$. This is given in the next theorem.

Theorem 4: With notations and assumptions as in theorem 3 and R diagonal

$$\{\omega | P_t \to 0\} =$$

$$= \{\omega | \sum_{s=t_0+1}^{\infty} [a^T\varphi(s)]^2 \text{ divergent for every constant column vector } a \neq 0.$$

To prove this the following lemma is needed.

Lemma: Let $\{P_t\}$ be a sequence of positively definite matrices such that $P_t \to P_\infty$ and $P_t - P_\infty$ positively semidefinite for all t. Then

$$P_\infty = 0 \Leftrightarrow a^T P_t^{-1} a \to \infty \quad t \to \infty \text{ for every constant column vector } a \neq 0.$$

Proof: The proof is given in Appendix.

Proof of theorem 4: Theorem 1 gives $P_t \to P_\infty$ a.e. for some $P_\infty \geq 0$. The formula in theorem 3 for computing P_{t+1} shows that $P_{t+1} - P_t \leq 0$ and so $P_t - P_\infty \geq 0$ for all t. Then the lemma gives

$P_t \to 0 \Leftrightarrow \tilde{a}^T P_t^{-1} \tilde{a} \to \infty \quad t \to \infty$ for every constant vector $\tilde{a} \neq 0$.

Now equation (5) for P_t is equivalent to

$$P_t^{-1} = P_0^{-1} + \sum_{s=t_0+1}^{t} \Phi(s)\ R^{-1}\ \Phi(s)^T$$

so that

$$\tilde{a}^T (P_t^{-1} - P_0^{-1}) \tilde{a} = \sum_{s=t_0+1}^{t} \tilde{a}^T \Phi(s)\ R^{-1}\ \Phi(s)^T\ \tilde{a} =$$

$$= \sum_{s=t_0+1}^{t} [\tilde{a}_1^T \ldots \tilde{a}_p^T] \begin{pmatrix} \varphi(s) & & 0 \\ & \ddots & \\ 0 & & \varphi(s) \end{pmatrix} \begin{pmatrix} \frac{1}{r_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{r_p} \end{pmatrix} \begin{pmatrix} \varphi(s)^T & & 0 \\ & \ddots & \\ 0 & & \varphi(s)^T \end{pmatrix} \begin{pmatrix} \tilde{a}_1 \\ \vdots \\ \tilde{a}_p \end{pmatrix} =$$

$$= \sum_{j=1}^{p} \sum_{s=t_0+1}^{t} \frac{1}{r_j} \left(\tilde{a}_j^T \varphi(s)\ \varphi(s)^T\ \tilde{a}_j\right)$$

Thus

$$\tilde{a}^T P_t^{-1} \tilde{a} \to \infty \text{ for every constant } \tilde{a} \neq 0$$

$$\Longleftrightarrow$$

$$\sum_{s=t_0+1}^{t} [a^T \varphi(s)]^2 \text{ divergent for every constant } a \neq 0$$

since $r_j > 0$ for all j, $1 \leq j \leq p$.

This completes the proof.

Theorem 4 shows that in the Gaussian case (with $v(\cdot)$ being white noise) the only condition needed for consistency a.e. is that

$$\sum_{s=t_0+1}^{\infty} [a^T \varphi(s)]^2$$

be divergent a.e. for every constant vector $a \neq 0$. This condition will be referred to as CC (Consistency Condition).

Now for simplicity consider single-input single-output systems. Then in the open-loop case CC is a condition on the input only, because to avoid divergence all components of the vector a corresponding to y-components in $\varphi(t)$ must be zero, since y contains also a white noise part. This fact is further discussed in Ljung, Wittenmark (1974). To see the relation between CC and the concept of persistently exciting consider the case $\varphi(t) = u(t-1)$. Then u is persistently exciting of order one only if

$$\lim_{N\to\infty} \frac{1}{N} \sum^{N} u(t)^2 > 0$$

whereas CC only demands that

$$\sum^{\infty} u(t)^2 \text{ diverges}$$

so that u(t) may e.g. decrease to zero with increasing t.

In the closed-loop case CC gives a condition on the feedback. If it is linear and constant, it must be of such a high order that not all of its terms are components in the vector $\varphi(t)$. If it is time-varying it must not converge too fast to a linear and constant feedback of low order.

Example: (from Ljung (1974))
Consider the system

$$y(t+1) + x_1 y(t) = x_2 u(t) + e(t+1)$$

with the time-varying feedback

$$u(t) = f(t)y(t)$$

where $f(t) \to f$ as $t \to \infty$. Then with $a^T = [\, a_1 \quad a_2 \,]$ CC is

$$\sum^{\infty} [a_1 + a_2 f(t)]^2 y(t)^2$$

diverges for every $a \neq 0$. Now $y(t) \not\to 0$ because of the noise, so there must be a subsequence for which $\{y(t)^2\}$ is bounded from below. Thus CC is satisfied if

$$\sum^{\infty} [f(t) - f]^2$$

diverges.

For the case when the minimization of $V_t(\cdot)$ is restricted to

a finite set of parameter values this result and the consistency condition (CC) was shown in Ljung (1974), cf also Ljung (1976).

Finally the question of non-consistency will be treated. In order to conclude non-consistency a.e. it is not sufficient that P_∞, the *a posteriori* covariance after all the measurements, is nonzero. But if the *a posteriori* distribution is continuous and x is a constant, then with $P_\infty > 0$ the probability will be zero for the estimate to take any particular value, especially the true one. The fact that x is a stochastic variable, is however a complication. The next theorem treats non-consistency in the Gaussian case and couples it to P_∞ being non-zero.

<u>Theorem 5</u>: With the assumptions of theorem three

$$P(P_\infty \neq 0, \quad \hat{x}_\infty = x) = 0$$

<u>Proof</u>: As in theorem 2 it is no restriction to consider the scalar case only. According to theorem 3 the conditional distribution of x given F_t is normal with mean $\hat{x}_t$ and covariance P_t. Introduce the sets $M_\gamma = \{\omega | P_\infty < \gamma\}$ and $M = \{\omega | P_\infty = 0\}$. Then

$$P(|x-\hat{x}_t| < \varepsilon | F_t) = \frac{1}{\sqrt{2\pi P_t}} \int_{-\varepsilon}^{\varepsilon} e^{-s^2/2P_t} ds \leq$$

$$\leq \begin{cases} 1 & \text{if} \quad \omega \in M_\gamma \\ \dfrac{1}{\sqrt{2\pi P_\infty}} \cdot 2\varepsilon \leq \dfrac{1}{\sqrt{2\pi\gamma}} \cdot 2\varepsilon = k(\gamma)\cdot\varepsilon & \text{if} \quad \omega \notin M_\gamma \end{cases}$$

Taking expectations on both sides gives

$$P(|x-\hat{x}_t| < \varepsilon) \leq k(\gamma)\cdot\varepsilon + P(M_\gamma)$$

for all $t > t_0$. Now $P(M_\gamma)$ can be made arbitrarily close to $P(M)$ by choosing γ small enough.

But $\hat{x}_t \to \hat{x}_\infty$ a.e. and so by Egorov's theorem (see e.g. Halmos (1950)) for any $\delta > 0$ there exists a set N with $P(N) > 1 - \delta$ such that $\hat{x}_t \to \hat{x}_\infty$ uniformly on N. Then there is a $T(\varepsilon)$ so that

$$\sup_{\omega \in N} \left| \hat{x}_t - \hat{x}_\infty \right| < \varepsilon$$

for all $t > T(\varepsilon)$.

This gives

$$P(\hat{x}_\infty = x) \leq P(|\hat{x}_\infty - x| < \varepsilon) \leq P(\{|\hat{x}_\infty - x| < \varepsilon\} \cap N) + \delta \leq$$

$$\leq P(|\hat{x}_t - x| < 2\varepsilon) + \delta$$

if $t > T(\varepsilon)$. Now the right member can be made less than $P(M) + 3\delta$ for any $\delta > 0$ by choosing first δ, then γ to make $P(M_\gamma) < P(M) + \delta$ then $\varepsilon\,(0 < \varepsilon < \delta/2k(\gamma))$ and finally $t > T(\varepsilon)$. Thus

$$P(\hat{x}_\infty = x) \leq P(M) = P(P_\infty = 0)$$

Then

$$P(\hat{x}_\infty = x,\ P_\infty \neq 0) = P(\hat{x}_\infty = x) - P(\hat{x}_\infty = x,\ P_\infty = 0) =$$

$$= P(\hat{x}_\infty = x) - P(P_\infty = 0) \leq 0$$

where the last equality is implied by theorem two. This completes the proof.

<u>Remark</u>: Theorems 2 and 5 together show that the sets $\{\omega | P_\infty = 0\}$ and $\{\omega | \hat{x}_\infty = x\}$ can differ only by a null-set.

Theorems 4 and 5 should be combined to show different cases of non-consistency. A constant and linear feedback of sufficiently low order is of course one case, since then $a^T \varphi(t) \equiv 0$ everywhere for some $a \neq 0$, so that CC is satisfied nowhere.

The only difficult cases are when the feedback converges too fast to a linear and constant one. Then the exact limit in convergence rate separating consistency from non-consistency will depend on the stability of the limiting closed-loop system.

Example (continued): Consider again the first-order example given above. If the closed-loop system is stable and

$$\sum^{\infty} [f(t)-f]^2$$

converges then $P_t \not\to 0$. But if the closed-loop system is unstable then $f(t)$ must converge faster in order to make

$$\sum^{\infty} [a_1 + a_2 f(t)]^2 \; y(t)^2$$

convergent, and the required convergence rate depends on how unstable the closed-loop system is, which in turn depends on f, x_1 and x_2.

5. CONCLUSIONS

The two main ideas and results of this paper are 1) the way of looking at the true system as taken from a set of systems at the beginning of each realization and 2) the coupling of consistency and non-consistency for the LS-method to the divergence or convergence of a certain series (CC, the Consistency Condition). This condition is shown to be sufficient *and* necessary in the Gaussian white noise case. It may be interpreted as a condition that the input should "shake" the system long enough, in the open-loop as well as in the closed-loop case.

It is interesting to note that the theorems do not require any conditions on the stability of the systems, as do most results previously given. However, in showing consistency using CC, unstable systems seem to require a "less exciting" input than do stable systems.

As for extensions, the case with time-varying noise covariance could be treated. This would effect only theorem four and CC would include the noise covariance. Theorems one and two are given in a general form, but their possible application to other cases has not been investigated. However, theorem four may be used for any method containing a P-equation as in the LS-case.

ACKNOWLEDGEMENT

The author is glad to express his gratitude to professor L Ljung for many fruitful discussions and good suggestions.

6. REFERENCES

Åström, K J (1968):
Lectures on the Identification Problem - The Least Squares Method. Report 6806, Dept of Automatic Control, Lund Institute of Technology.

Åström, K J, and Wittenmark, B (1971):
Problems of Identification and Control.
J of Mathematical Analysis and Applications, Vol 34, pp 90-113.

Chung, K L (1968):
A Course in Probability Theory. Harcourt, Brace & World Inc.

Halmos, P R (1950):
Measure Theory. D van Nostrand Company Inc.

Jazwinski, A H (1970):
Stochastic Processes and Filtering Theory.
Academic Press.

Ljung, L (1974):
On Consistency for Prediction Error Identification Methods. Report 7405, Dept of Automatic Control, Lund Institute of Technology.

Ljung, L (1976):
Consistency of the Least Squares Identification Method.
To appear in IEEE Trans on Automatic Control, October 1976.

Ljung, L, and Wittenmark, B (1974):
Asymptotic Properties of Self-Tuning Regulators.
Report 7404, Dept of Automatic Control, Lund Institute of Technology.

Kalman, R E (1960):
A New Approach to Linear Filtering and Prediction Problems. ASME J Basic Engineering, Vol 82, pp 35-45.

Mann, H B, and Wald, A (1943):
On the Statistical Treatment of Linear Stochastic Difference Equations. Econometrica, Vol 11, pp 173-220.

APPENDIX

Lemma: Let $\{P_t\}$ be a sequence of positively definite matrices such that $P_t \to P_\infty$ and $P_t - P_\infty$ positively semidefinite for all t. Then

$$P_\infty = 0 \Leftrightarrow a^T P_t^{-1} a \to \infty \qquad t \to \infty$$

for every constant column vector $a \neq 0$.

Proof of the lemma: First suppose that $P_\infty = 0$. Let λ_t be the smallest eigenvalue of P_t^{-1}. Then

$$a^T P_t^{-1} a \geq \lambda_t a^T a \to \infty \qquad t \to \infty$$

for every constant $a \neq 0$, since all eigenvalues of P_t tend to zero.

Next suppose $P_\infty \neq 0$, and assume that it is diagonal. This is no restriction since it is symmetric and thus can be diagonalized. At least one of the elements of P_∞ must be non-zero, say the (1,1)-element. Then put

$$P_\infty = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \ddots \end{pmatrix} \quad \text{and} \quad \overline{P} = \begin{pmatrix} \lambda_1 & & & 0 \\ & 0 & & \\ & & \ddots & \\ 0 & & & 0 \end{pmatrix}$$

so that $P_\infty \geq \overline{P}$. Also introduce $A_t = P_t - \overline{P}$ and $\tilde{A}_t$ with

$$A_t = \left(\begin{array}{c|cc} a_{11}^t & a_{12}^t & a_{13}^t \cdots\cdots \\ \hline a_{21}^t & a_{22}^t & a_{23}^t \cdots\cdots \\ a_{31}^t & a_{32}^t & a_{33}^t \\ \vdots & \vdots & \ddots \end{array}\right) = \left(\begin{array}{c|c} a_{11}^t & a_{12}^t \cdots\cdots \\ \hline a_{21}^t & \\ \vdots & \tilde{A}_t \\ \vdots & \end{array}\right)$$

Then $A_t \geq 0$.

Now

$$\det P_t = \left[\begin{array}{c|c} \lambda_1 + a_{11}^t & a_{12}^t \cdots\cdots \\ \hline a_{21}^t & \\ \vdots & \tilde{A}_t \\ \vdots & \end{array}\right] =$$

$$= \left[\begin{array}{c|c} \lambda_1 & a_{12}^t \cdots\cdots \\ \hline 0 & \\ \vdots & \tilde{A}_t \\ 0 & \end{array}\right] + \left[\begin{array}{c|c} a_{11}^t & a_{12}^t \cdots\cdots \\ \hline a_{21}^t & \\ \vdots & \tilde{A}_t \\ \vdots & \end{array}\right] =$$

$$= \lambda_1 \det \tilde{A}_t + \det A_t \geq \lambda_1 \det \tilde{A}_t$$

so that for the (1,1)-element of P_t^{-1}

$$(P_t^{-1})_{11} = \frac{\det \tilde{A}_t}{\det P_t} \leq \frac{\det \tilde{A}_t}{\lambda_1 \det \tilde{A}_t} = \frac{1}{\lambda_1}$$

Thus with $a^T = [\ 1 \quad 0 \quad 0 \ \ldots \ 0\ ]$

$$a^T P_t^{-1} a \leq \frac{1}{\lambda_1} a^T a \not\to \infty$$

as $t \to \infty$.