Non-asymptotic confidence ellipsoids for the least squares estimate

Erik Weyer
Department of Electrical and Electronic Engineering
University of Melbourne, Parkville VIC 3010, Australia
e.weyer@ee.mu.oz.au

M.C. Campi
Department of Electrical Engineering and Automation
University of Brescia, Via Branze 38, 25123 Brescia, Italy
campi@bsing.ing.unibs.it

Abstract

In this paper we consider the finite sample properties of least squares system identification, and we derive *non-asymptotic* confidence ellipsoids for the estimate. Unlike asymptotic theory, the obtained confidence ellipsoids are valid for a finite number of data points. The probability that the estimate belongs to a certain ellipsoid has a natural dependence on the volume of the ellipsoid, the data generating mechanism, the model order and the number of data points available.

Keywords: System identification, least squares, confidence ellipsoids, finite sample properties

1 Introduction

In this paper we consider the properties of least squares system identification when only a finite number of data points are available. The asymptotic properties of least squares identification are well understood, see e.g. Ljung (1999) or Söderström and Stoica (1989), but it is only recently that results addressing the finite sample properties have started appearing, e.g. Weyer et al (1999), Weyer and Campi (1999), Weyer (2000) and Campi and Weyer (2000).

In applications it is common to use the asymptotic confidence regions for the parameter estimate, even only a finite number of data points are available. In this paper we derive non-asymptotic confidence ellipsoids for the least squares estimate. It is shown that the confidence ellipsoids depend in a natural way on factors such as the model and system order, the pole locations and the number of data points available.

The main tool we make use of in order to derive the confidence ellipsoids is exponential inequalities. Earlier, using different techniques, Spall (1995) has considered uncertainty bounds for general M-estimators for a finite number of data points. His results are however difficult to use in the situation we consider here.

The paper is organised as follows. In the next section we introduce the identification setting. The main result is given in section 3 while technical results are given in the appendices. Due to space limitations only partial proofs are given.

2 Identification setting

2.1 The data generation mechanism

We assume that the observed data are generated by a linear system

$$y(t) = G_0(q^{-1})u(t) + H_0(q^{-1})e(t) \qquad (1)$$

where the input signal $u(t)$ is stochastic and generated by

$$u(t) = V_0(q^{-1})w(t) \qquad (2)$$

$w(t)$ is a sequence of independent Gaussian random variables with zero mean and variance σ_w^2. The noise process $e(t)$ is a sequence of independent Gaussian random variables with zero mean and variance σ_e^2. The assumptions on $u(t)$ and $e(t)$ are not crucial. The results can easily be extended to deterministic input signals and other types of iid noise sequences. $G_0(q^{-1})$, $H_0(q^{-1})$ and $V_0(q^{-1})$ are transfer functions in the backward shift operator q^{-1}, i.e $q^{-1}y(t) = y(t-1)$; however, for the sake of readability, we omit throughout to explicitly indicate the dependence on q^{-1}. Moreover, G_0, H_0 and V_0 can be written as

$$G_0 = \frac{B_0}{A_0}, \quad H_0 = \frac{C_0}{D_0}, \quad V_0 = \frac{R_0}{S_0}$$

where

$$\begin{aligned}
A_0 &= 1 + a_{01}q^{-1} + \cdots + a_{0n_0}q^{-n_0} \\
B_0 &= b_{01}q^{-1} + \cdots + b_{0n_0}q^{-n_0} \\
C_0 &= 1 + c_{01}q^{-1} + \cdots + c_{0n_0}q^{-n_0} \\
D_0 &= 1 + d_{01}q^{-1} + \cdots + d_{0n_0}q^{-n_0} \\
R_0 &= 1 + r_{01}q^{-1} + \cdots + r_{0n_1}q^{-n_1} \\
S_0 &= 1 + s_{01}q^{-1} + \cdots + s_{0n_1}q^{-n_1}
\end{aligned}$$

and n_0 and n_1 are upper bounds on the degrees. Moreover, we assume that the zeros of A_0, C_0, D_0, R_0 and S_0 are inside a circle of a known radius $\eta < 1$, i.e. we assume stability of the system with a known margin, and also that the transfer function between the noise sequence and the output has a stable inverse with the same stability margin. The zeros of B_0 is assumed to be inside a circle of known radius μ, where μ might be larger than 1, i.e. we allow for non-minimum phase zeros in the transfer function between $u(t)$ and $y(t)$, and finally we assume that $|b_{01}|$ is bounded by a known constant B.

2.2 Model Class

The model class considered is

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t) \tag{3}$$

where $v(t)$ is a disturbance and

$$\begin{aligned}
A(\theta) &= 1 + a_1 q^{-1} + \cdots + a_n q^{-n} \\
B(\theta) &= b_1 q^{-1} + \cdots + b_n q^{-n}
\end{aligned}$$

(3) can be written in linear regression form $y(t,\theta) = \phi^T(t)\theta + v(t)$ by introducing

$$\begin{aligned}
\phi(t) &= [-y(t-1), \cdots, -y(t-n), \\
&\quad\;\; u(t-1), \ldots, u(t-n)]^T \\
\theta &= [a_1, \ldots, a_n, b_1, \ldots, b_n]^T
\end{aligned}$$

Notice that the system itself does not need to belong to the model class.

2.3 The Identification Criterion

From a system identification perspective, the most important feature of the above model is its associated predictor which is given by

$$\hat{y}(t,\theta) = \phi^T(t)\theta$$

and the corresponding prediction error is

$$\epsilon(t,\theta) = y(t) - \hat{y}(t,\theta) \tag{4}$$

Ideally, one would like to choose θ such that the following theoretical identification cost

$$V(\theta) = E\epsilon^2(t,\theta) \tag{5}$$

is minimised. The value of θ which minimises (5) is given by

$$\theta^* = R^{-1}f \tag{6}$$

where

$$R = E\phi(t)\phi^T(t), \quad f = E\phi(t)y(t) \tag{7}$$

Since the data generation mechanism is unknown, one cannot compute the expected value (5) and the estimate (6). Instead the empirical version

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t,\theta) \tag{8}$$

is used, and the corresponding estimate is the well known least squares estimate

$$\hat{\theta}_N = R_N^{-1} f_N \tag{9}$$

where

$$R_N = \frac{1}{N}\sum_{t=1}^{N}\phi(t)\phi^T(t), \quad f_N = \frac{1}{N}\sum_{t=1}^{N}\phi(t)y(t) \tag{10}$$

Clearly, $\hat{\theta}_N$ can only be expected to be close to θ^* when the the number of data points tends to infinity, and this is indeed the case under mild assumptions, see e.g. Ljung (1999). However, we never have an infinite number of data points, and a question that arises naturally is to quantify the difference between $\hat{\theta}_N$ and θ^* for a finite N.

3 The main result

In this section we present the ellipsoidal confidence regions for the least squares estimate. First we bound the probability that the differences $R_N - R$ and $f_N - f$ exceed certain values (Theorem 3.1), and then we use these results to bound $\hat{\theta}_N - \theta^*$ (Theorem 3.2).

We have not made any attempt of optimising these bounds, and in some places we have made the bounds more conservative in order to get relatively simple expressions. The bounds are therefore looser than they need to be, but they illustrate how the confidence ellipsoids depend on important variables, and they show that in principle we can derive confidence ellipsoids for a finite number of data points.

Theorem 3.1 *For any finite N, $\epsilon > 0$, $\nu > 0$ we can explicitly compute values $\delta_{\phi\phi^T} \in (0,1]$ and $\delta_{\phi y} \in (0,1]$ such that*

$$Pr\left\{ |R_N - R| < \epsilon \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [1 \cdots 1] \right\} \geq 1 - \delta_{\phi\phi^T}$$

and

$$Pr\left\{ |f_N - f| < \nu \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right\} \geq 1 - \delta_{\phi y}$$

where the inequalities should be understood element by element.

Partial proof. See appendix A.

$\delta_{\phi\phi^T}$ and $\delta_{\phi y}$ are functions of $N, \epsilon, n, n_0, n_1, \eta$, μ, B, σ_w and σ_e. The functional dependencies of $\delta_{\phi\phi^T}$ and $\delta_{\phi y}$ are quite natural. In particular $\delta_{\phi\phi^T}$ and $\delta_{\phi y}$ tends to 1 when the bound on the pole positions $\eta \to 1$, and/or the system and model order $n, n_0, n_1 \to \infty$. This can be easily understood since under these conditions the prediction errors will have long range dependencies, and the probability that there is a large difference between the expected and empirical value increases. Also, as expected $\delta_{\phi\phi^T}$ and $\delta_{\phi y}$ tends to zero as $N \to \infty$, but notice that for small values of N, ϵ and ν, $\delta_{\phi\phi^T}$ and $\delta_{\phi y}$ may be equal to 1, in which case the Theorem does not yield any useful information.

We are now in the position that we can derive non-asymptotic confidence ellipsoids for the parameter estimate.

Theorem 3.2 *Assume that the data has been generated according to (1) and (2). Let $\hat{\theta}_N = R_N^{-1} f_N$ and $\theta^* = R^{-1} f$. If $R_N - 2n\epsilon I$ is positive definite, then*

$$(\hat{\theta}_N - \theta^*)^T (R_N - 2n\epsilon I)(\hat{\theta}_N - \theta^*) \leq$$

$$\frac{(\epsilon\sqrt{2n}||\hat{\theta}_N|| + \nu)^2 2n}{\lambda_{\min}(R_N - 2n\epsilon I)}$$

with probability at least $1 - \delta_{\phi\phi^T} - \delta_{\phi y}$ where $\delta_{\phi\phi^T}$ and $\delta_{\phi y}$ are given in Theorem 3.1. $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue, and n is the model order.

The shape of the non-asymptotic confidence ellipsoids are similar to those obtained from asymptotic theory under the assumption that the true system belongs to the model class. In the asymptotic case the ellipsoids are given by $(\hat{\theta}_N - \theta^*)^T R (\hat{\theta}_N - \theta^*)$, but since R is unknown it is in practice replaced by its sample mean R_N. The only difference between the ellipsoids is therefore that we subtract the diagonal matrix $2n\epsilon I$ in the non-asymptotic case. Note however, that even though the shape of the ellipsoids is similar, the probabilities we assign to the individual ellipsoids may be quite different in the asymptotic and finite sample case. The finite sample results tend to be on the conservative side.

Proof: From Theorem 3.1 it follows that

$$(R_N + \tilde{R})\theta^* = f_N + \tilde{f}$$

with probability at least $1 - \delta_{\phi\phi^T} - \delta_{\phi y}$ for some $\tilde{R}$ and $\tilde{f}$ satisfying

$$|\tilde{R}| \leq \epsilon \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [1 \cdots 1], \quad |\tilde{f}| \leq \nu \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

It follows that

$$\begin{aligned} &(\hat{\theta}_N - \theta^*)^T R_N (\hat{\theta}_N - \theta^*) = \\ &(\hat{\theta}_N - \theta^*)^T (R_N \hat{\theta}_N - R_N \theta^*) = \\ &(\hat{\theta}_N - \theta^*)^T (f_N - f_N - \tilde{f} + \tilde{R}\theta^*) = \\ &-(\hat{\theta}_N - \theta^*)^T \tilde{R} (\hat{\theta}_N - \theta^*) + \\ &(\hat{\theta}_N - \theta^*)^T (\tilde{R}\hat{\theta}_N - \tilde{f}) \end{aligned}$$

and hence

$$(\hat{\theta}_N - \theta^*)^T(R_N + \tilde{R})(\hat{\theta}_N - \theta^*) = (\hat{\theta}_N - \theta^*)^T(\tilde{R}\hat{\theta}_N - \tilde{f}) \tag{11}$$

Since $\tilde{R} + 2n\epsilon I$ is positive definite it follows that

$$(\hat{\theta}_N - \theta^*)^T(R_N + \tilde{R})(\hat{\theta}_N - \theta^*) \geq (\hat{\theta}_N - \theta^*)^T(R_N - 2n\epsilon I)(\hat{\theta}_N - \theta^*) \tag{12}$$

Next we observe that

$$(\hat{\theta}_N - \theta^*)^T(\tilde{R}\hat{\theta}_N - \tilde{f}) \leq \sum_{i=1}^{2n} |\hat{\theta}_i - \theta_i^*| \left(\sum_{j=1}^{2n} |\hat{\theta}_j|\epsilon + \nu\right)$$

From Hölder's inequality it follows that $(\sum_{j=1}^{2n} |\hat{\theta}_j|)^2 \leq 2n||\hat{\theta}||^2$. Hence

$$(\hat{\theta}_N - \theta^*)^T(\tilde{R}\hat{\theta}_N - \tilde{f}) \leq \sum_{i=1}^{2n} |\hat{\theta}_i - \theta_i^*|(\epsilon\sqrt{2n}||\hat{\theta}|| + \nu) \tag{13}$$

Thus, by combining (12), (11) and (13) it follows that

$$(\hat{\theta}_N - \theta^*)^T(R_N - 2n\epsilon I)(\hat{\theta}_N - \theta^*) \leq \sum_{i=1}^{2n} |\hat{\theta}_i - \theta_i^*|(\epsilon\sqrt{2n}||\hat{\theta}|| + \nu) \leq \sqrt{2n}||\hat{\theta} - \theta^*||(\epsilon\sqrt{2n}||\hat{\theta}|| + \nu)$$

This implies that

$$||\hat{\theta} - \theta^*|| \leq \frac{\sqrt{2n}(\epsilon\sqrt{2n}||\hat{\theta}|| + \nu)}{\lambda_{\min}(R_N - 2n\epsilon I)}$$

and hence

$$(\hat{\theta}_N - \theta^*)^T(R_N - 2n\epsilon I)(\hat{\theta}_N - \theta^*) \leq$$

$$\frac{(\epsilon\sqrt{2n}||\hat{\theta}_N|| + \nu)^2 2n}{\lambda_{\min}(R_N - 2n\epsilon I)}$$

■

4 Concluding remarks

In this paper we have derived non-asymptotic confidence ellipsoids for the least squares estimate. The shape of the ellipsoids is similar to that obtained using asymptotic theory, although the probabilities we assign to the ellipsoids can be quite different. The probability that the estimate belongs to a certain ellipsoids has a natural dependence on the volume of the ellipsoid, the data generating mechanism, the model order and the number of data points available.

References

[1] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes - Estimation and Prediction.* Lecture Notes in Statistics 110. Springer Verlag.

[2] Campi M. C. and E. Weyer (2000). "Finite sample properties of system identification methods". Submitted for publication.

[3] Ljung, L. (1999). *System Identification - Theory for the User.* Second edition. Prentice Hall.

[4] Spall, J.C. (1995). "Uncertainty Bounds for Parameter Identification with Small Sample Sizes." *Proc. of the 34th IEEE CDC*, New Orleans, Louisiana, US, pp. 3504-3515.

[5] Söderström, T. and P. Stoica (1989). *System Identification.* Prentice Hall.

[6] Weyer E. (2000). "Finite sample properties of system identification of ARX models under mixing conditions" *Automatica* Vol. 36, no. 9, pp. 1291-1299.

[7] Weyer E. and, M.C. Campi (1999). "Finite sample properties of system identification methods", *Proceedings of the 38th IEEE Conference on Decision and Control*, Phoenix, Arizona, USA, December 1999, pp. 510-515.

[8] Weyer E., R. C. Williamson, and I. M. Y. Mareels (1999). " Finite sample properties of linear model identification." *IEEE Transactions on Automatic Control*, Vol. 44, no. 7, pp. 1370-1383.

A Bounds on $R(N)$ and $f(N)$

The elements of $R(N)$ and $f(N)$ are of the forms $\frac{1}{N}\sum_{t=1}^{N} y(t-k)y(t-l)$, $\frac{1}{N}\sum_{t=1}^{N} y(t-k)u(t-l)$ and $\frac{1}{N}\sum_{t=1}^{N} u(t-k)u(t-l)$.

Due to space limitations we will only derive bounds for the difference between $\frac{1}{N}\sum_{t=1}^{N} y(t-k)y(t-l)$ and $Ey(t-k)y(t-l)$. The other bounds follow along the same lines.

Theorem A.1 *Let*

$$S_N^{yy} = \frac{1}{N}\sum_{t=1}^{N} y(t-k)y(t-l) - Ey(t-k)y(t-l)$$

Then

$$Pr\left\{|S_N^{yy}| \leq \epsilon\right\} \geq 1 - \delta_1 - \delta_2 - 2\delta_3$$

where

$$\delta_1 = 4\frac{e^{-\frac{N\epsilon_{ww}^2}{4\sigma_w^2(4\sigma_w^2+\epsilon_{ww})}}}{\left(1-e^{-\frac{N\epsilon_{ww}^2}{4\sigma_w^2(4\sigma_w^2+\epsilon_{ww})}}\right)^2}$$

$$\delta_2 = 4\frac{e^{-\frac{N\epsilon_{ee}^2}{4\sigma_e^2(4\sigma_e^2+\epsilon_{ee})}}}{\left(1-e^{-\frac{N\epsilon_{ee}^2}{4\sigma_e^2(4\sigma_e^2+\epsilon_{ee})}}\right)^2}$$

$$\delta_3 = 2\frac{e^{-\frac{N\epsilon_{we}^2}{2\sigma_w\sigma_e(2\sigma_w\sigma_e+\epsilon_{we})}}}{\left(1-e^{-\frac{N\epsilon_{we}^2}{2\sigma_w\sigma_e(2\sigma_w\sigma_e+\epsilon_{we})}}\right)^2}$$

$$\epsilon_{ww} \leq \frac{\epsilon(1-\eta)^{2n_0+2n_1+1}}{3\cdot 2^{2n_1}B^2(2(n_0+n_1)\eta+3(1-\eta))}\cdot\left(\frac{\eta}{\eta+\mu}\right)^{2n_0-2}$$

$$\epsilon_{ee} \leq \frac{\epsilon(1-\eta)^{2n_0+1}}{3\cdot 2^{n_0}(2n_0\eta+(1-\eta))}$$

$$\epsilon_{we} \leq \frac{\epsilon(1-\eta)^{2n_0+n_1+1}}{3\cdot 2^{n_0+n_1+1}B((2n_0+n_1)\eta+2(1-\eta))}\cdot\left(\frac{\eta}{\eta+\mu}\right)^{n_0-1}$$

Proof. Let

$$\begin{aligned} G_0V_0 &= g_1q^{-1}+g_2q^{-2}+\cdots \\ H_0 &= 1+h_1q^{-1}+h_2q^{-2}+\cdots \end{aligned}$$

and let

$$\begin{aligned} S_N^{ww}(i,j) = {} & \tfrac{1}{N}\textstyle\sum_{t=1}^N w(t-k-i)w(t-l-j) \\ & -\delta_{k+i-l-j}\sigma_w^2 \\ S_N^{ee}(i,j) = {} & \tfrac{1}{N}\textstyle\sum_{t=1}^N e(t-k-i)e(t-l-j) \\ & -\delta_{k+i-l-j}\sigma_e^2 \\ S_N^{we}(i,j) = {} & \tfrac{1}{N}\textstyle\sum_{t=1}^N w(t-k-i)e(t-l-j) \\ S_N^{ew}(i,j) = {} & \tfrac{1}{N}\textstyle\sum_{t=1}^N e(t-k-i)w(t-l-j) \end{aligned}$$

where $\delta_{k+i-l-j} = 1$ if $k+i = l+j$ and 0 otherwise.

Using (1) and (2) we find that

$$\begin{aligned} &\left|\frac{1}{N}\sum_{t=1}^N y(t-k)y(t-l) - Ey(t-k)y(t-l)\right| \leq \\ &\textstyle\sum_{i=1}^\infty\sum_{j=1}^\infty |g_i|\cdot|g_j|\cdot|S_N^{ww}(i,j)|+ \\ &\textstyle\sum_{i=0}^\infty\sum_{j=0}^\infty |h_i|\cdot|h_j|\cdot|S_N^{ee}(i,j)|+ \\ &\textstyle\sum_{i=1}^\infty\sum_{j=0}^\infty |g_i|\cdot|h_j|\cdot|S_N^{ew}(i,j)|+ \\ &\textstyle\sum_{i=0}^\infty\sum_{j=1}^\infty |h_i|\cdot|g_j|\cdot|S_N^{we}(i,j)| \end{aligned} \tag{14}$$

Suppose that $|S_N^{ww}(i,j)| \leq \epsilon_{ww}(i+j+1)$. Using the bound on the coefficients from Lemma B.1 and (19) and (20) we find that the first term on the "right hand" side of (14) is bounded by

$$\begin{aligned} &\textstyle\sum_{i=1}^\infty\sum_{j=1}^\infty |g_i|\cdot|g_j|\cdot|S_N^{ww}(i,j)| \leq \\ &2^{2n_1}B^2\left(1+\frac{\mu}{\eta}\right)^{2n_0-2}\frac{2(n_0+n_1)\eta+3(1-\eta)}{(1-\eta)^{2n_0+2n_1+1}}\epsilon_{ww} \end{aligned} \tag{15}$$

Similarly, by assuming that $|S_N^{ee}(i,j)| \leq \epsilon_{ee}(i+j+1)$ and $|S_N^{we}(i,j))| \leq \epsilon_{we}(i+j+1)$ we find that

$$\begin{aligned} &\textstyle\sum_{i=0}^\infty\sum_{j=0}^\infty |h_i|\cdot|h_j|\cdot|S_N^{ee}(i,j))| \leq \\ &2^{n_0}\frac{2n_0\eta+(1-\eta)}{(1-\eta)^{2n_0+1}}\epsilon_{ee} \end{aligned}$$

and

$$\begin{aligned} &\textstyle\sum_{i=1}^\infty\sum_{j=0}^\infty |g_i|\cdot|h_j|\cdot|S_N^{we}(i,j))| \leq \\ &2^{n_0+n_1}\left(1+\frac{\mu}{\eta}\right)^{n_0-1}B\frac{(2n_0+n_1)\eta+2(1-\eta)}{(1-\eta)^{2n_0+n_1+1}}\epsilon_{we} \end{aligned}$$

and it follows that

$$|S_N^{yy}| \leq \epsilon \tag{16}$$

Next we compute the probability that $|S_N^{ww}(i,j)| \leq \epsilon_{ww}(i+j+1)$ uniformly in i and j. Using (17) and (18) we find that this probability is at least $1-\delta$ where

$$\begin{aligned} \delta &= \sum_{i=0}^\infty\sum_{j=0}^\infty 4e^{-\frac{N\epsilon_{ww}^2(i+j+1)^2}{4\sigma_w^2(4\sigma_w^2+\epsilon_{ww}(i+j+1))}} \\ &= \sum_{m=0}^\infty (m+1)4e^{-\frac{N\epsilon_{ww}^2(m+1)^2}{4\sigma_w^2(4\sigma_w^2+\epsilon_{ww}(m+1))}} \\ &\leq 4\sum_{m=0}^\infty (m+1)e^{-\gamma(m+1)} \end{aligned}$$

where

$$\gamma = \frac{N\epsilon_{ww}^2}{4\sigma_w^2(4\sigma_w^2+\epsilon_{ww})}$$

Using (21) it follows that

$$Pr\left\{|S_N^{ww}(i,j)| \leq \epsilon_{ww}(i+j+1)\right\} \geq 1-\delta_1$$

where the probability is uniform in i and j.

Similarly we find that

$$Pr\left\{|S_N^{ee}(i,j)| \le \epsilon_{ee}(i+j+1)\right\} \ge 1-\delta_2$$

and

$$Pr\left\{|S_N^{ew}(i,j)| \le \epsilon_{we}(i+j+1)\right\} \ge 1-\delta_3$$

and the Theorem follows.

B Bounds on coefficients

Lemma B.1 *The coefficients of*

$$\begin{aligned} G_0V_0 &= g_1q^{-1} + g_2q^{-2} + \cdots \\ H_0 &= 1 + h_1q^{-1} + h_2q^{-2} + \cdots \end{aligned}$$

are bounded by

$$\begin{aligned} |g_k| &\le 2^{n_1} B \left(1+\frac{\mu}{\eta}\right)^{n_0-1} \cdot \\ &\quad \frac{k\cdots(k+n_0+n_1-2)}{(n_0+n_1-1)!}\eta^{k-1} \\ |h_k| &\le 2^{n_0}\frac{(k+1)\cdots(k+n_0-1)}{(n_0-1)!}\eta^k \end{aligned}$$

Proof: See Campi and Weyer (2000). ■

C Exponential inequalities

The main theorem we are going to make use of is the following one taken from Bosq (1998)

Theorem C.1 *Let $X_1,\ldots,X_N$ be independent zero mean real-valued random variables and let $S_N = \sum_{t=1}^N X_t$. Assume there exists a $c>0$ such that*

$$E|X_t|^k \le c^{k-2}k!EX_t^2 < \infty,$$

$$i = t,\ldots,N\; k = 3,4,\ldots$$

then

$$Pr\{|S_N| \ge \epsilon\} \le 2e^{-\frac{\epsilon^2}{4\sum_{t=1}^N EX_t^2 + 2c\epsilon}}$$

Corollary C.2 *Let $w(t)$ be a zero mean Gaussian variable with variance σ_w^2. Then*

$$Pr\left\{\left|\frac{1}{N}\sum_{t=1}^N w^2(t-k) - \sigma_w^2\right| \ge \epsilon(k,k)\right\} \le$$

$$2e^{-\frac{\epsilon^2(k,k)N}{4\sigma_w^2(2\sigma_w^2+\epsilon(k,k))}} \tag{17}$$

$$Pr\left\{\frac{1}{N}\left|\sum_{t=1}^N w(t-k)w(t-l)\right| \ge \epsilon(k,l)\right\} \le$$

$$4e^{-\frac{N\epsilon^2(k,l)}{4\sigma_w^2(4\sigma_w^2+\epsilon(k,l))}} \quad k \ne l \tag{18}$$

(17) follows from Theorem C.1, noting that $X_t = w^2(t) - \sigma_w^2$ satisfies the conditions in the Theorem with $c = 2\sigma_w^2$, $EX_t^2 = 2\sigma_w^4$.

In order to prove (18) for $k \ne l$ we use Theorem C.1 with $X_t = w(t-k)w(t-l)$. It follows that $EX_t = 0$, $EX_t^2 = \sigma_w^4$, $c = \sigma_w^2$. However, X_t, $t = 1,\ldots,N$ are not iid, but, we can group the time indices $\{1,2,\ldots,N\}$ into two set A_1 and A_2 such that X_t, $t \in A_1$ are iid random variables and X_t, $t \in A_2$ are iid random variables. For simplicity we assume that N is even and that A_1 and A_2 contain $N/2$ time indices each. Then we have

$$Pr\left\{\tfrac{1}{N}\left|\textstyle\sum_{t=1}^N w(t-k)w(t-l)\right| \ge \epsilon(k,l)\right\} \le$$

$$Pr\left\{\tfrac{1}{N}\left|\textstyle\sum_{t\in A_1} w(t-k)w(t-l)\right| \ge \tfrac{\epsilon(k,l)}{2}\right\} + Pr\left\{\tfrac{1}{N}\left|\textstyle\sum_{t\in A_2} w(t-k)w(t-l)\right| \ge \tfrac{\epsilon(k,l)}{2}\right\} =$$

$$2Pr\left\{\tfrac{1}{N}\left|\textstyle\sum_{t\in A_1} w(t-k)w(t-l)\right| \ge \tfrac{\epsilon(k,l)}{2}\right\} \le$$

$$4e^{-\frac{N\epsilon^2(k,l)}{4\sigma_e^2(4\sigma_e^2+\epsilon(k,l))}}$$

D Handy formulas

$$\sum_{k=0}^\infty \frac{(k+1)\cdots(k+n-1)}{(n-1)!}\eta^k = \frac{1}{(1-\eta)^n} \tag{19}$$

$$\sum_{k=0}^\infty \frac{k(k+1)\cdots(k+n-1)}{(n-1)!}\eta^k = \frac{n\eta}{(1-\eta)^{n+1}} \tag{20}$$

$$\sum_{m=0}^\infty (m+c_1)a^{m+c_2} = \frac{a^{c_2}}{(1-a)^2}(1+(c_1-1)(1-a)) \tag{21}$$