An Introduction to Inertial Sensor Stochastic Calibration

Stéphane Guerrier

Pennsylvania State University
Department of Statistics & Institute for CyberScience

This document was prepared with the help of
Dr. Roberto Molinari (Penn State), Dr. Mucyo Karemera (Penn State),
Gaetan Bakalli (U. Geneva), Haotian Xu (U. Geneva) & Yuming Zhang (Penn State)

Last update: April 2018

Outline

Outline

Introduction

General Framework:

- A **new framework for inertial sensor calibration.**
- It is based on the **Generalized Method of Wavelet Moments (GMWM)** of Guerrier et al. 2013, which is a new statistical approach to estimate the parameters of (complex) time series models.
- The GMWM is able to estimate efficiently time series models which are commonly used to describe the errors of inertial sensors.
- This calibration approach provides considerable improvements (in terms of navigation performance) compared to existing methods.
- This methodology is **robust** (potentially applicable for FDI purposes) and is able to **automatically select a suitable model** (or rank models).

Inertial Sensors

Inertial Measurement Unit (IMU):

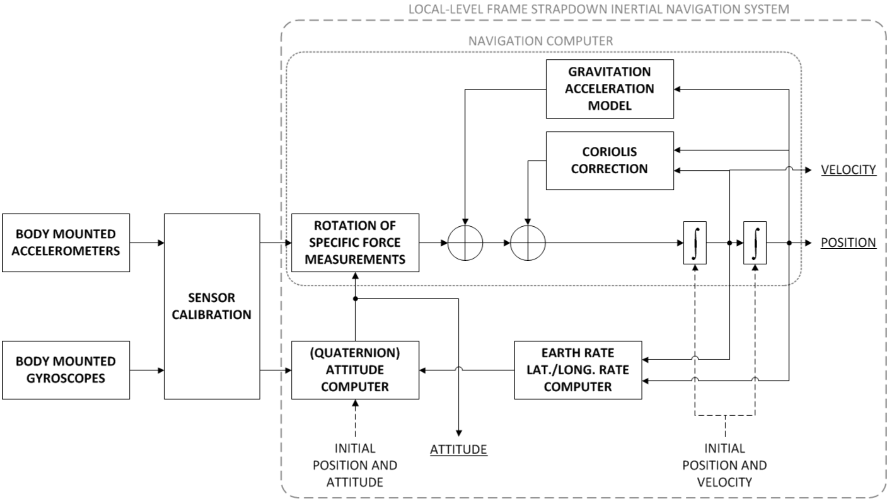
An IMU is composed of accelerometers and gyroscopes. Very widely employed as part of "integrated" navigation systems, a few examples:

- Robots
- (Small) **Unmanned Aerial Vehicles** (e.g. 3D modeling, search and rescue, etc)
- Autonomous underwater vehicles (e.g. water quality monitoring)
- Sport (e.g. ski, swimming, etc)
- Virtual reality (e.g. video games, movies, etc)

IMU in Navigation

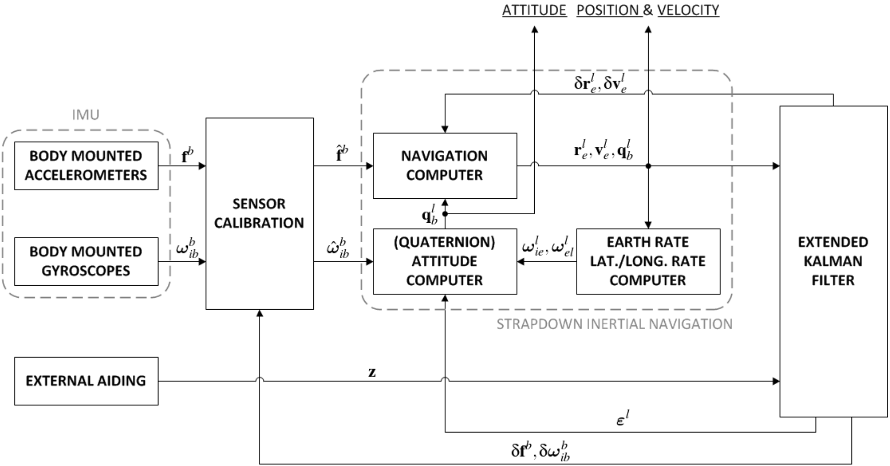
- **GPS clearly dominates** the current position and navigation market.
- It provides the whole range of navigation at very low cost.
- It is also highly portable, has low power consumption and is well suited for integration with other sensors.
- Unfortunately, **GPS does not work in all environments.**
- An IMU presents characteristics that are very different from GPS.
- It provides a **navigation solution when GPS-signals are unavailable**.
- It has higher sampling rate and **greatly improves orientation estimation**.
- The main challenge with IMUs (MEMS in particular) are their large sensor errors, which rapidly degrade the navigation performance.

Inertial Navigation



Source: Stebler 2012

IMU + other sensors (GPS)

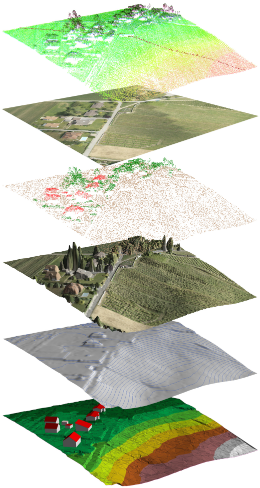


Source: Stebler 2012

Helicopter-based mobile mapping

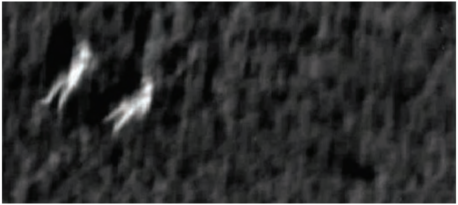
Ecole Polytechnique Fédérale de Lausanne (EPFL)

scan2map™
HANDHELD AIRBORNE MAPPING SYSTEM



UAV for search and rescue operations

Institut de Geomàtica (Barcelona), EPFL, ...



IMU Calibration

Errors in Inertial Sensors

- Possible causes:
  - Non-orthogonalities of the sensor axes
  - Environmental conditions (e.g. temperature)
  - Electronics
  - Dynamics
  - ...
- Error types:
  - Deterministic (calibration models, physical models, ...)
  - **Random error components** (typically latent time series models,...)

Correct stochastic sensor error modeling implies:

- Correct stochastic assumptions for inference
- **Better navigation or post-processing performance**

Effect on position of error model

Emulation setting:

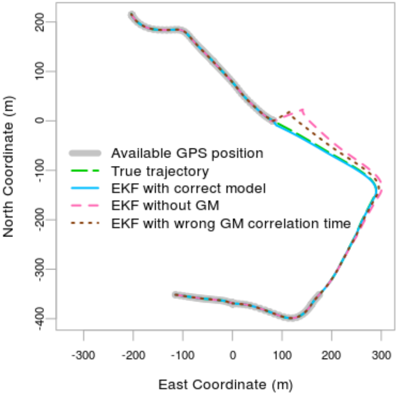
- Suppose the following model for inertial sensors (WN + GM):

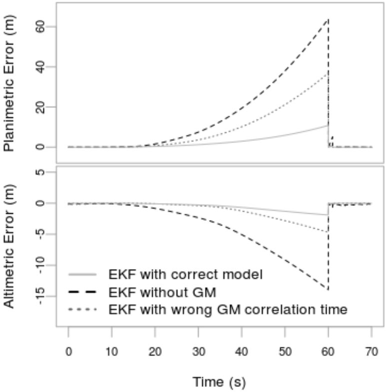
$$Y_t = \exp(-\beta \Delta t) Y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{GM}(1 - \exp(-2\beta\Delta t)))$$
$$X_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \;\; Z_t = X_t + Y_t.$$

- Consider the following three models:

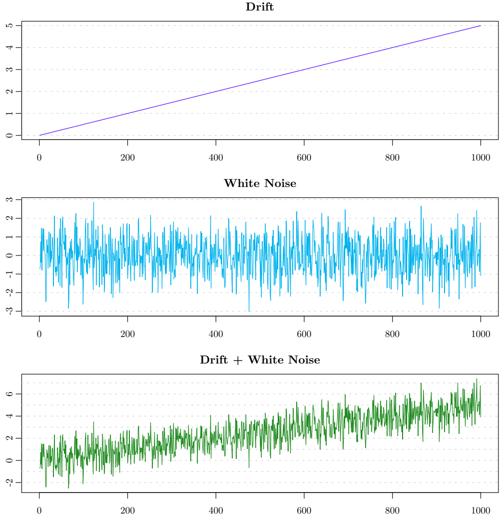
Sensor	**Scenario**	β	σ_{GM}	σ_{WN}
Acc.	Correct model	10^{-4}	50.0	70.0
	Wrong β	10^{-2}	50.0	70.0
	Without GM	–	–	70.0
Gyro.	Correct model	10^{-4}	10.0	30.0
	Wrong β	10^{-2}	10.0	30.0
	Without GM	–	–	30.0

Effect on position of error model





An easy latent time series model



Remarks:

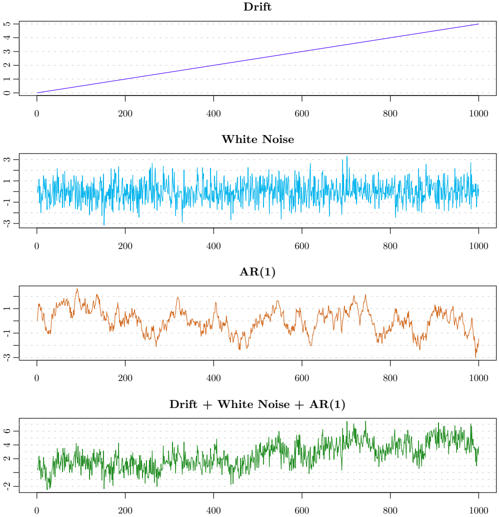
- Simple linear regression model:

$$y_t = \omega t + \varepsilon_t$$

$$\varepsilon_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

- MLE is perfectly fine.
- **What if we add an AR1 process?**

Adding an autoregressive process



Remarks:

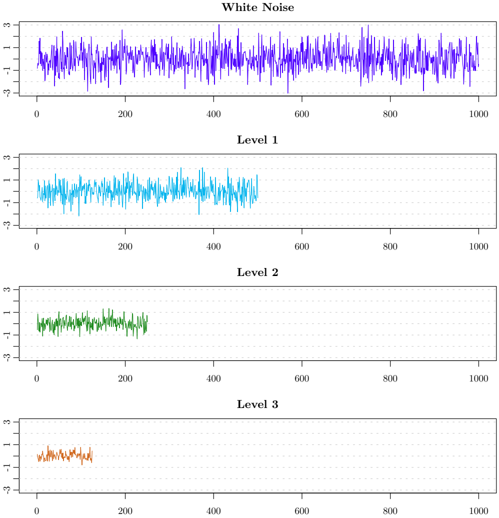
- Not a linear regression model but a **state space model**.
- Computing the likelihood is not an easy task (Kalman filter).
- **MLE (in fact EM-KF) fails**.

Estimation of Latent Time Series Models

Existing methods:

- Transforming into a “non-latent” model (e.g. ARMA)
  - Does not work in general.
  - Tends to diverge when one latent time series is “close” to unit root.
  - Difficult to “inverse”.
- MLE of an associated state-space (possibly using EM algorithm)
  - Computationally intensive.
  - Systematically diverges with “complex” models.
  - A lot of work is needed for a new model (see Stebler et al. 2011).
- “Graphical” method
  - Limited to a few possible models.
  - Not consistent in general (see Guerrier, Molinari, and Stebler 2016).
  - “Inefficient” (see Guerrier, Molinari, and Stebler 2016).

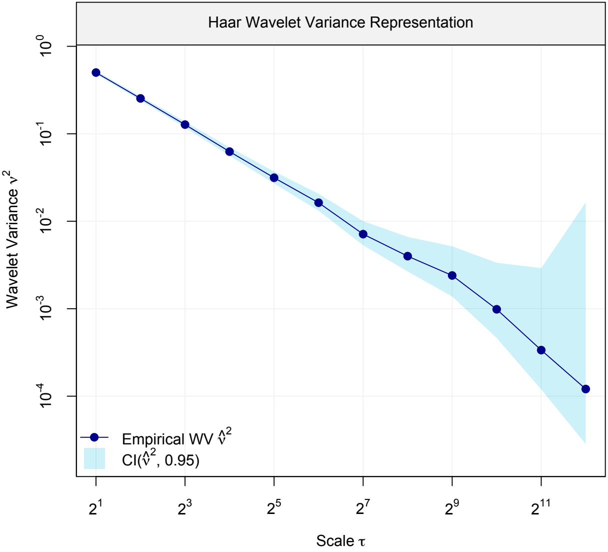
Looking differently at a time series using the Allan variance



Remarks:

- Proposed by *Allan, Proc. IEEE, 1966*
- Originally intended to study the stability of atomic clocks.
- Allan Variance is closely related to the (haar) Wavelet Variance.

The Wavelet (Allan) variance



Initial idea:

Match the WV:

- Exploit the relationship that exists between the model $F_{\boldsymbol{\theta}}$ and the WV $\boldsymbol{\nu}(\boldsymbol{\theta})$ (i.e. **mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\nu}(\boldsymbol{\theta})$**).
- "Inverse" this mapping by minimizing some discrepancy between empirical WV ($\hat{\boldsymbol{\nu}}$) and the theoretical WV $\boldsymbol{\nu}(\boldsymbol{\theta})$.
- This should provide an approximation of the point $\boldsymbol{\theta}(\hat{\boldsymbol{\nu}})$.

Wavelet Variance

Empirical WV:

- The WV (ν_j^2) is the **variance of wavelet coefficients** for the scale j.
- Wavelet coefficients ($W_{j,t}$) are weighted averages computed on the series Y_t.
- The weights are called wavelet filters h_j: e.g. the Haar wavelet filter.
- The wavelet filters give non-zero weights to observations at a given lag (window sizes of length L_j). Hence, there are as many WV as there are scales.
- The wavelet filters can be computed on consecutive windows, or on overlapping windows (to get $\widetilde{W}_{j,t}$ using $\tilde{h}_j$). Overlapping windows lead to more efficient estimators (such as the MODWT).

Wavelet Variance

Wavelet Variance

The Wavelet Variance (WV) is the variance of the wavelet coefficients, i.e.

$$\nu_j^2 = \mathrm{var}\left(W_{j,t}\right), \quad \text{where} \quad W_{j,t} = \sum_{l=0}^{L_j-1} h_{j,l} y_{t-l}, \; t \in \mathbb{Z}$$

and where $h_{j,l}$ are wavelet filters based on MODWT.

Remark

The Allan Variance (AV) is a **special case** of the WV, in fact $\phi_j^2 = 2\nu_j^2$ where ν_j^2 is based on Haar wavelet.

Estimation of the WV

MODWT estimator:

A consistent estimator for ν_j^2 is given by the MODWT estimator defined in Percival and Walden 2000

$$\hat{\nu}_j^2 = \frac{1}{M(T_j)} \sum_{t \in T_j} W_{j,t}^2.$$

Theorem: Asymptotic Normality

Serroukh, Walden, and Percival 2000 show that under suitable conditions

$$\sqrt{M(T_j)} \left(\hat{\nu}_j^2 - \nu_j\right) \xmapsto[T \to \infty]{\mathcal{D}} \mathcal{N}\left(0, S_{W_j}(0)\right).$$

A more general theorem

Theorem: Multivariate Extension

We extended this result to the multivariate case and demonstrated that under some regularity conditions

$$\sqrt{T}\left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\right) \underset{T \to \infty}{\overset{\mathcal{D}}{\longmapsto}} \mathcal{N}\left(0, \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\Sigma} = [\sigma^2_{kl}]_{k,l=1,\ldots,J}$.

Remark

This theorem generalizes Serroukh, Walden, and Percival 2000 result and enables to compute the (asymptotic) covariance between the WV (or the AV) at two different scales.

Theoretical WV

WV implied by $F_{\boldsymbol{\theta}}$:

Given a model $F_{\boldsymbol{\theta}}$ one can compute the theoretical WV as:

$$\nu_j = f\left(\boldsymbol{\theta}\right) = \int_{-1/2}^{1/2} |\widetilde{H}_j(f)|^2 S_{F_{\boldsymbol{\theta}}}(f) df.$$

Example:

Consider an AR1:

$$Y_t = \phi Y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right), \;\; t = 1, ..., T$$

With $\tau_j = 2^j$, the theoretical (Haar) WV of such process is given by

$$\nu_j = \frac{\left(\frac{\tau_j}{2} - 3\phi - \frac{\tau_j \phi^2}{2} + 4\phi^{\frac{\tau_j}{2}+1} - \phi^{\tau_j+1}\right)\sigma^2}{\frac{\tau_j^2}{2}\left(1-\phi\right)^2\left(1-\phi^2\right)}$$

WV of latent time series models

A very useful property:

Suppose we have

$$Y_t = X_t^{(1)} + ... + X_t^{(k)},$$

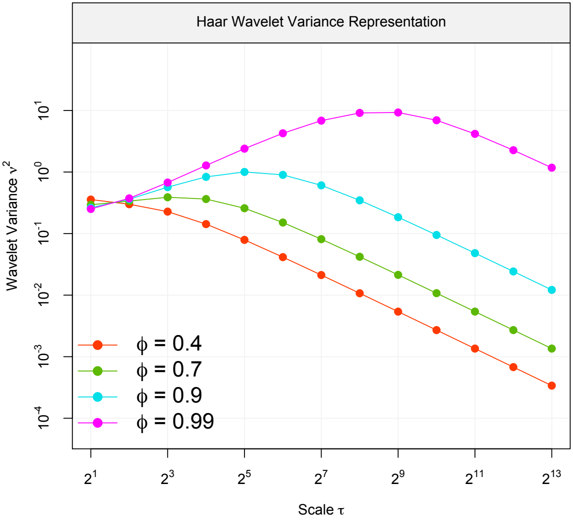
then the PSD of Y_t is

$$S_{Y_t} = S_{X_t^{(1)}} + ... + S_{X_t^{(k)}},$$

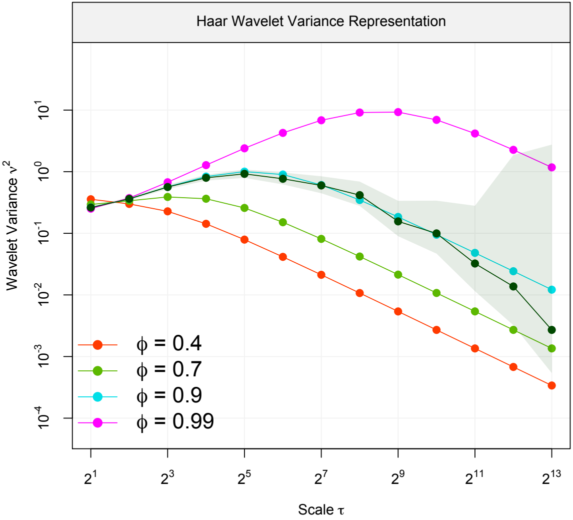
so the WV of Y_t is given by

$$\nu_{Y_t,j} = \int_{-1/2}^{1/2} |\widetilde{H}_j(f)|^2 \left(\sum_{i=1}^{k} S_{X_t^{(i)}} \right) df = \sum_{i=1}^{k} \nu_{X_t^{(i)},j}.$$

Principle of the GMWM



Principle of the GMWM



The GMWM estimator

Definition:

The GMWM estimator is the solution of the following optimization problem

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \; \left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})\right)^T \boldsymbol{\Omega} \left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})\right),$$

in which $\boldsymbol{\Omega}$, a positive definite weighting matrix, is chosen in a suitable manner such that the above quadratic form is convex.

Theorem: Consistency

$\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ (under some regularity conditions) for a large class of (latent or composite) models.

Asymptotic distribution of $\hat{\boldsymbol{\theta}}$

Theorem: Asymptotic Normality

We showed that (under some regularity conditions)

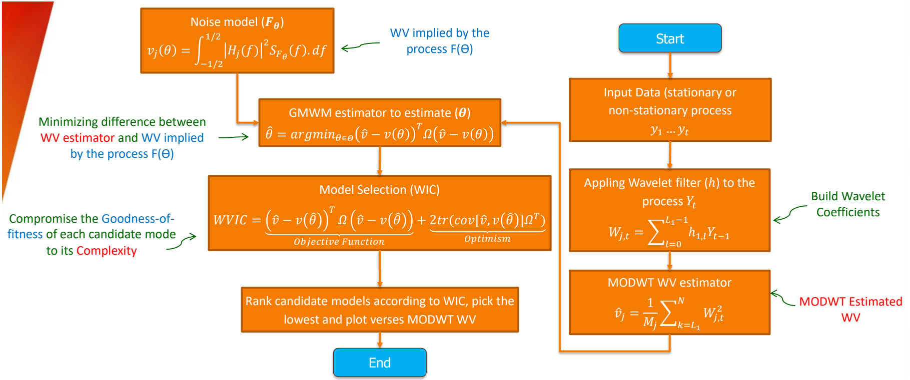
$$\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow[T\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \mathbf{V}\right),$$

where $\mathbf{V} = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T$, $\mathbf{B} = \left(\mathbf{D}^T\boldsymbol{\Omega}\mathbf{D}\right)^{-1}\mathbf{D}^T\boldsymbol{\Omega}$, $\mathbf{D} = \partial\boldsymbol{\nu}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ and $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\nu}}$.

Choosing $\boldsymbol{\Omega}$:

The most efficient estimator is (asymptotically) obtained by choosing $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, leading then to $\mathbf{V} = (\mathbf{D}^T\boldsymbol{\Sigma}^{-1}\mathbf{D})^{-1}$.

Flowchart GMWM



Thanks to Ahmed Radi

A small Example...

A simulated example:

Let $(Y_t) \ : t = 1, ..., 10^5$ be a simulated signal composed of a:

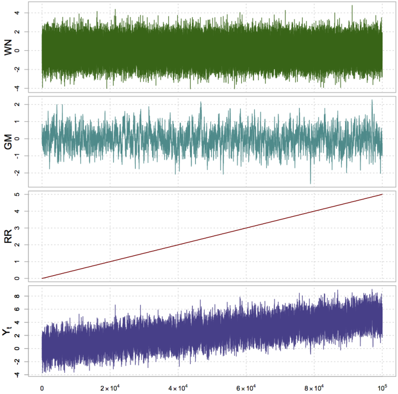
- First-order Gauss-Markov:
  $Y_t = \exp(-\beta \Delta t) Y_{t-1} + \varepsilon_t, \ \ \varepsilon_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_G^2 \left(1 - \exp(-2\beta\Delta t)\right)\right)$
- Gaussian White Noise: $X_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$
- Drift (rate ramp): $R_t = \omega t$

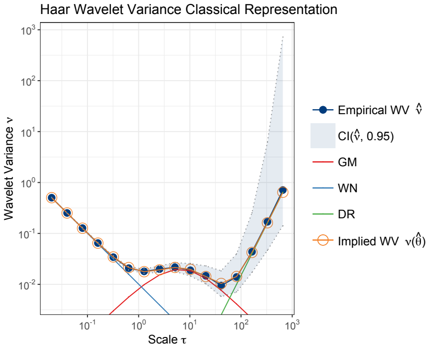
The observed process is therefore $Z_t = Y_t + X_t + R_t$ (which we write GM + WN + DR) and we have that $Z_t \sim F_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} = (\sigma, \sigma_G, \beta, \omega)$

A small Example...

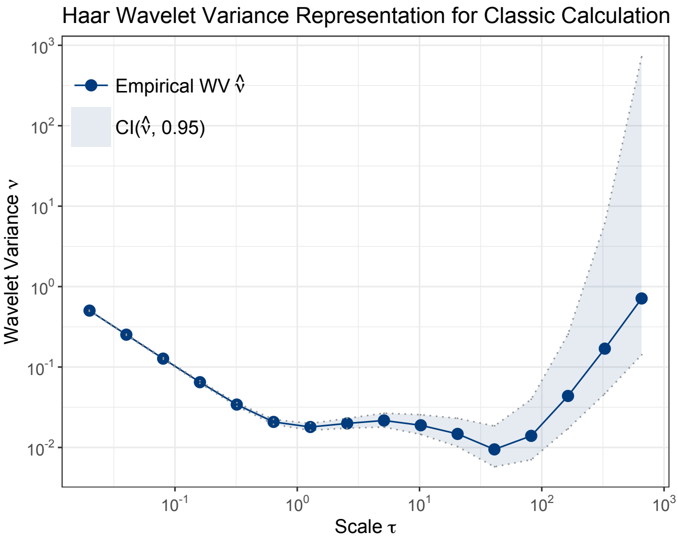
Simulation parameters:

$$\boldsymbol{\theta} = (\sigma^2, \beta, \sigma_G^2, \omega) = (1,\ 0.6,\ 0.1,\ 5 \cdot 10^{-5})$$

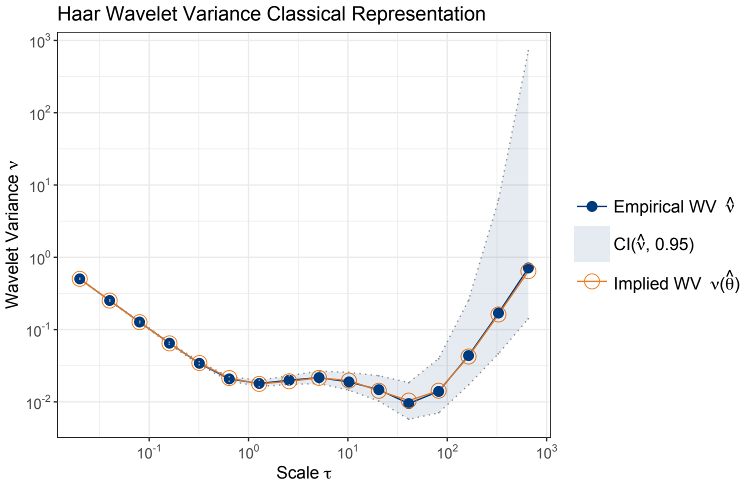




A small Example...



GMWM estimation results



GMWM estimation results

Estimated parameters:

	θ_0	$\hat{\theta}$	$\text{IC}(\theta_0, 0.95)$
σ^2	1.00	1.00	$(0.99; 1.01)$
β	0.60	0.58	$(0.55; 0.61)$
σ_G^2	10^{-1}	$1.07 \cdot 10^{-1}$	$(0.99 \cdot 10^{-1}; 1.12 \cdot 10^{-1})$
ω	$5 \cdot 10^{-5}$	$4.87 \cdot 10^{-5}$	$(4.67 \cdot 10^{-5}; 5.07 \cdot 10^{-5})$

Goodness of fit test:

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})\right)^T \boldsymbol{\Omega} \left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})\right) \xrightarrow[T \to \infty]{\mathcal{D}} \chi^2_{m-p}$$

In this case, the p-value of the test is ≈ 1. Therefore we cannot reject that $F_{\boldsymbol{\theta}}$ is the right model.

Simulation

Setting:

The model "White Noise + Gauss Markov + Drift" is in fact commonly used for inertial sensors. We simulate $B = 100$ signals of length $T = 6000$ from this model with $\boldsymbol{\theta}_0 = (\sigma, \sigma_G, \beta, \omega) = (2, 4, 0.05, 0.005)$

Comparison with MLE (EM-KF)

	GMWM		EM-KF	
	RMSE	R-RMSE	RMSE	R-RMSE
σ_G^2	0.96	0.06	74.57	4.66
β	$4.63 \cdot 10^{-3}$	0.09	0.04	0.85
σ^2	0.11	0.03	0.16	0.04
ω	$2.79 \cdot 10^{-4}$	0.06	0.12	23.58

Model Selection

Building upon the GMWM approach, it was possible to derive a criterion, called the **Wavelet Variance Information Criterion (WVIC)**, based on which it is possible to determine the most appropriate probabilistic model to predict the stochastic error signal of an inertial sensor.

Wavelet Variance Information Criterion (WVIC)

This criterion is defined as follows

$$\text{WVIC} = \mathbb{E}\left[\mathbb{E}_0\left[\left(\hat{\boldsymbol{\nu}}^0 - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right)^T \boldsymbol{\Omega}\left(\hat{\boldsymbol{\nu}}^0 - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right)\right]\right], \tag{1}$$

where $\mathbb{E}[\cdot]$ and $\mathbb{E}_0[\cdot]$ represent specific probabilistic expectations which measure how well the WV implied by the estimated model fits the WV observed on a future replicate.

However, it can be extremely useful to use the other replicates as “future” replicates.

Experiment Setup

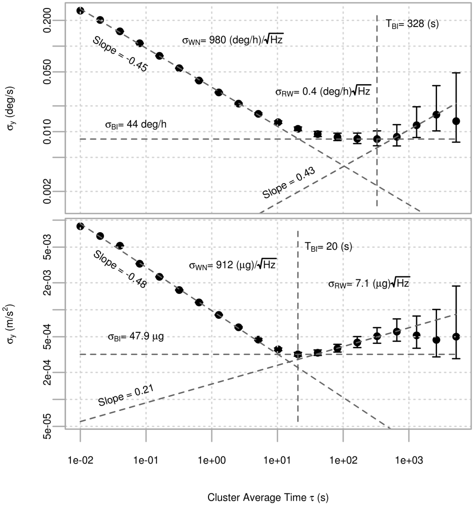
Data:

- Static data collected during 4.5 hours @100Hz
- Constant temperature condition
- Xsens MTi-G IMU

Sensor calibration approaches:

- Allan Variance
- KF-(Self)-tuning approach
- GMWM approach

Allan Variance Approach



KF-(Self)-tuning approach

Procedure:

- AV parameters used as initial approximation
- Analysis of position drift during artificial GNSS outages
- L1/L2 carrier-phased DGPS reference solution

Model:

$Y_t \sim F_{\theta}$ where F_{θ} is such that

$$Y_t = Y_{t,WN} + Y_{t,GM}$$

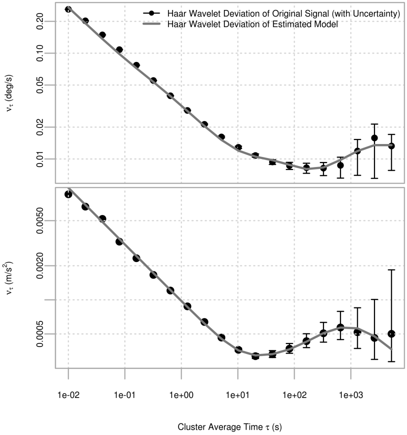
where $Y_{t,WN}$ and $Y_{t,GM}$ denote, respectively, a white noise and a Gauss-Markov process.

GMWM approach

Model:

$Y_t \sim F_\theta$ where F_θ is such that

$$Y_t = Y_{t,WN} + \sum_{k=1}^{3} Y_{t,GM}^{(k)}$$



Validation

Models comparison is non-trival...

- True model F_θ is **unknown...**
- Calibration on signal acquired in static conditions.

Proposed procedure:

1. INS/GNSS reference solution (L1/L2 DGPS + tactical/navigation grade IMU)
2. Emulation of synthetic IMU and GNSS signals along the reference
3. Addition of real static noise signal on synthetic signals
4. Introduction of artificial GNSS gaps
5. Processing procedure using closed-loop EKF to implement the models
6. Quality judged by analyzing the navigation error and EKF-predicted accuracy during inertial coasting model

Helicopter experiment:

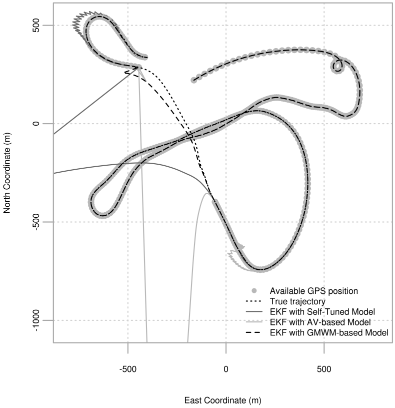
Reference:

Helicopter trajectory (Litton LN200 tactical-grade IMU @400Hz + Javad Legacy L1/L2 GNSS Rx @10Hz)

Helicopter experiment:

Reference:

Helicopter trajectory (Litton LN200 tactical-grade IMU @400Hz + Javad Legacy L1/L2 GNSS Rx @10Hz)



The GMWM computational platform

Outline

1. **Visualizing the signal(s)**:
   - Is the signal disturbed (contaminated)?
   - Which models could best describe the visualized WV?
2. **Estimating the model(s)**
   - The `gmwm.imu()` function for parameter estimation
   - The options for estimation
3. **Inference**
   - Confidence intervals for the parameters
   - Goodness-of-fit of the model to the signal
4. **Model Selection**
   - The efficient computation of the WVIC
   - Model selection from a set of user-specified models
   - Automatic model selection from a set of all sub-models of a single user-specified model

The GMWM package in action

Getting to know the package

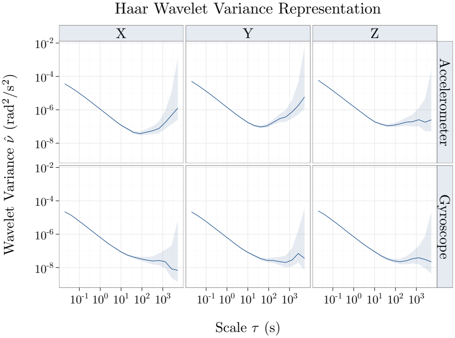
The package comes with multiple datasets

- To use the data within the `R` session:
  1. Load the `imudata` package (separate from the `gmwm` package)
  2. Make the dataset available in the session by typing `data(imu)`
- Load external data (e.g. binary data) with `read.imu()`

Visualizing the data

The function `wvar.imu()`

Once the calibration data has been entered, it is possible to plot the observed WV vs τ on a log-log scale



R code:

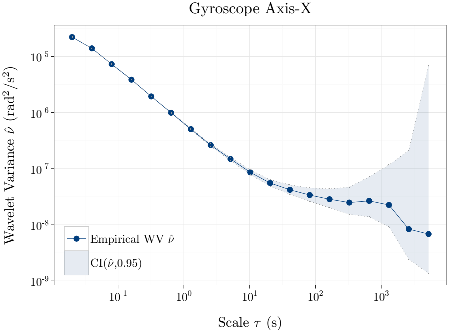
```
imu.obj = imu(imu6,
gyroscope = 1:3,
accelerometer = 4:6)
wv = wvar.imu(imu.obj)
plot(wv)
```

(Run time: 3.70 [sec])

Visualizing the data

The function `wvar()`

It is obviously possible to plot the WV from the axis of an accelerometer or gyroscope individually



R code:

```
WV.gx = wvar(imu6[,1])
plot(WV.gx)
```

(Run time: 0.65 [sec])

GMWM estimation

The `gmwm.imu()` function

The function tailor-made for IMU error modeling is `gmwm.imu()`

Main arguments of the `gmwm.imu()` function

- `model`: The structure of the model is specified through this argument
- `data`: The signal to be modelled
- `compute.v`: The method for computing the weighting matrix $\boldsymbol{\Omega}$
  - `fast`: Estimated diagonal matrix
  - `diag`: Use a diagonal matrix with the asymptotic variance of $\hat{\boldsymbol{\nu}}$
  - `bootstrap`: Estimate $\boldsymbol{\Omega}$ through parametric bootstrap

GMWM estimation

The `model` argument

The GMWM package can estimate each specific model or any latent model made by a combination of all or a subset of the following models

- `AR1()`: a first-order autoregressive process (reparametrization of Gauss-Markov process)
- `WN()`: white noise process
- `QN()`: quantization noise (rounding error)
- `RW()`: random walk process
- `DR()`: drift
- `AR()`: p-order autoregressive process
- `MA()`: q-order moving average process
- `ARMA()`: autoregressive-moving average processes

GMWM estimation

Latent model syntax

To specify a latent model we use the sign "+" between the available models while the `AR1()` model can be included more than once (say k times) and it can be specified as `k*AR1()`, for example

- `ARMA()+WN()+RW()`
- `3*AR1()+DR()`

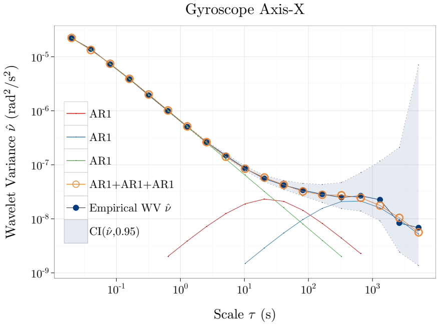
Convergence

If the `gmwm.imu()` function has problems of convergence, one can specify **starting values** for the parameters using the brackets in the syntax of each model (e.g. `WN(sigma2=0.5)`)

GMWM estimation

Visually assessing the fit

The function `plot()` applied to the object of a GMWM estimation allows to see how well the WV implied by the estimated model fits the observed WV



R code:

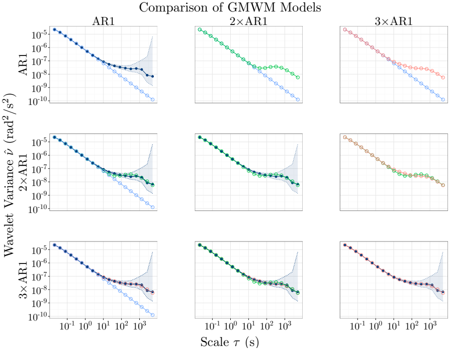
```
mod = gmwm.imu(3*AR1(),
imu6[,1]) plot(mod,
process.decomp = T)
```

(Run time: 0.56 [sec])

Model Comparison

Suppose we want to compare three models...

`AR1()`, `2*AR1()` and `3*AR1()`



R code:

```r
Xt = imu[,1]
m1 = gmwm.imu(AR1(),Xt)
m2 = gmwm.imu(2*AR1(),Xt)
m3 = gmwm.imu(3*AR1(),Xt)
compare.models(m1, m2, m3)
```

(Run time: 5.20 [sec])

Inference

Output of estimation

Aside from the value of the estimated parameters, the output of the function `gmwm.imu` provides other information for inference which include **confidence intervals** and **goodness of fit** test (GoF)

```
mod = gmwm.imu(2*AR1(), imu6[,1])
summary(mod, inference = T)
```

```
Model Information:
          Estimates       CI Low       CI High           SE
AR1     9.998700e-01 9.998336e-01 9.999064e-01 2.211547e-05
SIGMA2  5.223319e-11 4.343878e-11 6.102759e-11 5.346620e-12
AR1     1.265324e-01 1.265324e-01 1.265324e-01 1.765050e-08
SIGMA2  5.031185e-05 5.021566e-05 5.040805e-05 5.848209e-08

* The initial values of the parameters used in the minimization
of the GMWM objective function were  generated by the program
underneath seed: 1337.

Objective Function: 20.9297

Asymptotic Goodness of Fit:
Test Statistic: 903.8 on 15 degrees of freedom
The resulting p-value is: 0
```

WVIC model selection

Two options for model selection

- **Manual**: the function `rank.models()` allows the user to specify the set of models from which to select
- **Automatic**: the function `auto.imu()` allows to specify one single model from which all sub-models are generated to create a set from which to select

R code:

```
rank.models(3*AR1()+WN(),
2*AR1()+QN(), imu6[,1], nested =
F, bootstrap = F, model.type="imu",
robust = F)
```

```
mod.res = auto.imu(imu6, model
= 3*AR1()+WN()+RW()+QN()+DR(),
bootstrap = F, robust = F)
```

WVIC model selection

Automatic IMU model selection

Suppose we want to apply the `auto.imu()` function to the X-axis gyroscope in the `imu` dataset and consider all model combinations within the `4*AR1() + WN() + RW()` model

```
Xt = imu6[,1]
mod = 4*AR1()+WN()+RW()
mod.sel = auto.imu(Xt, model = mod)
summary(mod.sel)
```

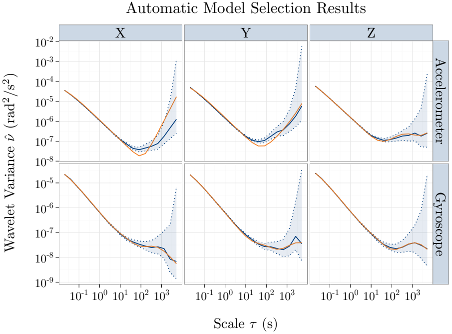
```
The model ranking for data column 1:
                              Obj Fun  Optimism   Criterion  GoF P-Value
1. AR1 AR1 AR1                17.2769    0.2974     17.5743            0
2. AR1 AR1 AR1 RW             17.2679    0.3139     17.5818            0
3. AR1 AR1 AR1 AR1            17.6022    0.4330     18.0351            0
4. AR1 AR1 RW                 20.1044    0.4871     20.5915            0
5. AR1 AR1                    20.9864    0.2070     21.1933            0
...
19. RW                     28420.4972    0.0160  28420.5132            0
```

(Run time: 50.25 [sec] ≈ 14x faster than the bootstrap option 685.4 [sec])

WVIC Model Selection

Visualizing the Automatic IMU Model Selection

We can observe the results of the `auto.imu()` function by requesting a plot. The plot will contain the empirical wavelet variance in addition to the best implied wavelet variance that we find. In this particular case, we used the defaults of the `auto.imu()` (63 models $\times$ 6 columns).

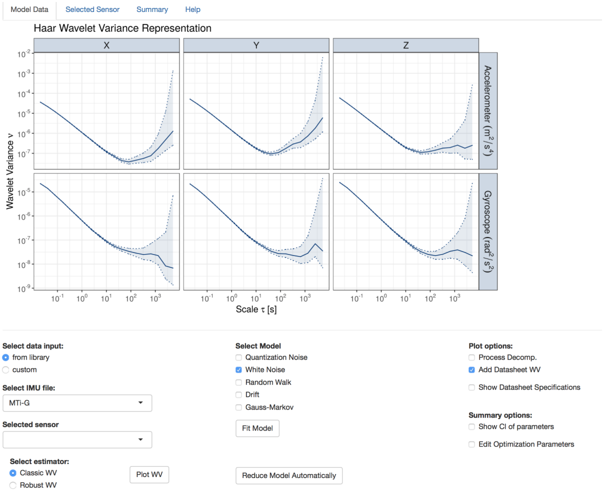


R code:

```
imu.obj = imu(imu6, gyroscope = 1:3,
accelerometer = 4:6, axis = c('X',
'Y', 'Z'))
auto.mod = auto.imu(imu.obj)
plot(auto.mod)
```

(Run time: 377.77 [sec])

Web-based Platform: gui4gmwm

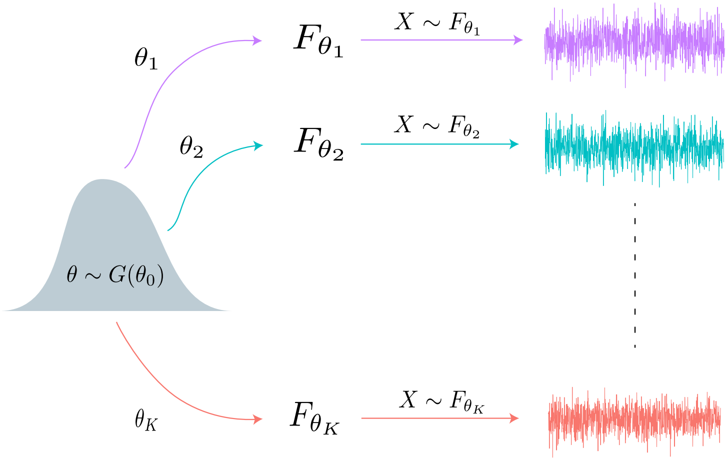


GMWM Extensions

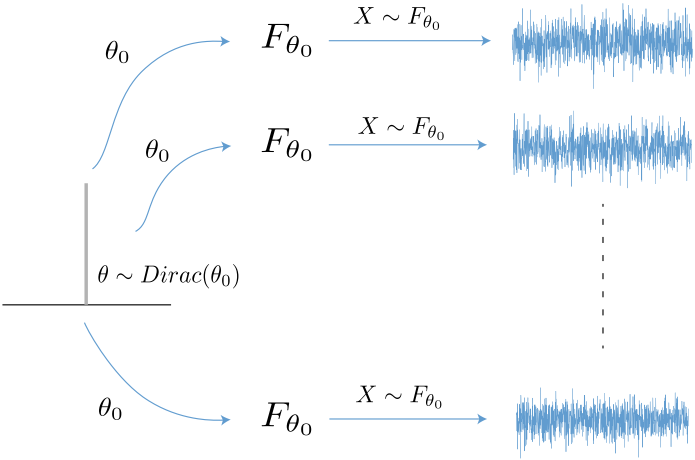
Extensions:

- Robust estimation (+ automatic outlier detection).
- Estimation and model selection based on multiple replicates.
- Extension to "near-stationarity" settings.
- Modelling dependence **between** sensors.
- Dynamic calibration (i.e. estimation parameters by taking into account external factors such as dynamics, temperature and atmospheric pressure).

Near Stationarity Setting



Stationarity Setting



GMWM with multiple replicates

MGMWM

Considering $k = 1, \ldots, K$, the number replicates recorded from the same IMU in static condition, we define the **Multisignal GMWM** estimator as the solution of the following minimization problem:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmin}} \, \frac{1}{K} \sum_{k=1}^{K} \|\hat{\boldsymbol{\nu}}_k - \boldsymbol{\nu}(\boldsymbol{\theta})\|_{\boldsymbol{\Omega}}^2. \tag{2}$$

Let $\hat{\nu}_{jk}$, respectively $\nu_j(\boldsymbol{\theta})$ be the j^{th} elements of the vectors $\hat{\boldsymbol{\nu}}_k$, the empirical WV computed on the k^{th} replicates of size $j = 1, \ldots, J_k$.

Definition: Near Stationarity

Near stationarity

We define a nearly stationary time series, as one which exhibit the following properties:

1. Same model, but with **different parameter values** for each sequences.
2. The vector of parameter $\boldsymbol{\theta}$ has a **probability distribution** $\boldsymbol{G}(\boldsymbol{\theta}_0)$, where $\mathbb{E}[G] = \boldsymbol{\theta}_0$.
3. The distribution $\boldsymbol{G}(\boldsymbol{\theta}_0)$ can be interpreted as **the internal sensor model**, which may account for unobserved factors (e.g. temperature).

Remark:

Near stationarity is a new statistical concept which **contradict the almost always assumed strong stationarity**. Further research is needed to understand this framework.

Two Estimators: Average GMWM vs MGMWM

Replicate Error Signals

Let us assume we have K independent replicates of the observed stochastic error of an inertial sensor. Can we use all this information?

Average GMWM vs MGMWM

We define $\hat{\boldsymbol{\theta}}^{\circ}$ as the Average GMWM (AGMWM) defined as:

$$\hat{\boldsymbol{\theta}}^{\circ} = \frac{1}{K} \sum_{k=1}^{K} \tilde{\boldsymbol{\theta}}_k \,,$$

with $\tilde{\boldsymbol{\theta}}_k = \text{argmin}_{\boldsymbol{\theta}_k \in \boldsymbol{\Theta}} \|\hat{\boldsymbol{\nu}}_k - \boldsymbol{\nu}(\boldsymbol{\theta}_k)\|^2_{\boldsymbol{\Omega}_k}$. Remeber that we define the Multisignal GMWM (MGMWM) as

$$\hat{\boldsymbol{\theta}}^{\star} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmin}} \, \frac{1}{K} \sum_{k=1}^{K} \|\hat{\boldsymbol{\nu}}_k - \boldsymbol{\nu}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Omega}_k}.$$

Two Estimators: Average GMWM vs MGMWM

Properties

It turns out that the MGMWM appears far more appropriate than the AGMWM for two main reasons:

- The MGMWM is more efficient than the AGMWM, i.e.

$$\frac{\text{tr}\left(\min_{\Omega_i} \text{var}\left[\hat{\boldsymbol{\theta}}^{\circ}\right]\right)}{\text{tr}\left(\min_{\Omega_i} \text{var}\left[\hat{\boldsymbol{\theta}}^{\star}\right]\right)} \xmapsto{\mathcal{P}} c > 1.$$

- From Jensen inequality implies that
  - If $\boldsymbol{\nu}(\boldsymbol{\theta})$ is linear, i.e for stochastic processes WN, DR and QN, or if $G(\boldsymbol{\theta}_0)$ is a Dirac function,

$$\hat{\boldsymbol{\theta}}^{\star} - \hat{\boldsymbol{\theta}}^{\circ} \xmapsto{\mathcal{P}} 0.$$

  - If $\boldsymbol{\nu}(\boldsymbol{\theta})$ is not linear i.e for stochastic processes RW and AR1, than

$$\hat{\boldsymbol{\theta}}^{\star} - \hat{\boldsymbol{\theta}}^{\circ} \xmapsto{\mathcal{P}} \delta \neq 0.$$

Testing for Near Stationarity

This frameworks allows to use the MGMWM objective function as a test statistics, which is defined below:

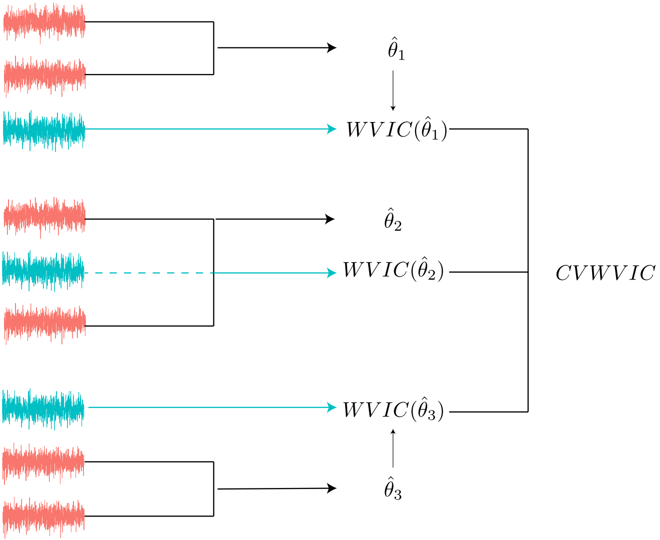
$$g(\hat{\boldsymbol{\theta}}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\|^2_{\hat{\boldsymbol{\Omega}}}.$$

Using this statistic we would therefore like to test the following null and alternative hypotheses

$$\mathcal{H}_0 : \boldsymbol{\theta}_k = \boldsymbol{\theta}_0, \; \forall \, k,$$
$$\mathcal{H}_a : \mathcal{H}_0 \text{ is false.}$$

Under $\mathcal{H}_0$, we fall in the case where the distribution G is a Dirac function and $\boldsymbol{\theta}_k = \boldsymbol{\theta}_0, \forall k$.

WVIC with Multiple Signal Replicates



CV-WVIC

Notation

- We partition the K replicates in two parts: the first part of size k_1 (with $1 \leq k_1 < K$), is used for **training purposes**, where $M = \binom{K}{k_1}$ denotes the size of every possible combination $m_i^{(k_1)}$ for $m = 1, \ldots, M$ and for $i = 1, \ldots, k_1$ whereas, the second part of size $k_2 = K - k_1$, for which $m_l^{(k_2)}$, for $l = 1, \ldots, k_2$, defines the complement of $m_i^{(k_1)}$, is used for **validation purpose**.
- Let $\mathcal{M}_\jmath$, $\jmath = 1, \ldots, \mathcal{J}$, denote the $\jmath^{th}$ model out of the set of $\mathcal{J}$ candidate models, with $\boldsymbol{\theta}_\jmath \subset \boldsymbol{\Theta}_\jmath \subseteq \mathbb{R}^p$ being the unknown parameter vector of $\mathcal{M}_\jmath$ and its associated parameter space.

CV-WVIC: Estimation and Validation Step

1. We use $m_i^{k_1}$, one specific combination of the error signals replicates of size k_1 to estimate the parameter vector $\boldsymbol{\theta}_{\jmath}^{(m_i)}$ for the $\jmath^{th}$ model in the following manner

$$\hat{\boldsymbol{\theta}}_{\jmath}^{(m_i)} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\jmath}}{\text{argmin}} \sum_{k \in m_i^{k_1}}^{k_1} ||\hat{\boldsymbol{\nu}}_k - \boldsymbol{\nu}(\boldsymbol{\theta}_{\jmath})||_{\boldsymbol{\Omega}}^2. \tag{3}$$

2. We than compute the MGMWM objective function, evaluated at the parameter estimated in Eq. (3), to deliver the WVIC computed on an independent sample $\boldsymbol{X}_t^{(k)}$ with $k \in c_{k_2}^{(m)}$:

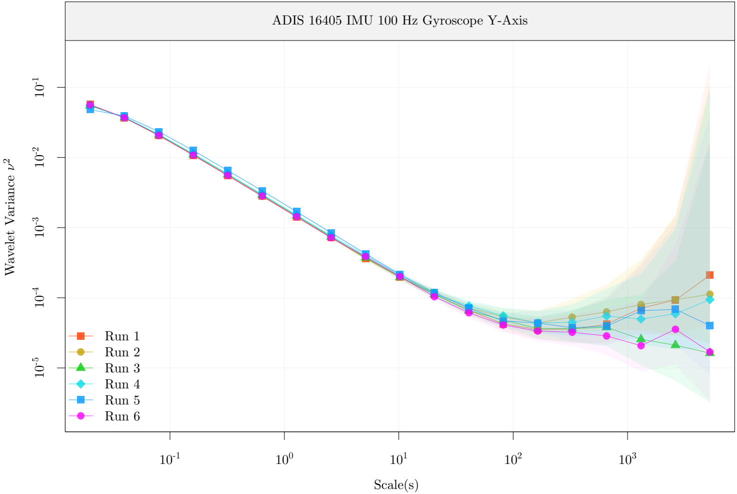
$$\widehat{\mathcal{C}}_{\jmath}^{(m_l)} = \sum_{k \in m_l^{k_2}}^{k_2} \|\hat{\boldsymbol{\nu}}_k - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}}_{\jmath}^{(m_i)})\|_{\boldsymbol{\Omega}}^2. \tag{4}$$

CV-WVIC

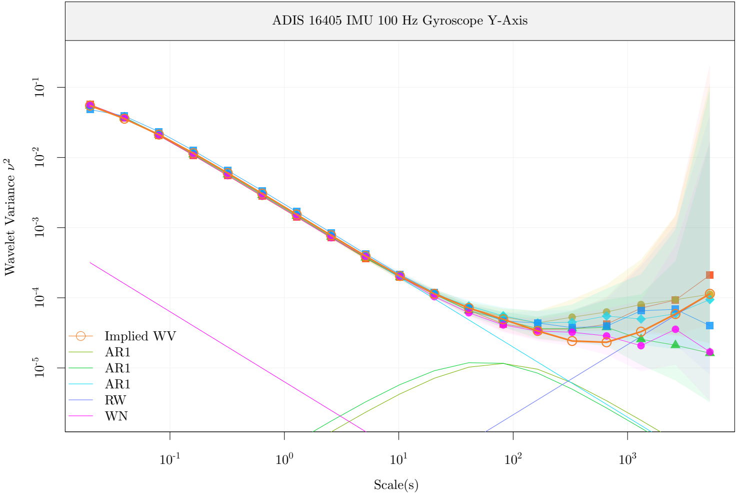
Repeating Eq. (3) and (4) for $m = 1, \ldots, M$, and computing the average $\widehat{\mathcal{C}}_\jmath = \frac{1}{M}\sum_{m=1}^{M} \widehat{\mathcal{C}}_\jmath^{(m_l)}$, for each model in the $\mathcal{J}$, will allow us to select the model which minimize $\widehat{\mathcal{C}}_\jmath$, i.e.

$$\hat{\jmath} = \underset{\jmath=1,\ldots,\mathcal{J}}{\text{argmin}}\, \widehat{\mathcal{C}}_\jmath. \tag{5}$$

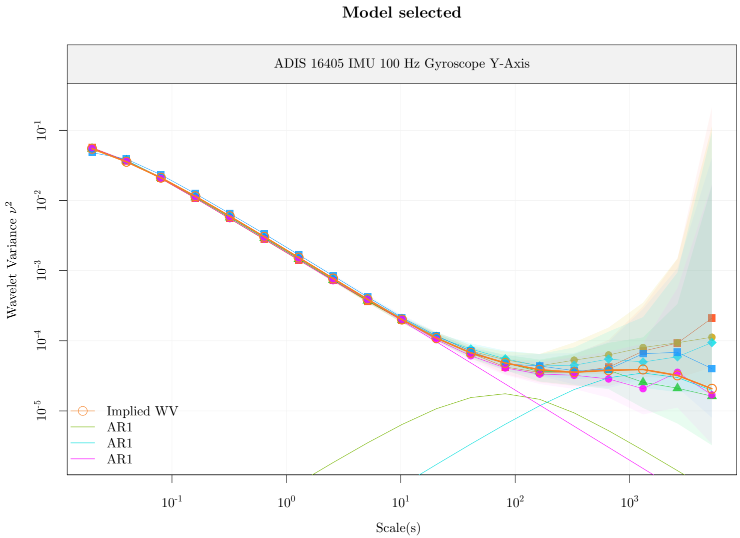
Case Study: The ADIS 16405 IMU



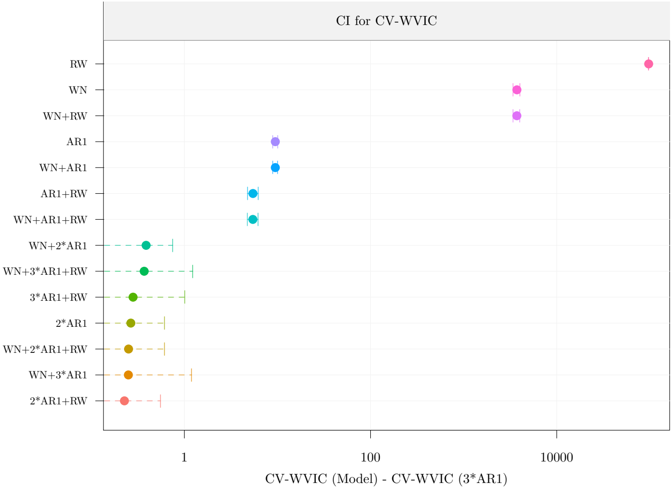
Case Study: The ADIS 16405 IMU



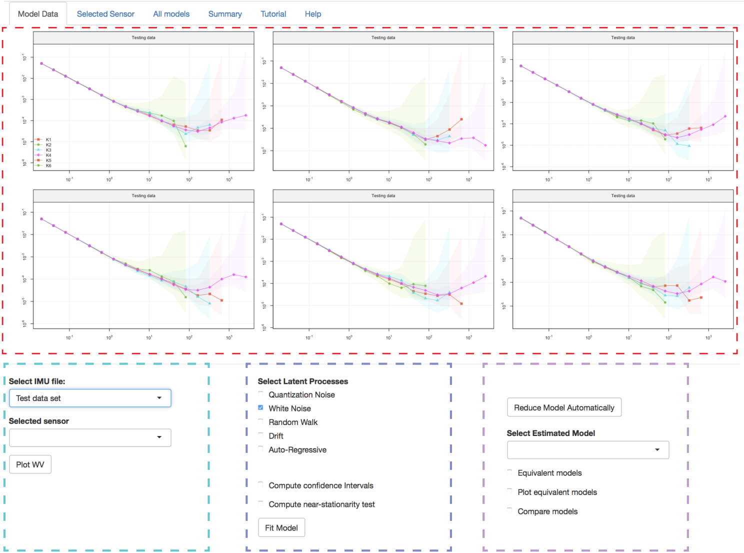
Case Study: The ADIS 16405 IMU



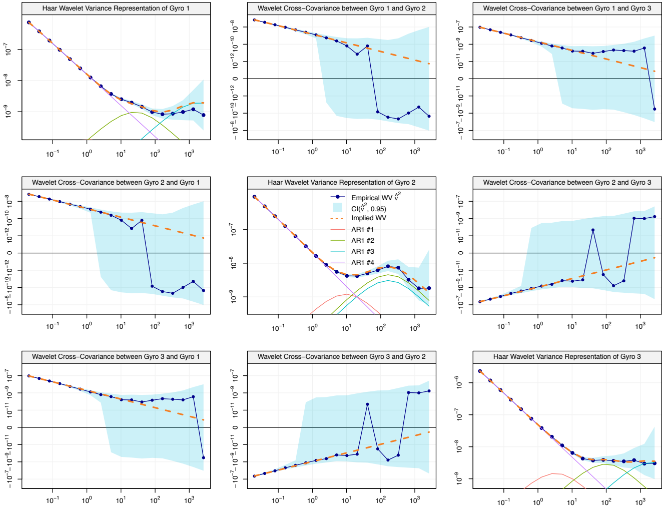
Case Study: The ADIS 16405 IMU



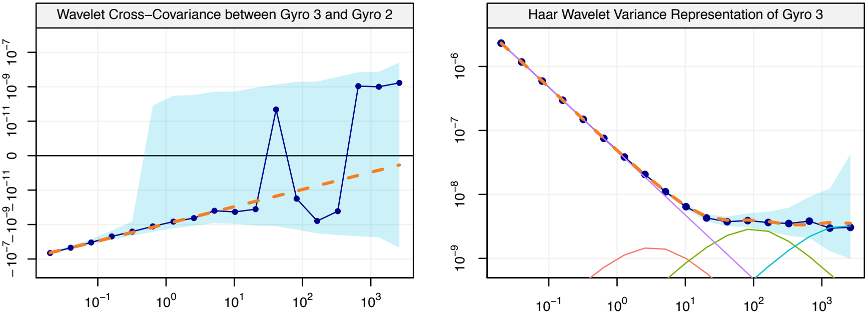
Web-based Platform: MGMWM



Modelling Dependence between Sensors

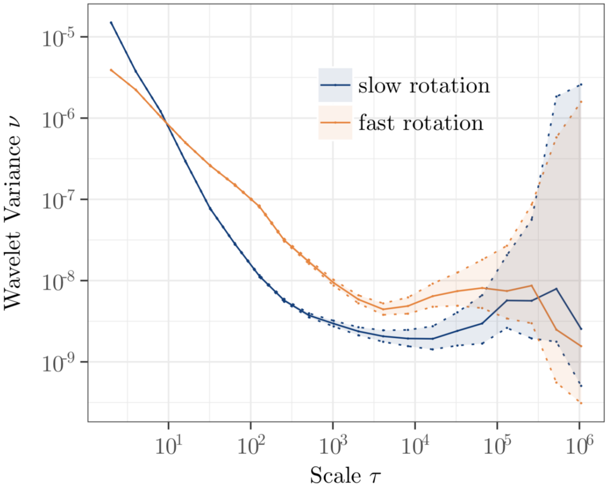


Modelling Dependence between Sensors

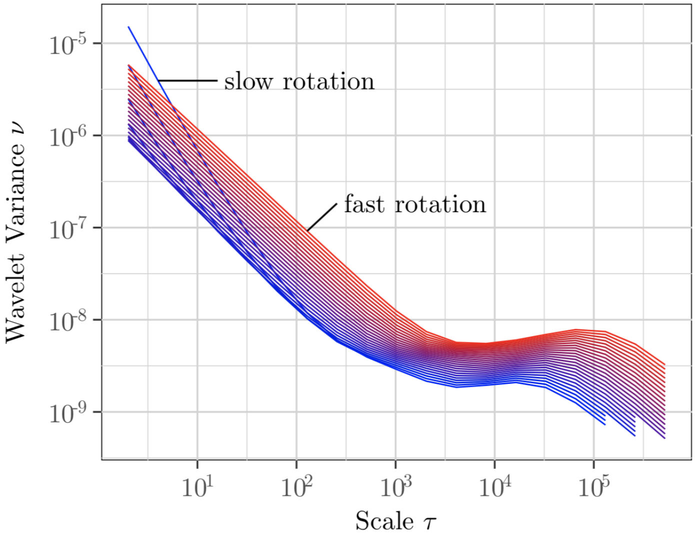


Dynamic Calibration

MEMS IMU Gyroscope rotating at $30\,°/s$ and $360\,°/s$



Dynamic Calibration



References

Guerrier, S., R. Molinari, and Y. Stebler (2016). "Theoretical Limitations of Allan Variance-based Regression for Time Series Model Estimation". In: *IEEE Signal Processing Letters* 23.5, pp. 597–601.

Guerrier, S. et al. (2013). "Wavelet-Variance-Based Estimation for Composite Stochastic Processes". In: *Journal of the American Statistical Association* 108.503.

Percival, D. B. and A. T. Walden (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Serroukh, A., A. T. Walden, and D. B. Percival (2000). "Statistical Properties and Uses of the Wavelet Variance Estimator for the Scale Analysis of Time Series". In: *Journal of the American Statistical Association* 95.449, pp. 184–196.

Stebler, Y. (2012). "Modeling and Processing Approaches for Integrated Inertial Navigation". PhD thesis. École Polytechnique Fédérale de Lausanne.

Stebler, Y. et al. (2011). "Constrained Expectation-Maximization Algorithm for Stochastic Inertial Error Modeling: Study of Feasibility". In: *Measurement Science and Technology* 22.8, p. 085204.

Outline

Introduction

Definition 2.1 (Time Series or TS).

A TS is a **stochastic process**,(i.e. a sequence of Random Variables (RV)), defined on a common probability space denoted as $(X_t)_{t=1,\ldots,T}$ (i.e. X_1, X_2, ..., X_T). Note that the time t is not continuous and belongs to discrete index sets. Therefore, we implicitly assume that:

- t is not random e.g. the time at which each observation is measured is known, and
- the time between two consecutive observations is constant.

Remark 1 (descriptive analysis).

In the classical time series theory, it is often useful to gain insight about a process by performing a descriptive analysis. While this approach may not be appropriate with inertial sensors, we shall briefly review it in the next few slides.

Introduction

Definition 2.2 (Descriptive Analysis).

Most time series analysis starts with displaying the data as a line plot on a graph. Time on the x-axis and variable on y-axis. Such graphs are often useful to assess various properties of the data at hand.

Time Series Graph/Plot:

- When recording values of the same variable over an extended period of time, it is difficult to discern any trend or pattern by simply looking at the values.
- However, when these data points are displayed on a plot (time on x-axis and X_t on y-axis), some features jump out.
- TS graph make trends easy to spot.
- These graphs more useful for small or moderate size data.

Descriptive Analysis

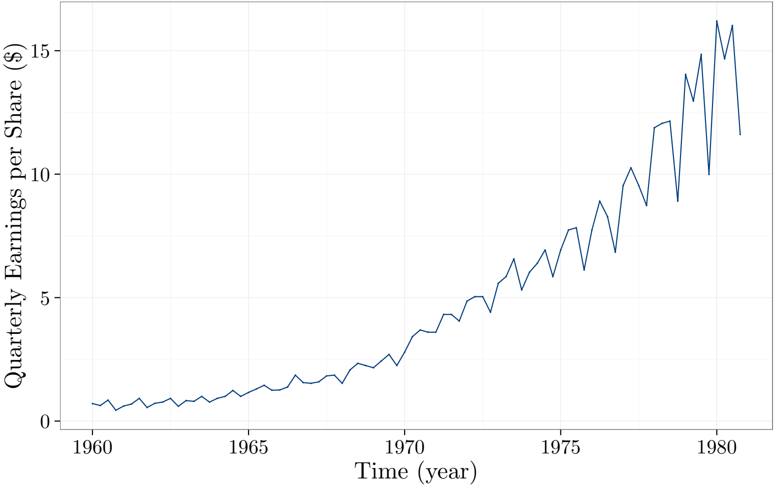
Question:

What do we want to check for in a time series data/graph?

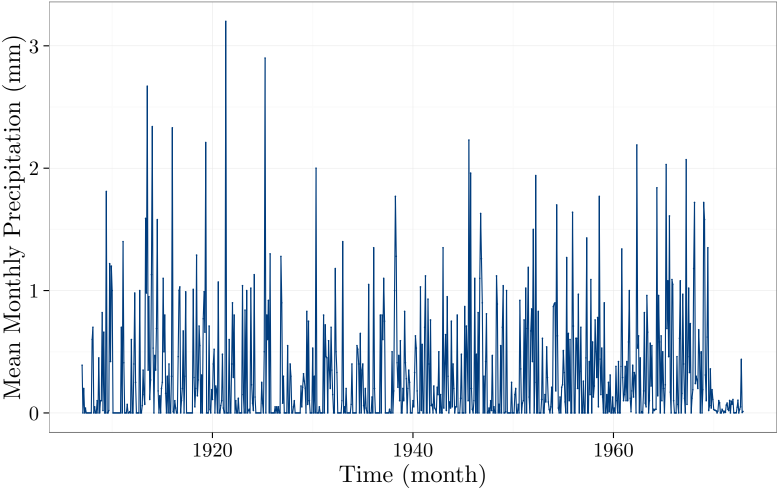
A possible answer:

- Trends:
  - Seasonal (e.g. business cycles)
  - Non-seasonal (e.g. impact of economic indicators on stock returns)
  - "Local" (e.g. vibrations observed before, during and after an earthquake)
- Changes in the **statistical properties**:
  - Mean (e.g. economic crisis)
  - Variance (e.g. earnings)
  - States (e.g bear/bull in finance)
- Model deviations (e.g. outliers)

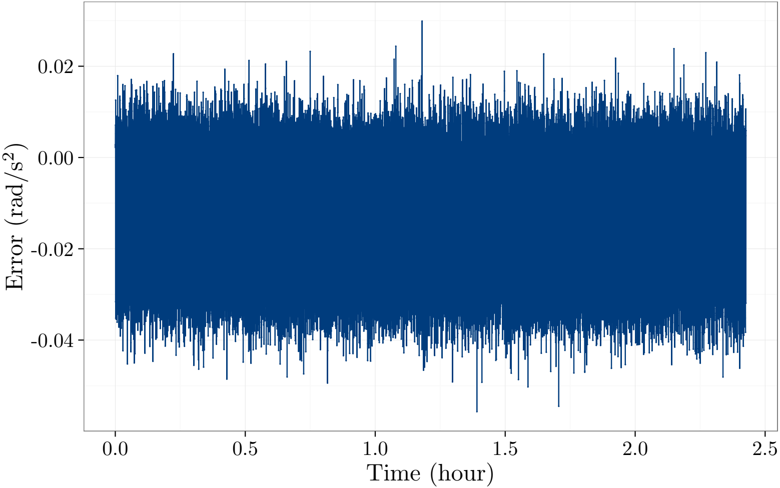
Example: Johnson and Johnson Quarterly Earnings



Example: Monthly Precipitation Data



Example: Inertial Sensor Data



Stochastic Processes Considered in this course

Definition 2.3 (Gaussian White Noise).

Gaussian White Noise (WN) with parameter $\sigma^2 \in \mathbb{R}^+$. This process is defined as

$$X_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

where "iid" stands for "independent and identically distributed".

Definition 2.4 (Quantization Noise).

Quantization Noise (QN) with parameter $Q^2 \in \mathbb{R}^+$. This process has a PSD of the form

$$S_X(f) = 4Q^2 \sin^2\left(\frac{\pi f}{\Delta t}\right) \Delta t, \;\; f < \frac{\Delta t}{2}.$$

Definition 2.5 (Drift).

Drift (DR) with parameter $\omega \in \Omega$ where Ω is either $\mathbb{R}^+$ or $\mathbb{R}^-$. This process is defined as $X_t = \omega t$.

Stochastic Processes Considered in this course

Definition 2.6 (Random walk).

Random walk (RW) with parameter $\gamma^2 \in \mathbb{R}^+$. This process is defined as

$$X_t = X_{t-1} + \epsilon_t \;\text{ where }\; \epsilon_t \overset{iid}{\sim} \mathcal{N}\left(0, \gamma^2\right) \;\text{ and }\; X_0 = 0.$$

Definition 2.7 (Auto-Regressive).

Auto-Regressive Process of Order 1 (AR1) process with parameter $\phi \in (-1, +1)$ and $\upsilon^2 \in \mathbb{R}^+$. This process is defined as

$$X_t = \phi X_{t-1} + Z_t, \quad Z_t \overset{iid}{\sim} \mathcal{N}(0, \upsilon^2).$$

Stochastic Processes Considered in this course

Definition 2.8 (Gauss Markov).

Gauss Markov Process of Order 1 (GM) process with parameter $\beta \in \mathbb{R}$ and $\sigma_G^2 \in \mathbb{R}^+$. This process is defined as

$$X_t = \exp(-\beta \Delta_t) X_{t-1} + Z_t, \quad Z_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_G^2(1 - \exp(-2\beta \Delta t)))$$

where Δt denotes the time between X_t and X_{t-1}.

Remark 2 (GM and AR1).

A GM process is a one-to-one reparametrization of an AR1 process. In this course, we shall only discuss AR1 processes but all results remain valid for GM processes.

Stochastic Processes Considered in this course

Definition 2.9 (Composite stochastic processes).

A composite stochastic process is a sum of latent processes. In this course, we will always assume that these latent processes are independent.

Example:

The composite stochastic process: "2*AR1 + WN" is given:

$$
\begin{aligned}
Y_t &= \phi_1 Y_{t-1} + Z_t, \quad Z_t \overset{iid}{\sim} \mathcal{N}(0, \upsilon_1^2) \\
W_t &= \phi_2 W_{t-1} + U_t, \quad U_t \overset{iid}{\sim} \mathcal{N}(0, \upsilon_2^2) \\
Q_t &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\
X_t &= Y_t + W_t + Q_t,
\end{aligned}
$$

where **only** (X_t) is observed.

Main purpose of TS analysis: Forecasting

- **Forecasting** is one of the main purpose of time series analysis. The question can be described as: if $(X_t)_{t=1,\ldots,T}$ is an identically distributed sequence but is *not independent*, what is the "best" predictor for X_{T+h} for $h > 0$ (i.e. an estimator of $\mathbb{E}[X_{T+h}|X_T, \ldots])$?
- *A simple answer is that it depends on the "dependence" between $X_1, \ldots, X_T$!*
- How could we measure this dependence?
- A first step is to extend the notation of covariance and correlation to time dependent sequences. We will refer to these notions as **autocovariance** and **autocorrelation**.
- The notion of autocovariance is an important one in time series analysis as it is closely linked to the concept of stationarity. Informally speaking, the latter creates a framework in which averages are "meaningful" (we will come back to this).

Review of Independence and Dependence

Definition 2.10 (Independence of two events).

Two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definition 2.11 (Independence of two random variables).

Two random variables X and Y with Cumulative Distribution Functions (CDF) $F_X(x)$ and $F_Y(y)$ (respectively) are **independent** if and only if their joint CDF $F_{X,Y}(x, y)$ is such that $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.

Definition 2.12 (iid sequence).

The sequence $X_1, X_2, \ldots, X_T$ is said to be **iid** if and only if

$$\mathbb{P}(X_i < x) = \mathbb{P}(X_j < x) \;\; \forall x \in \mathbb{R}, \forall i, j \in \{1, \ldots, T\}, \;\; \text{and}$$

$$\mathbb{P}(X_1 < x_1, X_2 < x_2, \ldots, X_T < x_T) = \mathbb{P}(X_1 < x_1) ... \mathbb{P}(X_T < x_T),$$

for any $T \geq 2$ and $x_1, ..., x_T \in \mathbb{R}$.

Measuring (linear) dependence

Dependence between T RV is difficult to measure at one shot! So we consider just two random variables, X_t and X_{t+h}. Then, one common (linear) measure of dependence is the covariance between X_t and X_{t+h}, which is defined below.

Definition 2.13 (AutoCovariance).

The covariance between X_t and X_{t+h}, defined as the *AutoCovariance* or simply ACV, is denoted using the function $\gamma_X(t, t+h)$, i.e.

$$\gamma_X(t, t+h) \equiv \text{Cov}(X_t, X_{t+h}) = \mathbb{E}(X_t X_{t+h}) - \mathbb{E}(X_t)\mathbb{E}(X_{t+h}),$$

where

$$\mathbb{E}(X_t) = \int_{-\infty}^{\infty} x\, f(x) dx$$

$$\mathbb{E}(X_t, X_{t+h}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2) dx_1 dx_2,$$

where $f(x_1, x_2)$ denotes the joint density of X_t and X_{t+h}.

Measuring (linear) dependence

Remark 3 (Scale dependence).

Just as any covariance, the $\gamma_X(t, t+h)$ is "scale dependent" and therefore $\gamma_X(t, t+h) \in \mathbb{R}$.

- If $\mid \gamma_X(t, t+h) \mid$ is "close" to 0, then they are "less dependent".
- If $\mid \gamma_X(t, t+h) \mid$ is "far" from 0, X_t and X_{t+h} are "more dependent".

Remark 4 (ACV and independence).

$\gamma_X(t, t+h) = 0$ does not imply X_t and X_{t+h} are independent. However, if X_t and X_{t+h} are joint normally distributed then $\gamma_X(t, t+h) = 0$ implies that X_t and X_{t+h} are independent.

Measuring (linear) dependence

A measure of dependence related to the ACV is the autocorreation. This is arguably the most commonly used metric in time series analysis.

Definition 2.14 (Autocorrelation).

The correlation between X_t and X_{t+h} is defined as the *autocorrelation* or simply ACF and is denoted using the function $\rho_X(t, t+h)$, i.e.

$$\rho_X(t, t+h) = \text{corr}(X_t, X_{t+h}) = \frac{\text{cov}(X_t, X_{t+h})}{\sqrt{\text{var}(X_t)}\sqrt{\text{var}(X_{t+h})}}$$

Remark 5 (Scale invariance).

Just as any correlation, $\rho_X(t, t+h)$ is scale free. Moreover, if $\rho_X(t, t+h)$ is "close" to ± 1 then this implies that there is "strong" (linear) dependence between X_t and X_{t+h}.

Measuring (linear) dependence

Remark 6 (Notation).

The notation $\gamma_X(t, t+h)$ and $\rho_X(t, t+h)$ is often simplified to $\gamma(t, t+h)$ and $\rho(t, t+h)$ when not ambiguous (i.e. only one time series is considered).

Remark 7 (Linear dependence and real dependence).

Covariance and correlation measure linear dependence. They are less helpful to measure monotonic dependence and they are much less helpful to measure nonlinear dependence. Nonlinear measures of dependence exist but we will not discuss this subject in this class. Here is an example:

Remark 8 (Causation).

Correlation *does NOT* imply causation. For example, if $\rho(t, t+h) \neq 0$ it does not imply that $X_t \to X_{t+h}$ is causal. Actually, real causation doesn't exist in Statistics but there exist approximated metric to measure this concept such as Granger causality (see Granger 1969). This idea is clearly illustrated in the image below.

Estimation of in the context of time series

Motivation:

Consider the simple (but strange!) model:

$$X_t \sim \mathcal{N}\left(0, Y_t^2\right) \;\; \text{where } Y_t \text{ is unobserved and such that } \; Y_t \overset{iid}{\sim} \mathcal{N}\left(0, 1\right).$$

In this case, it is clear that the estimation of $\text{var}(X_t)$ is difficult (in fact X_t^2 is your best guess!) since only X_t is useful to estimate $\text{var}(X_t)$. This process is an example of a **non-stationary process** (we will see why in the next slides). On the other hand, if we consider the **stationarity process** such as:

$$X_t = \theta W_{t-1} + W_t \;\; \text{where} \;\; W_t \overset{iid}{\sim} \mathcal{N}\left(0, 1\right).$$

Then, one can guess that a natural estimator of $\text{var}(X_t)$ is simply $\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T} X_i^2$, because our hope is that averages are "**meaningful**" for such processes. In the next slides, we will formalize this idea through the concepts of stationarity.

Strong and Weak Stationarity

There exist two forms of stationarity, which are defined below:

Definition 2.15 (Strong Stationarity).

The joint probability distribution of $(X_t)_{t\in\mathbb{N}}$ is invariant under a shift in time, i.e.

$$\mathbb{P}(X_t \leq x_1, \ldots, X_{t+k} \leq x_k) = \mathbb{P}(X_{t+h} \leq x_1, \ldots, X_{t+h+k} \leq x_k)$$

for any time shift h and any $x_1, x_2, \cdots, x_k$ belong to the domain of $X_t, \cdots, X_{t+k}$ and $X_{t+h}, \cdots, X_{t+h+k}$.

Definition 2.16 (Weak Stationarity).

The mean and autocovariance of the stochastic process are finite and invariant under a shift in time, i.e.

$$\mathbb{E}\left[X_t\right] = \mu < \infty,$$
$$\mathbb{E}\left[X_t^2\right] = \mu_2 < \infty,$$
$$\text{cov}(X_t, X_{t+h}) = \text{cov}(X_{t+k}, X_{t+h+k}) = \gamma(h).$$

Strong and Weak Stationarity

Why does stationarity matter?

The stationarity of X_t is important because it provides a framework in which averaging makes sense. Unless properties like mean and covariance are either fixed or "evolve" in a known manner, the concept of averaging is essentially meaningless.

Remark 9 (Implication on the ACV and ACF).

If a process is weakly stationary or strongly stationary and $\text{cov}(X_t, X_{t+h})$ exist for all $h \in \mathbb{Z}$. Then we have the ACV and ACF only depends on the lag between observations, i.e.

$$\gamma(t, t+h) = \text{cov}(X_t, X_{t+h}) = \text{cov}(X_{t+k}, X_{t+h+k}) = \gamma(t+k, t+h+k) = \gamma(h),$$
$$\rho(t, t+h) = \text{corr}(X_t, X_{t+h}) = \text{corr}(X_{t+k}, X_{t+h+k}) = \rho(t+k, t+h+k) = \rho(h).$$

Strong and Weak Stationarity

Remark 10 (Properties of the ACV and ACF).

Remark 9 implies that the ACV and ACF have the following properties:

- $\gamma(0) = \text{var}\left[X_t\right] \geq 0$ and $\rho(0) = 1$.
- $\gamma(h) = \gamma(-h)$ and $\rho(h) = \rho(-h)$ (therefore they are both even functions).
- $|\gamma(h)| \leq \gamma(0)$ and $|\rho(h)| \leq 1$ for all $h \in \mathbb{Z}$.

The first two properties are direct for the properties of the covariance and correlation (i.e. $\text{cov}(X, X) = \text{var}(X)$ and $\text{cov}(X, Y) = \text{cov}(Y, X)$). However, the third property is less obvious and is a consequence of the Cauchy-Schwarz inequality, i.e.

$$\mathbb{E}^2\left[XY\right] \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]. \tag{6}$$

Using (6), we have

$$\begin{aligned}\gamma(h)^2 &= \left(\text{cov}\left(X_t, X_{t+h}\right)\right)^2 = \left(\mathbb{E}\left[\left(X_t - \mu\right)\left(X_{t+h} - \mu\right)\right]\right)^2 \\ &\leq \mathbb{E}\left[\left(X_t - \mu\right)^2\right]\mathbb{E}\left[\left(X_{t+h} - \mu\right)^2\right] = \gamma(0)^2,\end{aligned}$$

which verifies the last properties.

Remarks: Strong and Weak Stationarity

Remark 11.

Neither type of stationarity implies the other one. This is being illustrated in the two examples presented in Appendix A ▸ Go to Appendix A. Note however that if X_t is Normal (Gaussian) with $\sigma^2 = \text{var}(X_t) < \infty$, then weak stationarity implies strong stationarity.

Remark 12.

From the definition of (weak) stationarity, it is easy to see that a WN or a QN processes is stationary while a RW process is not. The stationarity of an AR1 (or GM) is less obvious. In fact, an AR1 is stationary if $|\phi| < 1$. The derivation of this property is given in Appendix B ▸ Go to Appendix B.

Linear Processes

Definition 2.17 (Linear Processes).

A stochastic process (X_t) is said to be a linear process if it can be expressed as a linear combination of an iid sequence (which here is Gaussian for convenience), i.e.:

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

where $W_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Remark 13 (Properties of linear processes).

All linear processes are stationary and such that:

$$\mathbb{E}[X_t] = \mu,$$

$$\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{h+j}.$$

Linear Processes

Remark 14 (Convergence of linear processes).

The latter condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ is required to ensure that the series has a limit and is related to the absolutely summable covariance structure (defined below).

Definition 2.18 (Absolutely summable covariance structure).

A process (X_t) is said to have an absolutely summable covariance structure if $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$.

Remark 15 (All linear process have an absolutely summable covariance structure).

Interestingly, the condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ is actually stronger than $\sum_{h=-\infty}^{\infty} |\gamma_X(h)| < \infty$. Indeed, we have that

$$\sum_{h=-\infty}^{\infty} |\gamma_X(h)| \leq 2\gamma(0) \left(\sum_{j=-\infty}^{\infty} |\psi_j| \right)^2 < \infty.$$

A Fundamental Representation

A Fundamental Representation

Autocovariances and autocorrelations also turn out to be very useful tools as they are one of the *fundamental representations of time series*.

If we consider a zero mean normally distributed process, it is clear that its joint distribution is fully characterized by the autocovariances $\mathbb{E}[X_t X_{t+h}]$ since the joint probability density only depends on these covariances.

Once we know the autocovariances we know everything there is to know about the process and therefore: **if two processes have the same autocovariance function, then they are the same process**.

Another Fundamental Representation

Fundamental Representation: the Power Spectral Density

For the same processes considered in the previous slide, another fundamental representation of a time series is given by the Power Spectral Density (PSD) which can be defined as

$$S_X(f) = \int_{-\infty}^{\infty} \gamma_X(h)\, e^{-ifh} dh,$$

where f is a frequency. Hence, the PSD is a Fourier transform of the autocovariance function which describes the variance of a time series over frequencies (with respect to lags h).

Given that the definition of the PSD, as for the autocovariance function, once we know the PSD we know everything there is to know about the process and therefore: **if two processes have the same PSD, then they are the same process**.

Estimation Problems with Dependent Data

Estimation in the context of time series is not as straightforward as in the iid case. In order to "warm-up", let us start with the easiest case: the empirical mean.

Let (X_t) be a stationary time series, therefore we have that $\mu_t = \mathbb{E}[X_t] = \mu$ and the value of μ can be estimated by the sample mean, i.e.

$$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t.$$

Using the properties of stationary process we have that:

$$\text{var}\left(\bar{X}\right) = \frac{1}{T^2}\,\text{cov}\left(\sum_{t=1}^{T} X_t, \sum_{s=1}^{T} X_s\right) = \frac{\gamma(0)}{T}\sum_{h=-T}^{T}\left(1 - \frac{|h|}{T}\right)\rho(h). \tag{7}$$

The derivation of (7) is instructive and is given in Appendix C ▸ Go to Appendix C. Moreover, some simulation-based and analytical methods to estimate var $\left(\bar{X}\right)$ are discussed in Appendix D ▸ Go to Appendix D.

Estimation of $\gamma(h)$ and $\rho(h)$

We define here the "classical" estimator of $\gamma(h)$ and $\rho(h)$ as the sample autocovariance and autocorrelation functions. In the following section, we shall study the properties of these estimators:

Definition 2.19 (Sample autocovariance function).

The sample autocovariance function is defined as

$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}\left(X_t - \bar{X}\right)\left(X_{t+h} - \bar{X}\right)$$

with $\hat{\gamma}(h) = \hat{\gamma}(-h)$ for $h = 0, 1, ..., k$, where k is a fixed integer.

Definition 2.20 (Sample autocorrelation function).

The sample autocorrelation function is defined as $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$, with $\hat{\rho}(h) = \hat{\rho}(-h)$ for $h = 0, 1, ..., k$, where k is a fixed integer.

Outline

Statistical Estimators

In this section, we shall present an introduction to the study of the properties of statistical estimators. We will mainly focus on asymptotic properties and start by consider a general class of estimators.

Definition 3.1 (Extremum Estimator).

Many estimators have a **common structure**, which is often useful to study their asymptotic properties. One structure or framework is the class of estimators that maximize some objective function, referred to as extremum estimators, which can can be defined as follows:

$$\hat{\boldsymbol{\theta}} \equiv \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; \hat{Q}_n(\boldsymbol{\theta}), \tag{8}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ denote, respectively, the parameter vector of interest and its set of possible values.

Remark:

The vast majority of statistical estimators can be represented as extremum estimator. This is for example for the case for least squares, maximum likelihood or (generalized) method of moment estimators.

Example: Least Squares Estimator

Consider the linear model $\boldsymbol{y} = \mathbf{X}\beta_0 + \boldsymbol{\varepsilon}$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a full-ranked constant matrix, $[\boldsymbol{\varepsilon}]_i \overset{iid}{\sim} (0, \sigma_\varepsilon^2)$ and $\beta \in \mathcal{B} \subseteq \mathbb{R}^p$. Let $\hat{\beta}$ denote the Least Squares Estimator (LSE) of β_0, i.e.

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}.$$

Moreover, this estimator is an extremum estimator since it can expressed as:

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\text{argmax}} \; -||\boldsymbol{y} - \mathbf{X}\beta||_2^2,$$

similarly to our definition given in (8).

Example: Maximum Likelihood Estimator

Let $Z_1, \ldots, Z_n$ be an iid sample with pdf $f(z|\boldsymbol{\theta}_0)$. The Maximum Likelihood Estimator (MLE) is given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; \frac{1}{n} \sum_{i=1}^{n} \log \left[f\left(z_i | \boldsymbol{\theta} \right) \right]. \tag{9}$$

Therefore, the MLE can be seen as an example of extremum estimator with

$$\hat{Q}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log \left[f\left(z_i | \boldsymbol{\theta} \right) \right].$$

Remark: In (9) we are actually using a *normalized* log-likelihood instead of the actual log-likelihood. This has (in the vast majority of cases) no impact on the estimator but the normalized form is more convenient to use when we let $n \to \infty$.

Example: Generalized Method of Moment

Consider the same iid sample as in the previous example and suppose that there is a "moment function" vector $\mathbf{g}(z|\boldsymbol{\theta})$ such that $\mathbb{E}[\mathbf{g}(z|\boldsymbol{\theta}_0)] = 0$. Then, a possible estimator of $\boldsymbol{\theta}_0$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; -\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(z_i|\boldsymbol{\theta})\right]^T \widehat{\mathbf{W}} \left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(z_i|\boldsymbol{\theta})\right], \tag{10}$$

where $\widehat{\mathbf{W}}$ is an positive definite matrix of appropriate dimension. Such estimators are called Generalized Method of Moments (GMM) estimators. They belong to the class of extremum estimators.

Remark: Instead of (10) we will often consider an alternative (but equivalent) definition of such estimator, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; -||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}},$$

where $||\mathbf{x}||^2_{\mathbf{A}} = \mathbf{x}^T\mathbf{A}\mathbf{x}$, and where $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$ denote, respectively, the empirical and model based moments.

Example: A simple GMM Estimator

Let $Z_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$ and $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0^2)^T$. Suppose we wish to estimate $\boldsymbol{\theta}_0$ by matching the first three empirical moments with their theoretical counterparts. In this case, a reasonable moment function or condition defining a GMM estimator is given by:

$$\mathbf{g}(Z|\boldsymbol{\theta}) = \begin{bmatrix} Z - \mu \\ Z^2 - \left(\mu^2 + \sigma^2\right) \\ Z^3 - \left(\mu^3 + 3\mu\sigma^2\right) \end{bmatrix}.$$

However, it can be noticed that $\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(Z_i|\boldsymbol{\theta}) = \hat{\gamma} - \gamma(\boldsymbol{\theta})$, where $\hat{\gamma}$ and $\gamma(\boldsymbol{\theta})$ denote, respectively, the empirical and model-based moments, i.e.

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} Z_i \\ Z_i^2 \\ Z_i^3 \end{bmatrix}, \quad \gamma(\boldsymbol{\theta}) = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \\ \mu^3 + 3\mu\sigma^2 \end{bmatrix}.$$

Then, we can write our GMM estimator of $\boldsymbol{\theta}_0$ as:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \; ||\hat{\gamma} - \gamma(\boldsymbol{\theta})||_{\widehat{\mathbf{W}}}^2 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; -||\hat{\gamma} - \gamma(\boldsymbol{\theta})||_{\widehat{\mathbf{W}}}^2. \tag{11}$$

Consistency of Statistical Estimators

In the next slides we will discuss the conditions for the consistency of extremum estimator, which is often denoted as $\hat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$. We start by defining consistency.

Definition: Consistency

The estimator $\hat{\boldsymbol{\theta}}$ is said to be consistent if it converges in probability to $\boldsymbol{\theta}_0$, i.e.

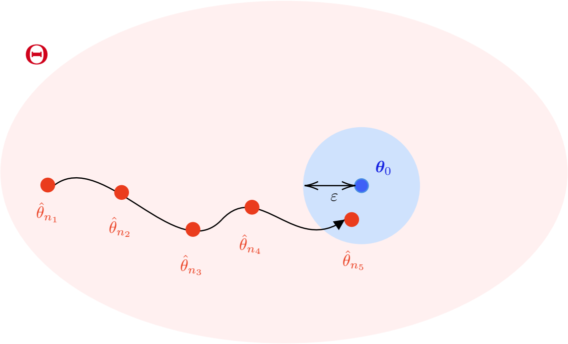
$$\lim_{n\to\infty} \Pr\left(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon\right) = 0,$$

for all $\varepsilon > 0$.

Consistency - Interpretation

Interpretation:

In Layman's term consistency simply means that if n is "large enough" $\hat{\boldsymbol{\theta}}$ will be **arbitrarily close to θ_0** (i.e. inside of an hypersphere of radius ε centered at $\boldsymbol{\theta}_0$). This also means the procedure (i.e. our estimator) based on unlimited data will be able to identify the underlying truth (i.e. $\boldsymbol{\theta}_0$).



Related Theorems

Consistency is often proven using the following two important results.

Theorem 3.2 (Weak Law of Large Number).

Suppose X_i are iid random variables with finite mean μ (i.e. $\mathbb{E}[X_i] = \mu$) and finite variance. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$.

Theorem 3.3 (Continuous Mapping Theorem).

Suppose $Y_n \xrightarrow{\mathcal{P}} \mu$, then $g(Y_n) \xrightarrow{\mathcal{P}} g(\mu)$ if $g(\cdot)$ is a continuous function.

A simple example on the consistency is presented in Appendix E Go to Appendix E.

Consistency of Extremum Estimators

When considering real-life problems the approach based on Theorem 3.2 presented in Appendix E is in general not flexible enough and we generally rely on the results such the one presented below.

Theorem 3.4 (Consistency of Extremum Estimators).

If there is a function $Q_0(\boldsymbol{\theta})$ such that:

(C.1) $Q_0(\boldsymbol{\theta})$ *is uniquely maximized in* $\boldsymbol{\theta}_0$,

(C.2) $\boldsymbol{\Theta}$ *is compact*[a],

(C.3) $Q_0(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$,

(C.4) $\hat{Q}_n(\boldsymbol{\theta})$ *converges uniformly in probability to* $Q_0(\boldsymbol{\theta})$[b].

then we have $\hat{\boldsymbol{\theta}} \xmapsto{\mathcal{P}} \boldsymbol{\theta}$.

---

[a] Compact means that $\boldsymbol{\Theta}$ is both closed (i.e. containing all its limit points) and bounded (i.e. all its points are within some fixed distance of each other).

[b] $\hat{Q}_n(\boldsymbol{\theta})$ is said to converges uniformly in probability to $Q_0(\boldsymbol{\theta})$ if $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xmapsto{\mathcal{P}} 0$

Remarks: Consistency of Extremum Estimators

This theorem is an important result, which provides a general approach to prove the consistency of a large class of estimators. A few remarks on the conditions of this result:

- Condition (C.1) is substantive and there are well-known examples where it fails. We will discuss further on how this assumption can (in some cases) be verified in practice.
- Condition (C.2) is also substantive as it requires that there exist some known bounds on the parameters. In practice, this assumption is often neglected although it is in most cases unrealistic to assume it.
- Conditions (C.3) and (C.4) are often referred to as "standard regularity conditions". They are typically satisfied. The verification of these conditions will be discussed further in this document.

Theorem 3.4: Sketch of the proof

The basic idea of the proof is the following. Under Condition (C.1) we have

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; Q_0(\boldsymbol{\theta}).$$

Condition (C.4) implies that

$$\hat{Q}_n(\boldsymbol{\theta}) \xmapsto{\mathcal{P}} Q_0(\boldsymbol{\theta}),$$

therefore, it seems logical that

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; \hat{Q}_n(\boldsymbol{\theta}) \xmapsto{\mathcal{P}} \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \; Q_0(\boldsymbol{\theta}) = \boldsymbol{\theta}_0.$$

Unfortunately, without additional conditions this simple proof is not correct. The main reason is that additional conditions (i.e. uniform convergence as well as Conditions (C.2) and (C.3)) are needed to ensure the validity of the above convergence given in orange. A formal proof of this results is given in Appendix F and is of course good to know! ▸ Go to Theorem Appendix F

Verification of Condition (C.1)

In general the verification of Condition (C.1) is difficult and is often assumed in the statistical literature. We present here two results, namely Lemma 3.5 from Newey and McFadden 1994 and Theorem 2 from Komunjer 2012, which allow the verification of this condition for GMM-type estimator (on which we shall focus here). A discussion on the verification of this condition for other estimators can for example be found in Chapter 7 of Baltagi 2008. The proof of Lemma 3.5 is given in Appendix G ▸ Go to Appendix Appendix G.

Lemma 3.5 (Identification GMM).

If $\mathbf{W} > 0$ *(i.e. positive definite) (where* $\mathbf{W}$ *is such that* $\widehat{\mathbf{W}} \xmapsto{\mathcal{P}} \mathbf{W}$*),* $\mathbf{g}_0(\boldsymbol{\theta}) = \mathbb{E}[\mathbf{g}(z|\boldsymbol{\theta})]$, $\mathbf{g}_0(\boldsymbol{\theta}_0) = \mathbf{0}$ *and* $\mathbf{g}_0(\boldsymbol{\theta}) \neq \mathbf{0}$ *if* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ *then* $Q_0(\boldsymbol{\theta}) = -\mathbf{g}_0(\boldsymbol{\theta})^T \mathbf{W} \mathbf{g}_0(\boldsymbol{\theta})$ *has a unique maximum at* $\boldsymbol{\theta}_0$*.*

Remark:

If we write a GMM estimator as in (11) the condition $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ can be replaced by $\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Verification of Condition (C.1)

Therefore, Lemma 3.5 shows that a GMM estimator is Condition (C.1) in the case where $\mathbf{g}_0(\boldsymbol{\theta}) = \mathbf{0}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (or alternatively $\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$). The following theorem (which is a "simplified" version for Theorem 2 of Komunjer 2012) provides us with a way to verify this new condition.

Theorem 3.6 (Homeomorphism).

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$. Let $\mathbf{g}^*(\boldsymbol{\theta})$ denote a subset of p elements of $\mathbf{g}_0(\boldsymbol{\theta}) \in \mathbb{R}^q$, $q \geq p$ such that:

- $\mathbf{g}^*(\boldsymbol{\theta})$ *is in $\mathcal{C}^2$ (i.e. $\mathbf{g}^*(\boldsymbol{\theta})$ can be differentiated twice).*
- *For every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $J(\boldsymbol{\theta})$ is nonnegative (or alternatively nonpositive), where $J(\boldsymbol{\theta}) \equiv \det\left(\frac{\partial}{\partial \boldsymbol{\theta}^T}\mathbf{g}^*(\boldsymbol{\theta})\right)$.*
- $||\mathbf{g}^*(\boldsymbol{\theta})|| \to \infty$ *whenever* $||\boldsymbol{\theta}|| \to \infty$.
- *For every $s \in \mathbb{R}^p$ the equation $\mathbf{g}^*(\boldsymbol{\theta}) = s$ has countably many (possibly zero) solutions in $\boldsymbol{\Theta}$.*

Then, $\mathbf{g}^*(\boldsymbol{\theta})$ is a Homeomorphism (i.e. $\mathbf{g}^*(\boldsymbol{\theta})$ is continuous and one-to-one).

Discussion on Theorem 3.6

Remarks:

- A direct consequence of Lemma 3.5 and Theorem 3.6 is that any GMM estimator with $\mathbf{W} > 0$ satisfying the conditions of Theorem 3.6 satisfies Condition (C.1) of Theorem 3.4.
- In addition, if one can show that $\mathbf{g}^c(\boldsymbol{\theta})$ is in $\mathcal{C}$ (where $\mathbf{g}^c(\boldsymbol{\theta})$ denotes the element of $\mathbf{g}_0(\boldsymbol{\theta})$ that are not in $\mathbf{g}^*(\boldsymbol{\theta})$) then Condition (C.3) of Theorem 3.4 is also verified.
- Note that when considering a GMM estimator of the form used in (11), one can simply verify the conditions of Theorem 3.4 with $\mathbf{g}(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta})$.
- In Theorem 2 in Komunjer 2012 it is actually assumed that $\boldsymbol{\theta} \in \mathbb{R}^p$ while we assume that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We used this simplification to avoid an overly technical treatment of this topic. In fact, we assume here that there exist a one-to-one function $h(\cdot)$ such that $h : \mathbb{R}^p \mapsto \boldsymbol{\Theta}$. This condition is typically verified in practice.

Example: Proving Condition (C.1)

In this example we revisit the example presented in ▸ Example GMM. Let us say that $\widehat{\mathbf{W}}$ is such that $\widehat{\mathbf{W}} \overset{\mathcal{P}}{\longmapsto} \mathbf{W} > 0$. Then, we showed that:

$$\gamma(\boldsymbol{\theta}) = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \\ \mu^3 + 3\mu\sigma^2 \end{bmatrix}.$$

Therefore, we define

$$\mathbf{g}^*(\boldsymbol{\theta}) = \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix} \quad \text{and} \quad g^c(\boldsymbol{\theta}) = \left[\mu^3 + 3\mu\sigma^2\right].$$

Since the elements of $\mathbf{g}^*(\boldsymbol{\theta})$ are polynomial in $\boldsymbol{\theta}$, the condition $\mathbf{g}^*(\boldsymbol{\theta}) \in \mathcal{C}^2$ is trivially satisfied. Next, we define

$$\mathbf{A}(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \boldsymbol{\theta}^T} \gamma(\boldsymbol{\theta}) \in \mathbb{R}^{3\times 2} \quad \text{and} \quad \widetilde{\mathbf{A}}(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \boldsymbol{\theta}^T} \mathbf{g}^*(\boldsymbol{\theta}) \in \mathbb{R}^{2\times 2}.$$

Example: Proving Condition (C.1)

Then, the determinant of $\widetilde{\mathbf{A}}(\boldsymbol{\theta})$ is equal to 1 as illustrated in the diagram below:

$$\mathbf{A}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}^T} \boldsymbol{\gamma}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & 0 \\ 2\mu & 1 \\ 3\left(\mu^2 + \sigma^2\right) & 3\mu \end{bmatrix}, \quad \det\left(\boxed{\widetilde{\mathbf{A}}(\boldsymbol{\theta})}\right) = 1.$$

Finally, the last two conditions of 3.6 are trivially satisfied since $||\mathbf{g}^*(\boldsymbol{\theta})||$ can only diverge if $||\boldsymbol{\theta}||$ diverges and since $\mathbf{g}^*(\boldsymbol{\theta}) = s$ has (one) countably solutions in $\boldsymbol{\Theta}$.

Then, it follows from Lemma 3.5 and Theorem 3.6 that the function $Q_0(\boldsymbol{\theta}) = -||\boldsymbol{\gamma}(\boldsymbol{\theta}_0) - \boldsymbol{\gamma}(\boldsymbol{\theta})||^2_{\mathbf{W}}$ is uniquely maximized in $\boldsymbol{\theta}_0$. ●

Verification of Condition (C.4)

In general, establishing the uniform convergence of $\hat{Q}_n(\boldsymbol{\theta})$ to $Q_0(\boldsymbol{\theta})$, i.e.

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \xmapsto{\mathcal{P}} 0, \tag{12}$$

is not easy. A common strategy is the following:

- Show that $\hat{Q}_n(\boldsymbol{\theta}) \xmapsto{\mathcal{P}} Q_0(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (i.e. pointwise convergence).
- Show that $\hat{Q}_n(\boldsymbol{\theta})$ is almost surely Lipschitz continuous, i.e.

$$\sup_{\boldsymbol{\theta}_1,\, \boldsymbol{\theta}_2\in\boldsymbol{\Theta}} |\hat{Q}_n(\boldsymbol{\theta}_1) - \hat{Q}_n(\boldsymbol{\theta}_2)| \leq H||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||,$$

  where H is random variable which is almost surely bounded, i.e. there exist a constant c such that $|H| < c$ almost surely.

Verification of Condition (C.4)

Then, the following theorem (which is a slightly adapted version of the Arzela-Ascoli Theorem) allows the verification of Condition (C.4).

Theorem 3.7 (modified Arzela-Ascoli).

Suppose that $\boldsymbol{\Theta}$ is compact, for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we have $\hat{Q}_n(\boldsymbol{\theta}) \overset{\mathcal{P}}{\longmapsto} Q_0(\boldsymbol{\theta})$ and $\hat{Q}_n(\boldsymbol{\theta})$ is almost surely Lipschitz continuous. Then, we have $\hat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$.

Remark:

In this course, we will not discuss how to prove that $\hat{Q}_n(\boldsymbol{\theta})$ is almost surely Lipschitz continuous and assume it for simplicity (more discussion on this topic can for example be found in Newey and McFadden 1994). Nevertheless, it worth mentioning that this condition is almost always satisfied in practice and is therefore "reasonable" to assume for simplicity.

In Appendix H we provide a practice example on how to prove (C.4) ▸ Go to Appendix H.

Law of Large Number for Dependent Process

In general showing that $\hat{Q}_n(\boldsymbol{\theta}) \overset{\mathcal{P}}{\longmapsto} Q_0(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is generally done using one version of the law of large number (see e.g. Appendix H).

However, when X_i which are not iid random variables, Theorem 3.2 cannot be applied. The following theorem (taken from Proposition 7.5 of Hamilton 1994) generalizes this result for (weak) stationary processes with absolutely summable covariance structure (see Definition 2.18). A simple version of this proof based on Chebychev's inequality is given in Appendix I ▸ Go to Appendix I.

Theorem 3.8 (Weak Law of Large Number for Dependent Process).

Suppose (X_t) is a (weak) stationary process with absolutely summable autocovariance structure, then

$$\bar{X}_T \overset{\mathcal{P}}{\longmapsto} \mathbb{E}[X_t].$$

Example: Consistency of the Sample Mean

Consider a stationary AR1 process and suppose we wish to study whether its sample mean converges in probability to its expected value. This time we consider consider a non-zero mean AR1, i.e.

$$(X_t - \mu) = \phi\,(X_{t-1} - \mu) + Z_t,$$

where $\mu = \mathbb{E}[X_t]$, $Z_t \overset{iid}{\sim} \mathcal{N}(0, \nu^2)$ and $\nu^2 < \infty$. This process can also be written as a linear process (see Definition 2.17):

$$X_i - \mu = \sum_{k=0}^{\infty} \phi^k Z_{i-k},$$

Since the process is stationary for $|\phi| < 1$ (see Appendix B), we have

$$\gamma_h = \frac{\nu^2 \phi^{|h|}}{1 - \phi^2},$$

for $h \in \mathbb{Z}$.

Example: Consistency of the Sample Mean

Then, we have that

$$\sum_{h=-\infty}^{\infty} |\gamma_k| = \sum_{h=-\infty}^{\infty} \frac{\nu^2 |\phi|^{|h|}}{1-\phi^2} < 2 \lim_{n\to\infty} \frac{\nu^2(1-|\phi|^{n+1})}{(1-\phi^2)(1-|\phi|)} = \frac{2\nu^2}{(1-\phi^2)(1-|\phi|)} < \infty,$$

implying that the process has an absolutely summable covariance structure (see Definition 2.18). Therefore by applying Theorem 3.8 we can verify that

$$\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t \xmapsto{\mathcal{P}} \mathbb{E}[X_t] = \mu.$$

Consistency of $\hat{\gamma}(h)$ and $\hat{\rho}(h)$

In Definitions 2.19 and 2.20 we defined the sample autocovariance and autocorrelation functions. In the following corollary we show that these estimators are both consistent. The proof of these results can be found in most time series textbooks but is also given in Appendix J to illustrate the use of Theorem 3.8 Go to Appendix J.

Corollary 3.9 (Consistency of $\hat{\gamma}(h)$ and $\hat{\rho}(h)$).

Let (X_t) be such that:

- *(X_t) is weakly stationary,*
- *(X_t^2) has an absolutely summable covariance structure,*

then for all $|h| < \infty$ we have

$$\hat{\gamma}(h) \xmapsto{\mathcal{P}} \gamma(h),$$

$$\hat{\rho}(h) \xmapsto{\mathcal{P}} \rho(h).$$

Asymptotic Normality

The Central Limit Theorem (CLT) takes one step further than the law of large number. It identifies the limiting distribution of the (properly scaled) sum of random variables as the normal distribution, which allows us to do the statistical inference (confidence interval and hypothesis testing). The scale will tell us how fast this approximation converges to the normal distribution. These results are generally based on CLT with two most significant results used to prove asymptotic normality.

Theorem 3.10 (CLT for iid sequences).

Suppose X_i are iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$.

Above result can be extended to finite dimensional multivariate iid sequences by the Cramér-Wold device as follows.

Theorem 3.11 (CLT for iid multivariate sequences).

Suppose $\boldsymbol{X}_i$ are iid random variables with $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{\mu} \in \mathbb{R}^d$ and $\text{cov}(\boldsymbol{X}_i) = \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. Then, we have $\sqrt{n}\left(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Asymptotic Normality of Extremum Estimators

General results on the asymptotic normality of extremum estimators can be found in Newey and McFadden 1994. In this course, we shall restrict our attention to a simple example of GMM estimators Go to example. We will see that the asymptotic normality of extremum estimators is implied by combining the following results and techniques:

- the consistency of $\hat{\boldsymbol{\theta}}$,
- the Central Limit Theorem (see Theorem 3.11),
- Slutsky's Theorem (which allows to "mix" convergence in probability and distribution) and
- Taylor expansions.

Asymptotic Normality of Extremum Estimators

In the simple example we considered on GMM estimators, we have used Theorem 3.11 that:

$$\sqrt{n}\left(\hat{\gamma} - \gamma(\boldsymbol{\theta}_0)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{bmatrix} Z_i - \mu_0 \\ Z_i^2 - \mu_0^2 - \sigma_0^2 \\ Z_i^3 - \mu_0^3 - 3\mu_0\sigma_0^2 \end{bmatrix} \xmapsto{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{13}$$

where $\boldsymbol{\Sigma} = \text{cov}\left(\begin{bmatrix} Z_i & Z_i^2 & Z_i^3 \end{bmatrix}^T\right)$. In this example, the extremum estimator we considered is defined using the objective function

$$Q_n(\boldsymbol{\theta}) = -||\hat{\gamma} - \gamma(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}}.$$

Central Limit Theorem

Then we can relate the asymptotic normality of the GMM estimator with the asymptotic normality of the moment conditions (i.e. Equation (13)) we have just shown. Since $\hat{\boldsymbol{\theta}}$ maximizes $Q_n(\boldsymbol{\theta})$, it satifies the following first order condition, i.e.

$$\left.\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}_3,$$

which is equivalent to

$$\left.\frac{\partial}{\partial \boldsymbol{\theta}} \left((\hat{\gamma} - \gamma(\boldsymbol{\theta})) \right)^T \widehat{\mathbf{W}} \left(\hat{\gamma} - \gamma(\boldsymbol{\theta}) \right) \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}_3. \tag{14}$$

Using Taylor expansion for $\gamma(\hat{\boldsymbol{\theta}})$ around the true $\boldsymbol{\theta}_0$, we obtain

$$\hat{\gamma} - \gamma(\hat{\boldsymbol{\theta}}) = \hat{\gamma} - \gamma(\boldsymbol{\theta}_0) + \left.\frac{\partial \left(\hat{\gamma} - \gamma(\boldsymbol{\theta}) \right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) + o_p\left(\frac{1}{\sqrt{n}} \right). \tag{15}$$

Under certain regularity conditions, we have (using Theorem 3.3) that

$$\left.\frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{\gamma} - \gamma(\boldsymbol{\theta}) \right)\right|_{\boldsymbol{\theta}} = \left.-\frac{\partial}{\partial \boldsymbol{\theta}} \gamma(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \xmapsto{p} \left.-\frac{\partial}{\partial \boldsymbol{\theta}} \gamma(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Central Limit Theorem

Then plug Equation (15) into Equation (14), and under certain regularity condition and using Slutsky's Theorem we obtain

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xmapsto{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \mathbf{D}^T \boldsymbol{\Sigma} \mathbf{D}\right), \tag{16}$$

where

$$\mathbf{D} = \left[|| \frac{\partial}{\partial \boldsymbol{\theta}} \left(\boldsymbol{\gamma}(\boldsymbol{\theta})\right) ||_{\mathbf{W}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right]^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\gamma}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)^T \mathbf{W}.$$

The derivation of (16) is due to the fact that:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\overbrace{\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^T \widehat{\mathbf{W}} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right) \right]^{-1}}^{\xrightarrow{p} \left[|| \frac{\partial}{\partial \boldsymbol{\theta}} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})) ||_{\mathbf{W}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right]^{-1}}$$

$$\underbrace{\left(\frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})\right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^T \widehat{\mathbf{W}}}_{\xrightarrow{p} \left(\frac{\partial}{\partial \boldsymbol{\theta}} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})) |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)^T \mathbf{W}} \sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta}_0)\right) + o_p(1).$$

Central Limit Theorem and α-Mixing

For dependent process, the validity of CLT requires the process to be "mixing" or "asymptotic independent". Suppose two events G and H are independent, then

$$|\Pr(G \cap H) - \Pr(G)\Pr(H)| = 0.$$

Based on this idea, many dependence measures have been developed. The most used α-mixing coefficient is one of these dependence measures which can be easily verified for certain stochastic processes.

Definition: Mixing Coefficients

For a stochastic process $\{X_i\}_{i \in \mathbb{Z}}$, define the strong- or α-mixing coefficients as

$$\alpha(t_1, t_2) = sup\{|\Pr(A \cap B) - \Pr(A)\Pr(B)| : \; A \in \mathcal{F}_{-\infty}^{t_1}, B \in \mathcal{F}_{t_2}^{\infty}\},$$

where $\mathcal{F}_{-\infty}^{t_1} = \sigma(X_{-\infty}, \ldots, X_{t_1})$ and $\mathcal{F}_{t_2}^{\infty} = \sigma(X_{t_2}, \ldots, X_{\infty})$ are σ-algebras generated by corresponding random variables.

If the process is stationary, then $\alpha(t_1, t_2) = \alpha(t_2, t_1) = \alpha(|t_1 - t_2|) \equiv \alpha(\tau)$. If $\alpha(\tau) \to 0$ as $\tau \to \infty$, then the process is strong-mixing or α-mixing.

Central Limit Theorem and α-Mixing

Theorem 3.12 (Central Limit Theorem for α-Mixing Process).

Let (X_t) be a strictly stationary process with $\mathbb{E}[X_t] = 0$. $S_n \equiv \sum_{t=1}^{n} X_t$ is the partial sum process with $\sigma_n^2 \equiv \text{var}(S_n)$. Suppose (X_t) is α-mixing, and that for $\delta > 0$

$$\mathbb{E}\left[|X_t|^{2+\delta}\right] \leq \infty, \text{ and } \sum_{n=0}^{\infty} \alpha(n)^{\delta/2+\delta} \leq \infty.$$

Then

$$\lim_{n\to\infty} \frac{\sigma_n^2}{n} = \mathbb{E}\left[|X_t|^2\right] + 2\sum_{k=1}^{\infty} \mathbb{E}\left[X_1 X_k\right] \equiv \sigma^2.$$

If $\sigma^2 > 0$, (X_t) obeys both the central limit theorem with variance σ^2, and the functional central limit theorem.

Remark 16 (Implication of α-mixing).

α-mixing $\Rightarrow$ the covariance structure is absolutely summable (see Definition 2.18).

Implication on the estimation of $\gamma(h)$ and $\rho(h)$

Using the results previously presented, let us quickly revisit the estimation and the inference of $\gamma(h)$ and $\rho(h)$.

Remark 17 (Significant of $\hat{\rho}(h)$).

If (X_t) is a white noise then $\hat{\rho}(h)$ should be equal to 0 if $h \neq 0$. In practice, this is of course not the case due the estimation error of $\hat{\rho}(h)$. The next result gives us a way to assess whether the data comes from a completely random series or whether correlations are statistically significant at some lags.

Theorem 3.13 (Distribution of $\hat{\rho}(h)$ in iid case).

If (X_t) is white noise (with finite variance) and $h = 1, ..., H$ where H is fixed but arbitrary we have that

$$\sqrt{T}\left(\hat{\rho}(h) - \rho(h)\right) \xrightarrow{D} \mathcal{N}\left(0, 1\right).$$

Implication on the estimation of $\gamma(h)$ and $\rho(h)$

Remark 18 (Confidence intervals for $\hat{\rho}(h)$).

Theorem 3.13 implies that an approximate confidence interval for $\hat{\rho}(h)$ (in the iid case) is given by

$$\text{CI}(\rho(h), \alpha) = \hat{\rho}(h) \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{T}}$$

for $0 < h < k < \infty$ and where $z_{1-\frac{\alpha}{2}} \equiv \boldsymbol{\Phi}^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the $\left(1 - \frac{\alpha}{2}\right)$ quantile of a standard normal distribution. Typically, for $\alpha = 0.05$ one would consider the following confidence interval:

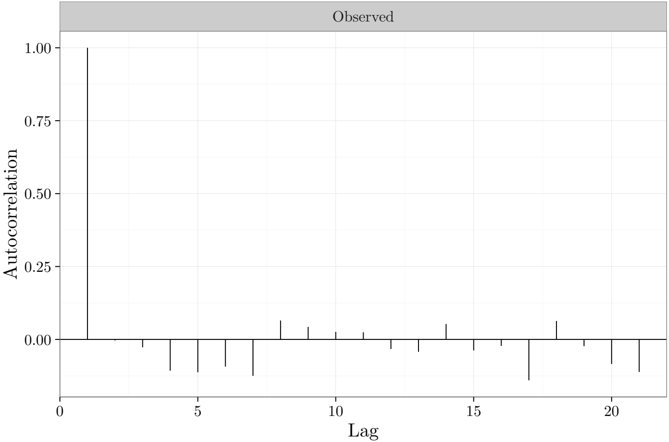
$$\text{CI}(\rho(h), 0.05) = \hat{\rho}(h) \pm \frac{2}{\sqrt{T}}$$

Remark 19 (Proof of Theorem 3.13).

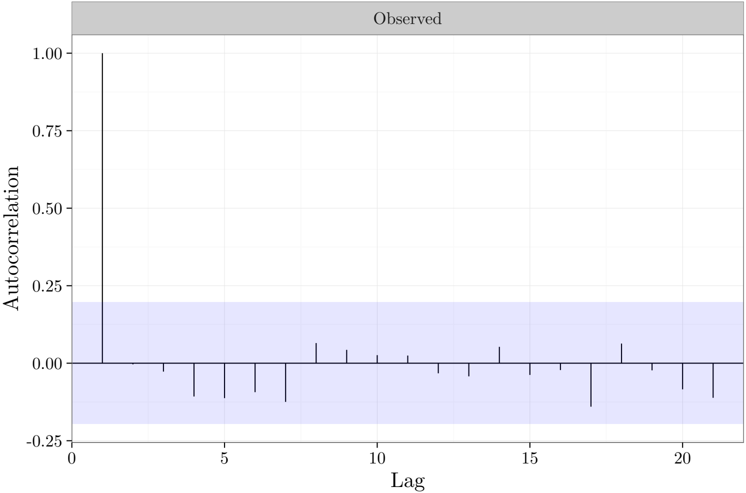
The proof of Theorem 3.13 is straightforward from the CLT and Delta method. It is therefore omitted from this document but can for example be found in Hamilton 1994.

Example: Sample Autocorrelation Function

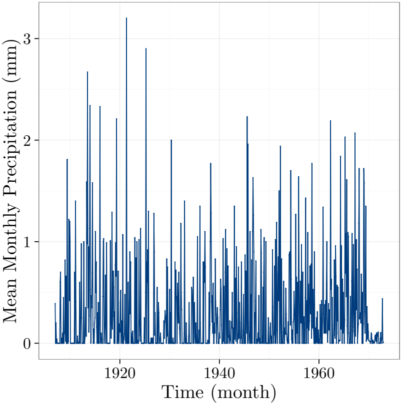
Consider the estimated ACF of a simulated WN process of length $T = 100$.

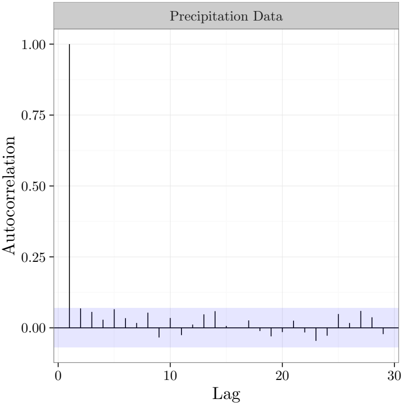


An Example: a White Noise Process



Example: ACF of Precipitation Data





Remark:

The "ACF" plot suggests an absence of linear dependence in this dataset.

References

Baltagi, Badi H (2008). *A Companion to Theoretical Econometrics*. John Wiley & Sons.

Hamilton, J. D. (1994). *Time Series Analysis*. Vol. 2. Princeton university press Princeton.

Komunjer, I. (2012). "Global Identification in Nonlinear Models with Moment Restrictions". In: *Econometric Theory* 28.4, p. 719.

Newey, W. K. and D. McFadden (1994). "Large Sample Estimation and Hypothesis Testing, V in Handbook of Econometrics". In: vol. 4. Elsevier, Amsterdam.

Outline

Estimation of composite stochastic processes

In general, inertial sensor stochastic calibration involves the estimation of the parameters of composite stochastic processes. These models are typically difficult to estimate because of their latent features. In the definition below we characterize the **class of composite stochastic processes** we shall consider here.

Definition 4.1 (Common class of processes for IMU calibration).

Let (X_t) be a sum of latent independent stochastic process such that:

- (X_t) is made of a sum which includes a subset or all processes in the set $\{$**QN**, **WN**, **AR1**, **RW**, **DR**$\}$ (i.e. see Definitions 2.3 to 2.7), where processes in the subset $\{$**QN**, **WN**, **RW**, **DR**$\}$ are included only once and process **AR1** can be included k times $(0 \leq k < \infty)$.
- Let $\mathcal{Q}$ denote an arbitrary compact subset of $\mathbb{R}^+$. Then, the innovation process for processes **WN**, **RW** and **AR1** have respective variances σ^2, γ^2 and ν^2 such that σ^2, γ^2 and $\nu^2 \in \mathcal{Q}$ and processes **QN** and **DR** have $Q^2 \in \mathcal{Q}$ and $|\omega| \in \mathcal{Q}$ respectively.

Estimation of composite stochastic processes

Estimation techniques for composite stochastic processes

There are three main class of estimation techniques that can be used for the estimation of the class of processes defined in the previous slide (i.e. Definition 4.1). The methods are the following:

- **Maximum likelihood based methods:** while these methods are in theory optimal there applicability is extremely limited due numerical reasons and tends to perform badly in practice.
- **Allan variance-based methods:** this class of method is the most popular approach for IMU calibration. However, they typically lead to inconsistent estimators and their finite sample performance is often much lower than GMWM-based technique.
- **GMWM and related methods:** In our (very biased) opinion, these techniques are currently the best choice for the estimation of the parameters of the class of processes considered in Definition 4.1. As we will see, this method is in fact a formalized version Allan variance-based methods.

Maximum likelihood based methods

Maximum likelihood based approaches are generally inappropriate for the estimation of the class of processes considered Definition 4.1. In this course, we shall avoid a technical discussion on likelihood based method and refer the readers to Stebler et al. 2011 and Guerrier, Molinari, and Balamuta 2016 for more details. Instead, we will consider an example to illustrate the numerical issues of this technique.

Suppose we wish to estimate a composite process composed of a WN and an AR1 process, i.e.

$$\begin{aligned} &Y_t = \phi_0 Y_{t-1} + Z_t, \quad Z_t \overset{iid}{\sim} \mathcal{N}\left(0, \nu_0^2\right), \\ &U_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_0^2\right), \quad X_t = Y_t + U_t. \end{aligned} \tag{17}$$

Then, we want to estimate the parameter $\boldsymbol{\theta}_0 = \left[\phi_0 \;\; \nu_0^2 \;\; \sigma_0^2\right]^T$.

Maximum likelihood based methods

Since the process is Gaussian we have

$$\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\right),$$

where $\mathbf{X} \equiv [X_1, ..., X_T]^T$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \equiv \text{cov}(\mathbf{X})$. Since U_t and Y_t are independent, we have

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \text{cov}(\mathbf{X}) = \text{cov}(\mathbf{Y}) + \text{cov}(\mathbf{U}) = \sigma_0^2 \mathbf{I}_T + \frac{\nu_0^2}{1-\phi_0^2}\left[\phi_0^{|i-j|}\right]_{i,j=1,\ldots,T}, \quad (18)$$

where $\mathbf{Y} \equiv [Y_1, ..., Y_T]^T$, $\mathbf{U} \equiv [U_1, ..., U_T]^T$ and where $\mathbf{I}_T$ denotes the identity matrix of dimension T. Note that the form of $\text{cov}(\mathbf{Y})$ is due to the autocovariance of an AR1 which has been discussed in Appendix B.

Based on (18) we can now write the log-likelihood function of the modeled considered here which, up to a constant, can be expressed as

$$\mathcal{L}\left(\boldsymbol{\theta}|\mathbf{X}\right) = -\log\left(\det\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})\right)\right) - \mathbf{X}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}.$$

Maximum likelihood based methods

Therefore, we can following maximum likelihood estimator for $\boldsymbol{\theta}_0$:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; \mathcal{L}\left(\boldsymbol{\theta} | \mathbf{X}\right) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmin}} \; \log\left(\det\left(\boldsymbol{\Sigma}(\boldsymbol{\theta})\right)\right) + \mathbf{X}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{X}. \tag{19}$$

Unfortunately, the applicability of this estimator is essentially impossible when $T > 10^5$ since every evaluation of this the function $\mathcal{L}\left(\boldsymbol{\theta} | \mathbf{X}\right)$ requires to invert a $T \times T$ matrix, which entails a considerable (and often unrealistic) computational burden.

An alternative approach to compute maximum likelihood estimator for $\boldsymbol{\theta}_0$ is based on the EM-algorithm of Dempster, Laird, and Rubin 1977. If the process (Y_t) (or U_t) was observed we could easily estimate the parameters of (17) by considering separately the likelihood of both processes. Since (17) is a state-space model we could use the following approach:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmax}} \; \mathcal{L}\left(\boldsymbol{\theta} | \mathbf{X}, \hat{\mathbf{Y}}(\boldsymbol{\theta})\right),$$

where $\hat{\mathbf{Y}}(\boldsymbol{\theta})$ denotes the estimation of $\mathbf{Y}$ (i.e. the states) based on a Kalman filter assuming $\boldsymbol{\theta}$ to be the correct parameter vector. Unfortunately, this approach suffers for the same computational limitations as (19).

Introduction: Allan Variance

- The Allan Variance (AV) is a statistical technique originally developed in the mid-1960s to study the stability of precision oscillators (see e.g. Allan 1966).
- It can provide information on the types and magnitude of various superimposed noise terms (i.e. composite stochastic processes).
- This method has been adapted to characterize the properties of a variety of devices including inertial sensors (see El-Sheimy, Hou, and Niu 2008).
- The AV is a measure of variability developed for long term memory processes and can in fact be interpreted as a Haar wavelet coefficient variance (see Percival and Guttorp 1994). We will discuss this connection further on.

Definition: Allan Variance

Definition 4.2 (Allan Variance).

We consider the AV at dyadic scales (τ_j) starting from local averages of the process which can be denoted as

$$\bar{X}_t^{(j)} \equiv \frac{1}{\tau_j} \sum_{i=1}^{\tau_j} X_{t-\tau_j+i} \, ,$$

where $\tau_j \equiv 2^j$, $j \in \{x \in \mathbb{N} \, : \, 1 \leq x < \log_2(T) - 1\}$ therefore determines the number of consecutive observations considered for the average. Then, the AV is defined as

$$\text{AVar}_j\left(X_t\right) \equiv \frac{1}{2}\, \mathbb{E}\left[\left(\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)}\right)^2\right].$$

Remark 20 (Alternative scale definition).

The definition of the AV is actually valid for $\tau_j = \left\lfloor 2^j \right\rfloor$ with $j \in \{x \in \mathbb{R} \, : \, 1 \leq x < \log_2(T) - 1\}$. In some case, it could use to consider this alternative definition (see e.g. El-Sheimy, Hou, and Niu 2008) but we shall restrict ourself here to the case where $j \in \{x \in \mathbb{N} \, : \, 1 \leq x < \log_2(T) - 1\}$.

Properties of the Allan Variance

Remark 21 (Notation of the Allan Variance).

For notational simplicity, we may sometimes replace $\text{AVar}_j(X_t)$ by simply ϕ_j^2 when the dependence of the AV to the process (X_t) is evident.

As highlighted earlier, the AV is, among others, a widely and commonly used approach in engineering for sensor calibration as it is linked to the properties of the process (X_t) as shown in the following lemma (see e.g Percival and Walden 2006 for a proof).

Lemma 4.3 (AV connection to PSD).

For a stationary process (X_t) with PSD $S_X(f)$ we have

$$\phi_j^2 \equiv \text{AVar}_j(X_t) = 4\int_0^\infty \frac{\sin^4(\pi f \tau_j)}{(\pi f \tau_j)^2} S_X(f) df.$$

Therefore, this result establishes a direct connection between the AV and PSD. A natural question is therefore whether the mapping PSD $\mapsto$ AV is one-to-one. Greenhall 1998 (see Theorem 1) showed that this is actually not the case. This is illustrated in the following slides.

Spectral Ambiguity of the AV

Consider two (independent) stochastic processes (X_t) and (Y_t) with respective PSD $S_X(f)$ and $S_Y(f)$. Suppose that $S_X(f) \neq S_Y(f)$, then the two processes will have the same AV if

$$\Delta \equiv \int_0^\infty \frac{\sin^4(\pi f \tau_j)}{(\pi f \tau_j)^2} \Phi(f) df = 0,$$

where $\Phi(f) \equiv S_X(f) - S_Y(f)$. To show that it is possible that $\Delta = 0$ when $\Phi(f) \neq 0$, we will use the following critical identity:

$$\sin^4(x) = \sin^2(x) - \frac{1}{4}\sin^2(2x). \tag{20}$$

First, we note that Δ may be expressed using (20) as follows:

$$\begin{aligned}
\Delta &= \int_0^\infty \frac{\sin^4\left(\tau\pi f\right)}{\left(\tau\pi f\right)^2}\Phi(f)df \\
&= \lim_{n\to-\infty} \int_{2^n}^\infty \frac{\sin^2\left(\tau\pi f\right) - \frac{1}{4}\sin^2\left(2\tau\pi f\right)}{\left(\tau\pi f\right)^2}\Phi(f)df.
\end{aligned}$$

Spectral Ambiguity of the AV

Second, by the change of variable $u = 2f$ in the second term we obtain

$$\Delta = \lim_{n \to -\infty} \left[\int_{2^n}^{\infty} \frac{\sin^2(\tau\pi f)}{(\tau\pi f)^2} \Phi(f) df - \frac{1}{2} \int_{2^{n+1}}^{\infty} \frac{\sin^2(\tau\pi u)}{(\tau\pi u)^2} \Phi(f) du \right].$$

Now suppose that $\Phi(f) = 2\Phi(2f)$. In this case, we have $\Phi(f) = 2\Phi(u)$ and therefore we obtain

$$\Delta = \lim_{n \to -\infty} \int_{2^n}^{2^{n+1}} \frac{\sin^2(\tau\pi f)}{(\tau\pi f)^2} \Phi(f) df = 0.$$

Remark 22.

This result demonstrates that the mapping from PSD to Allan variance is not necessarily one-to-one. Greenhall 1998 showed that in the continuous case (i.e. $\tau_j \in \mathbb{R}$) $\Delta = 0$ if and only if $\Phi(f) = 2\Phi(2f)$. However, the "only if" part of this results (while conjectured) is unknown in the discrete case and is currently being investigated.

Properties of the Allan Variance

One reason of explaining the widespread use of the Allan variance for sensor calibration is due to the following additivity property, which is particularly convenient to identify composite stochastic processes (see Definition 2.9).

Corollary 4.4 (Additivity of the AV).

Consider two (independent) stochastic processes (X_t) *and* (Y_t) *with respective PSD* $S_X(f)$ *and* $S_Y(f)$*. Suppose that we observe the process* $Z_t = X_t + Y_t$*. Then, we have*

$$\text{AVar}_j\left(Z_t\right) = \text{AVar}_j\left(X_t\right) + \text{AVar}_j\left(Y_t\right).$$

Proof: The proof of this result is direct from Lemma 4.3. Indeed, since $S_Z(f) = S_X(f) + S_Y(f)$, we have

$$\begin{aligned}
\text{AVar}_j\left(Z_t\right) &= 4\int_0^\infty \frac{\sin^4(\pi f \tau_j)}{(\pi f \tau_j)^2} S_Z(f)df \\
&= 4\int_0^\infty \frac{\sin^4(\pi f \tau_j)}{(\pi f \tau_j)^2} S_X(f)df + 4\int_0^\infty \frac{\sin^4(\pi f \tau_j)}{(\pi f \tau_j)^2} S_Y(f)df \\
&= \text{AVar}_j\left(X_t\right) + \text{AVar}_j\left(Y_t\right).
\end{aligned}$$

■

Properties of the Allan Variance

While Lemma 4.3 is an important results which is very convenient to determine the theoretical AV of a certain stochastic process. However, the applicability of this results is often limited since the integral defined in (4.3) can be intractable. An alternative to Lemma 4.3 has been proposed by Zhang 2008 and is far advantageous from a computational standpoint.

Lemma 4.5 (AV connection to ACF).

For a stationary process (X_t) with variance σ_X^2 and ACF $\rho(h)$ we have

$$\text{AVar}_j\left(X_t\right) = \frac{\sigma_X^2}{\tau_j^2}\left(\tau_j\left[1-\rho(\tau_j)\right] + \sum_{i=1}^{\tau_j-1} i\left[2\rho(\tau_j - i) - \rho(i) - \rho(2\tau_j - i)\right]\right).$$

The proof of this result is instructive and is presented in Xu et al. 2017.

Properties of the Allan Variance

Remark 23.

Using Lemma 4.5, the exact form of the AV for different stationary processes, such as the general class of ARMA models, can easily be derived. Moreover, Zhang 2008 provided the theoretical AV for non-stationary processes such as the random walk and ARFIMA models for which the AV, as mentioned earlier, represents a better measure of uncertainty compared to other methods.

Remark 24.

Lemma 4.5 was extended to non-stationary processes in Xu et al. 2017.

Example: In Appendix K we provide an example on the derivation of the theoretical AV of an MA(1) process (see ▸ Go to Appendix K)

Estimation of the Allan Variance

Several estimators of the AV have been introduced in the literature. The most commonly is (probably) the Maximum-Overlapping AV (MOAV) estimator proposed by Percival and Guttorp 1994, which is defined as follows:

Definition 4.6 (Maximum-Overlapping AV Estimator).

The MOAV is defined as:

$$\hat{\phi}_j^2 \equiv \widehat{\text{AVar}}_j\left(X_t\right) = \frac{1}{2\left(T - 2\tau_j + 1\right)} \sum_{k=2\tau_j}^{T} \left(\bar{X}_k^{(j)} - \bar{X}_{k-\tau_j}^{(j)}\right)^2. \tag{21}$$

We will now study the properties of this estimator through the following lemmas.

Consistency of the Maximum-Overlapping AV Estimator

Lemma 4.7 (Consistency).

Let (X_t) be such that:

- $(X_t - X_{t-1})$ *is a (strongly) stationary process,*
- $(X_t - X_{t-1})^2$ *has absolutely summable covariance structure,*
- $\mathbb{E}\left[(X_t - X_{t-1})^4\right] < \infty$,

Then, we have

$$\widehat{\text{AVar}}_j\left(X_t\right) \xmapsto{\mathcal{P}} \text{AVar}_j\left(X_t\right).$$

The proof of Lemma 4.7 is given in Appendix L Go to Appendix L

Remark 25 (Connection to Wavelet Variance).

This result is closely related by the results of Percival 1995 on the wavelet variance. We shall explore the connection between the AV and wavelet variance in the next section.

Asymptotic Normality of the MOAV Estimator

Compare to consistency, the asymptotic normality requires stronger conditions given in the following lemma.

Lemma 4.8 (Asymptotic normality).

Let (X_t) be such that:

- $(X_t - X_{t-1})$ *is a (strongly) stationary process.*
- $(X_t - X_{t-1})$ *is strong mixing process with mixing coefficient* $\alpha(n)$ *such that* $\sum_{n=1}^{\infty} \alpha(n)^{\delta/2+\delta} < \infty$ *for some* $\delta > 0$.
- $\mathbb{E}\left[(X_t - X_{t-1})^{4+\delta}\right] < \infty$ *for some* $\delta > 0$.

Then, under these conditions we have that

$$\sqrt{T}\left(\widehat{\text{AVar}}_j\left(X_t\right) - \text{AVar}_j\left(X_t\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_T^2/T),$$

where $\sigma_T^2 \equiv \sum_{h=-\infty}^{\infty} \text{cov}\left(\left(\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)}\right)^2, \left(\bar{X}_{t+h}^{(j)} - \bar{X}_{t+h-\tau_j}^{(j)}\right)^2\right)$.

The proof of Lemma 4.8 is given in Appendix M Go to Appendix M

Confidence Interval of the MOAV Estimator

Based on the asymptotic normality results (Lemma 4.8), we can construct the $1-\alpha$ confidence intervals for $\widehat{\text{AVar}}_j\left(X_t\right)$ as

$$\text{CI}\left(\text{AVar}_j\left(X_t\right)\right) = \left[\widehat{\text{AVar}}_j\left(X_t\right) \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_T}{T}\right],$$

where $z_{1-\frac{\alpha}{2}} \equiv \mathbf{\Phi}^{-1}\left(1-\frac{\alpha}{2}\right)$ is the $\left(1-\frac{\alpha}{2}\right)$ quantile of a standard normal distribution.

However, the so called "Long-Run Variance" σ_T^2 is usually unknown. Many methods have been proposed to consistently estimate it under mild conditions (see e.g. Newey and West 1986).

Remark 26.

Gaussian-based confidence intervals are often problematic with the AV as the lower limit of CI can very well be negative. Next, we will discuss an alternative method to construct the CI for such statistic.

Allan Variance log-log Representation

As illustrated in Lemmas 4.3 and 4.5 the AV depends on the properties of the stochastic process (X_t). We will see that "log-log" representation of the AV is often useful for the identify various processes that may compose (X_t).

For example, let's suppose that X_t is a white noise process. We showed in Appendix K that the theoretical AV of such process is given

$$\phi_j^2 \equiv \text{AVar}_j(X_t) = \frac{\sigma^2}{\tau_j}.$$

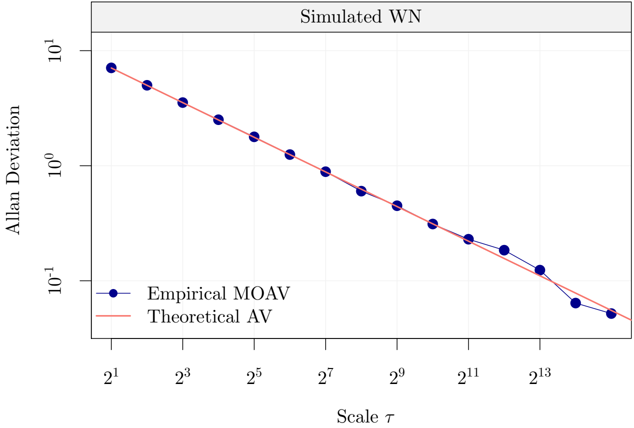
Therefore, we have that the Allan Deviation or AD (i.e. $\sqrt{\text{AVar}_j(X_t)}$ or ϕ_j) is such that

$$\log(\phi_j) = \log\left(\sqrt{\frac{\sigma^2}{\tau_j}}\right) = \log(\sigma) - \frac{1}{2}\log(\tau_j). \tag{22}$$

Thus, the log of the AD is linear in τ_j with a slope of $-1/2$ and with intercept $\log(\sigma)$. Let us start by considering a simple simulated example.

Allan Deviation of a WN process

Simulation based on a white noise process with $\sigma^2 = 10^2$ and $T = 10^5$.



Allan Variance log-log Representation

Suppose now that (X_t) is composite stochastic process composed of a WN (see 2.3) and RW (see 2.6). For simplicity, we assume that $X_t = Y_t + W_t$ where Y_t is a white noise process with variance σ^2 and W_t a random walk with variance γ^2. We already know that

$$\log\left(\text{AVar}_j(Y_t)\right) = \log\left(\sigma\right) - \frac{1}{2}\log(\tau_j).$$

and it can be shown (using for example Lemma 4.3) that

$$\text{AVar}_j(W_t) = \frac{1}{3}\gamma^2\tau_j,$$

and therefore we obtain

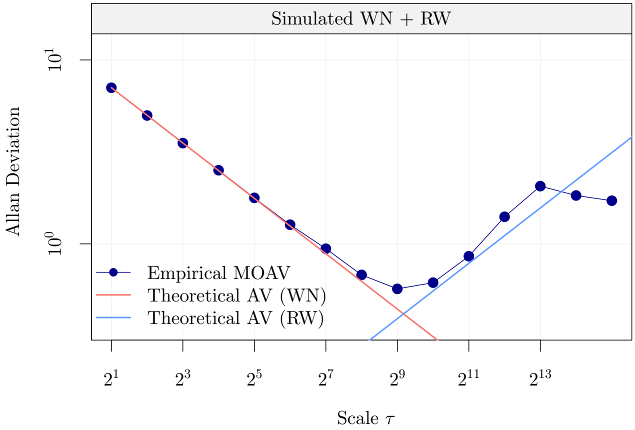
$$\log\left(\sqrt{\text{AVar}_j(W_t)}\right) = \log\left(\sqrt{\frac{1}{3}\gamma^2\tau_j}\right) = \log\left(\frac{1}{\sqrt{3}}\gamma\right) + \frac{1}{2}\log(\tau_j).$$

Thus, the log of the AD of (Z_t) is also linear in τ_j with a slope of $+1/2$. By Corollary 4.4 we also have that

$$\text{AVar}_j(X_t) = \text{AVar}_j(Y_t) + \text{AVar}_j(W_t) = \frac{\sigma^2}{\tau_j} + \frac{1}{3}\gamma^2\tau_j. \tag{23}$$

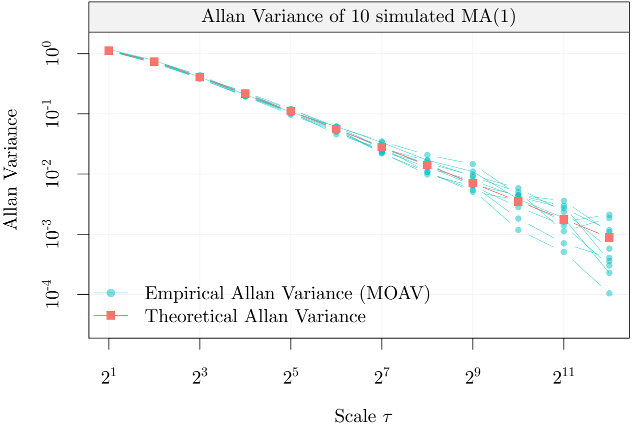
Allan Deviation of a WN + RW process

Simulation based on a white noise process with $\sigma^2 = 10^2$ and a random walk process with $\gamma^2 = 0.03^2$ and $T = 10^5$.



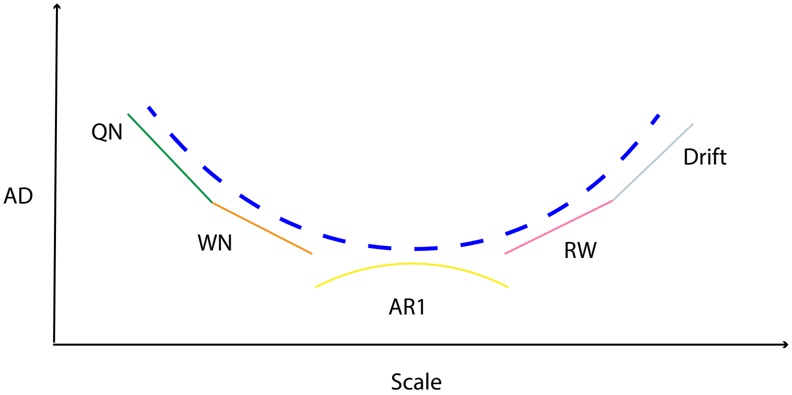
Simulated Example: MA(1)

Using the formula we derived in ▸ Example Theoretical AV MA(1). We simulated 10 MA(1) with $\theta = 0.9$ and $\sigma^2 = 1$. Their empirical AV (i.e. MOAV) are presented below together with the theoretical AV of this process in red.



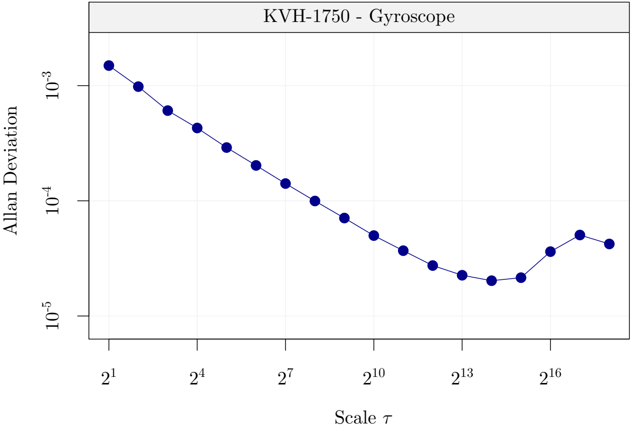
Parameter Estimation through the Allan Variance

The AV is a powerful technique to identify the processes considered in Definition 4.1. Indeed, the five processes we are considering are characterized in log-log representation as follows: *(i)* QN - linear with slope of -1; *(ii)* WN - linear with slope of -1/2; *(iii)* AR1 - curved shape with a slope $\in [-1/2, 1/2]$; *(iv)* RW - linear with slope of 1/2; *(v)* DR - linear with slope of 1.



Real Data: MOAV KVH-1750 (Gyro), 30 mins at 500Hz

The shape of the AD suggests that a WN + RW may be a reasonable approximation. But how could we estimate the parameters of this model?



AV-based Estimation

Main idea:

We have seen that there exist a mapping from $\boldsymbol{\theta}_0$ (model's parameters) to the AV (or AD), i.e.

$$\boldsymbol{\theta}_0 \mapsto \text{PSD or AVS} \mapsto \text{AV}.$$

In general, the mapping from the PSD to the AV is not one-to-one, nevertheless, one may hope that in some case this mapping can be "inverted" (in some sense) so that the AV may be used to estimated $\boldsymbol{\theta}_0$. This is central of AV linear regression approach.

Let us assume that we want to estimate the parameter of a QN, WN, RW or DR process. Then the process (X_t) is such that it has a linear representation in a $\log(\phi_{\tau_j}) - \log(\tau_j)$ plot for the set of scales $\boldsymbol{\eta} \in \boldsymbol{\mathcal{G}}$ where

$$\boldsymbol{\mathcal{G}} = \left\{ \{\tau_k, \ldots, \tau_{k+h}\} \, | k, h \in \mathbb{N}^+, k+h \leq J \right\}$$

denotes all possible sets which contain adjacent scales having cardinality $|\boldsymbol{\eta}| > 0$.

AV-based Estimation

Then, there exists a linear relationship between $\log\left(\phi_{\tau_j}(\theta)\right)$ and $\log\left(\tau_j\right)$ we can write

$$\log\left(\phi_{\tau_j}\left(\theta\right)\right) = g\left(\theta\right) + \lambda \log\left(\tau_j\right), \;\; \forall \tau_j \in \boldsymbol{\eta}, \tag{24}$$

where the function $g(\cdot)$ as well as the constant λ are known and depend on the model. For a white noise, we have for example $g(\sigma) = \log(\sigma)$ and $\lambda = -1/2$, see (22). This linear relationship leads to the following "least-squares" estimator of θ, noted as $\hat{\theta}_{AV}$ and defined as

$$\hat{\theta}_{AV} \equiv \underset{\theta \in \Theta}{\text{argmin}} \sum_{\tau_j \in \boldsymbol{\eta}} \left[\log\left(\hat{\phi}_{\tau_j}\right) - g\left(\theta\right) - \lambda \log\left(\tau_j\right)\right]^2, \tag{25}$$

which is an extremum estimator (see Definition 3.1) and admit the solution

$$\hat{\theta}_{AV} = g^{-1}\left\{\frac{1}{|\boldsymbol{\eta}|} \sum_{\tau_j \in \boldsymbol{\eta}} \left[\log\left(\hat{\phi}_{\tau_j}\right) - \lambda \log\left(\tau_j\right)\right]\right\}, \tag{26}$$

where $|\boldsymbol{\eta}|$ denotes the cardinality of the set $\boldsymbol{\eta}$. The derivation of (26) is instructive and is given in Appendix N ▸ Go to Appendix N

AV-based Estimation

Let us consider a simple example of the use of (26) with a WN process. We already derived that

$$\log\left(\text{AVar}_j(Y_t)\right) = \log\left(\sigma\right) - \frac{1}{2}\log(\tau_j)$$

implying that $g(\sigma) = \log(\sigma)$ (and thus $g^{-1}(\sigma) = \exp(\sigma)$) and $\lambda = -1/2$. This leads to the following estimator of σ,

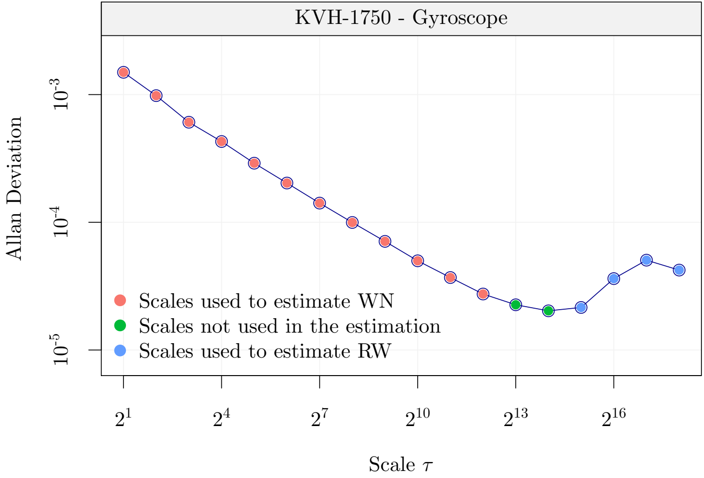
$$\hat{\sigma}_{AV} = \exp\left\{\frac{1}{|\boldsymbol{\eta}|}\sum_{\tau_j \in \boldsymbol{\eta}} \log\left(\hat{\phi}_{\tau_j}\right) + \frac{1}{2}\log\left(\tau_j\right)\right\}. \tag{27}$$

Similarly for a RW, we obtain

$$\hat{\gamma}_{AV} = \sqrt{3}\exp\left\{\frac{1}{|\boldsymbol{\eta}|}\sum_{\tau_j \in \boldsymbol{\eta}} \log\left(\hat{\phi}_{\tau_j}\right) - \frac{1}{2}\log\left(\tau_j\right)\right\}. \tag{28}$$

Using these results let's estimate the parameters on the KVH-1750 gyro.

Reasonable Model: WN + RW



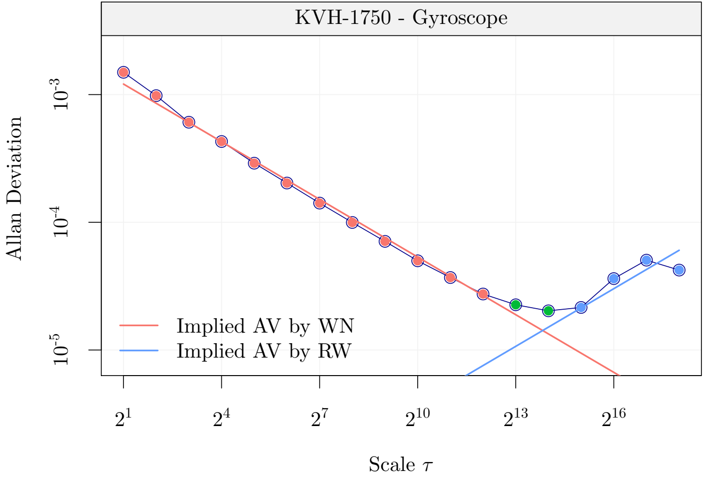
Example: MOAV KVH-1750 (Gyro) - Estimation

Using (27) and (28) on the selected scales, we obtain the following estimates:

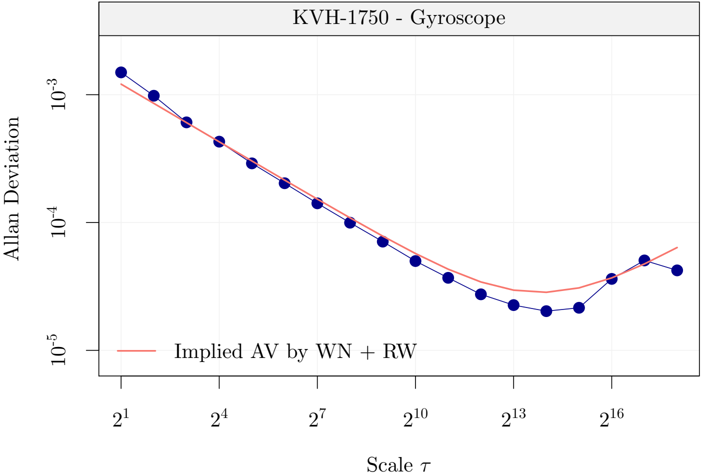
$$\hat{\sigma}_{AV} = \frac{1}{2} \exp\left\{ \frac{1}{12} \sum_{j=1}^{12} \log\left(\hat{\phi}_j\right) + \frac{1}{2} \log\left(2^j\right) \right\} \approx 1.710 \cdot 10^{-3}$$

$$\hat{\gamma}_{AV} = \sqrt{3} \exp\left\{ \frac{1}{4} \sum_{j=15}^{18} \log\left(\hat{\phi}_j\right) - \frac{1}{2} \log\left(2^j\right) \right\} \approx 2.043 \cdot 10^{-7}.$$

Example: MOAV KVH-1750 (Gyro) - Estimation



Example: MOAV KVH-1750 (Gyro) - Estimation



Properties of AV-based Estimation

Before studying the properties of the AV-based estimation technique, we will assess whether the MOAV estimator (see Definition 4.6 is consistent (see Lemma 4.7) and asymptotically normal (see Lemma 4.8) under the setting of given Definition 4.1. Lemma 1 of Guerrier, Molinari, and Stebler 2016 answered that question by showing that all the conditions of Lemmas 4.7 and 4.8 are satisfied this setting.

Since $\hat{\theta}_{AV}$ admits the analytically solution given in (26) the consistency of the AV-based estimator can be directly studied using the consistency of the AV and the continuous mapping theorem (see Theorem 3.3). Indeed, using the latter we have

$$\hat{\theta}_{AV} = g^{-1}\left\{\frac{1}{|\boldsymbol{\eta}|}\sum_{\tau_j \in \boldsymbol{\eta}}\left[\log\left(\hat{\phi}_{\tau_j}\right) - \lambda \log\left(\tau_j\right)\right]\right\}$$

$$\xmapsto{\mathcal{P}} g^{-1}\left\{\frac{1}{|\boldsymbol{\eta}|}\sum_{\tau_j \in \boldsymbol{\eta}}\left[\log\left(\phi_{\tau_j}\right) - \lambda \log\left(\tau_j\right)\right]\right\} = \theta^*,$$

and therefore if we can show that $\theta^* = \theta_0$ then the estimator is consistent.

Properties of AV-based Estimation

Corollary 4.9 (Consistency of AV-based Estimation).

The AV-based estimation technique is consistent if the process (X_t) is composed of either a QN, WN or DR process.

Proof: Simply show that $\theta^* = \theta_0$.

Corollary 4.10 (Inconsistency of AV-based Estimation).

The AV-based estimation technique is not consistent if the process (X_t) includes at least one RW process or contains more than one latent process.

Proof: It can be show at any model $\theta^* > \theta_0$, implying the consistency of the method.

A formal proof of these results are given in Guerrier, Molinari, and Stebler 2016.

Properties of AV-based Estimation

Remark 27 (Why is the AV-based estimation inconsistent?).

The main reason for the inconsistency of this approach is that the effect of multiple superimposed stochastic processes cannot (perfectly) separated in the log-log representation. This suggests the following alternative definition of $\hat{\theta}_{AV}$,

$$\hat{\theta}^{*}_{AV} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{j=1}^{J} \left(\hat{\phi}_j - \phi_j(\boldsymbol{\theta})\right)^2,$$

where $\phi_j(\boldsymbol{\theta})$ denotes the theoretical AV under the assumption that $\boldsymbol{\theta}$ is the correct parameter vector. We will see that this approach is a special case of the GMWM estimator and it leads to consistent and asymptotically normally distributed estimators.

Example: MOAV KVH-1750 (Gyro) - Estimation with $\hat{\theta}^*_{AV}$

Let us revisit the KVH-1750 (Gyro) dataset using this alternative AV-estimation approach. We show in (23) that

$$\mathrm{AVar}_j(X_t) = \mathrm{AVar}_j(Y_t) + \mathrm{AVar}_j(W_t) = \frac{\sigma^2}{\tau_j} + \frac{1}{3}\gamma^2\tau_j.$$

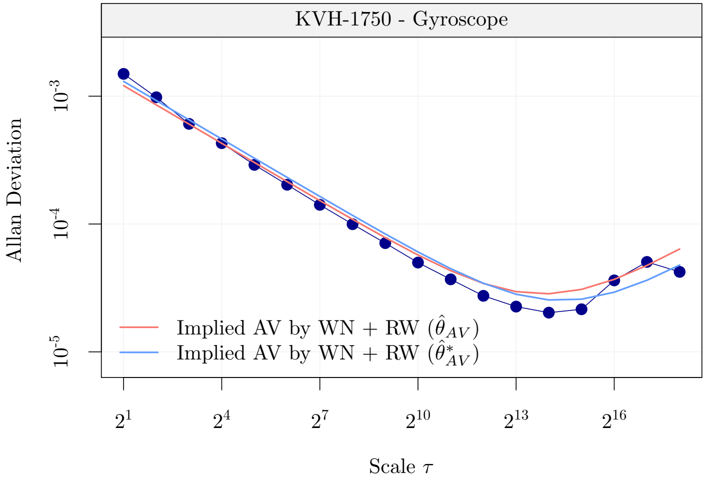
and therefore we can define the following estimator:

$$\hat{\boldsymbol{\theta}}^*_{AV} = [\hat{\sigma}^*_{AV} \quad \hat{\gamma}^*_{AV}]^T = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\mathrm{argmin}} \sum_{j=1}^{J} \left(\phi_j^2 - \frac{\sigma^2}{\tau_j} - \frac{1}{3}\gamma^2\tau_j\right)^2,$$

corresponding to the point estimate:

$$\hat{\sigma}^*_{AV} \approx 1.847 \cdot 10^{-3}$$
$$\hat{\gamma}^*_{AV} \approx 1.493 \cdot 10^{-7}.$$

Example: MOAV KVH-1750 (Gyro) - Estimation with $\hat{\theta}^*_{AV}$



References

Allan, D. W. (1966). "Statistics of Atomic Frequency Standards". In: *Proceedings of the IEEE* 54.2, pp. 221–230.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

El-Sheimy, N., H. Hou, and X. Niu (2008). "Analysis and Modeling of Inertial Sensors using Allan Variance". In: *IEEE Transactions on Instrumentation and Measurement* 57.1, pp. 140–149.

Greenhall, C. A. (1998). "Spectral Ambiguity of Allan Variance". In: *IEEE Transactions on Instrumentation and Measurement* 47.3, pp. 623–627.

Guerrier, S., R. Molinari, and J. Balamuta (2016). "Discussion on Maximum Likelihood-Based Methods for Inertial Sensor Calibration". In: *IEEE Sensors Journal* 16.14, pp. 5522–5523.

Guerrier, S., R. Molinari, and Y. Stebler (2016). "Theoretical Limitations of Allan Variance-based Regression for Time Series Model Estimation". In: *IEEE Signal Processing Letters* 23.5, pp. 597–601.

Newey, Whitney K and Kenneth D West (1986). *A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix*.

Percival, D. B. and P. Guttorp (1994). "Long-memory processes, the Allan variance and wavelets". In: *Wavelet Analysis and its Applications*. Vol. 4. Elsevier, pp. 325–344.

Percival, D. B. and A. T. Walden (2006). *Wavelet methods for time series analysis*. Cambridge university press.

Percival, D. P. (1995). "On Estimation of the Wavelet Variance". In: *Biometrika* 82.3, pp. 619–631.

Stebler, Y. et al. (2011). "Constrained Expectation-Maximization Algorithm for Stochastic Inertial Error Modeling: Study of Feasibility". In: *Measurement Science and Technology* 22.8, p. 085204.

Outline

Definition: Wavelet Coefficients

Definition 5.1 (Wavelet Coefficients).

In a similar way to the AV, we can define the Wavelet Variance (WV) at dyadic scales (τ_j) for $j \in \{x \in \mathbb{N} : 1 \leq x < \log_2(T) - 1\}$. To do so, we first need to define the wavelet filters $h_{j,l}$ as "weights" having the following properties

$$\sum_{l=0}^{L_j-1} h_{j,l} = 0, \; \sum_{l=0}^{L_1-1} h_{1,l}^2 = \frac{1}{2} \text{ and } \sum_{l=-\infty}^{\infty} h_{1,l} h_{1,l+2m} = 0,$$

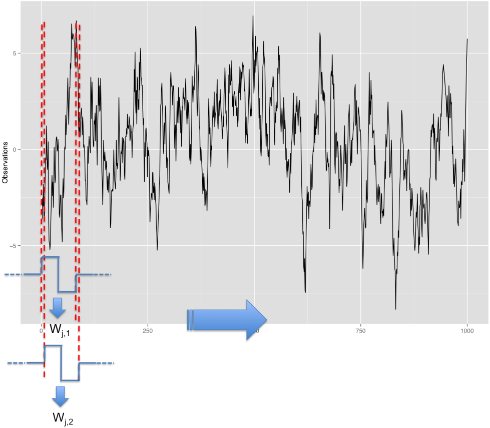
where $m \in \mathbb{N}^+$, $L_j = (2^j - 1)(L_1 - 1) + 1$ is the length of the filter at level j and L_1 is the length of the first level filter $h_{1,l}$.

Then, the wavelet coefficients $W_{j,t}$ are defined as

$$W_{j,t} = \sum_{l=0}^{L_j-1} h_{j,l} X_{t-l}$$

Example: Wavelet Coefficients

The wavelet coefficients can be applied in different ways but the most useful (in the present context) is the Maximum Overlap Discrete Wavelet Transform (MODWT):



Definition: Wavelet Variance

Definition 5.2 (Wavelet Variance).

Once we have defined the wavelet coefficients $W_{j,t}$, we can now define the WV as being the variance of the wavelet coefficients at level j

$$\nu_j^2 \equiv \mathrm{Var}[W_{j,t}].$$

Lemma 5.3 (Relation between WV and AV).

The WV has an exact relationship to the AV when using the ***Haar wavelet filter*** *$h_{j,l}$, i.e. $\nu_j^2 = 2\,\mathrm{AVar}_j$.*

The proof of Lemma 5.3 is given in Percival and Guttorp 1994.

Remark 28 (Theoretical WV).

Lemma 5.3 implies that the (Haar) WV is equivalent (up to a scaling factor) to the AV. Therefore, the theoretical WV can computed using Lemmas 4.5 and 4.3.

Wavelet Variance Estimator

Definition 5.4 (Wavelet Variance Estimation).

An unbiased estimator for the WV issued from a certain wavelet transform is given by the **MODWT estimator** (see Percival 1995)

$$\hat{\nu}_j^2 = \frac{1}{M_j(T)} \sum_{t=L_j}^{T} W_{j,t}^2$$

where $M_j(T) = T - L_j + 1$.

Remark 29 (Alternative WV Estimators).

There exist other estimators of the WV, in particular robust estimators that perform well when the observed time series suffers from the presence of outliers or other forms of contamination. However, when the observed time series is uncontaminated, $\hat{\nu}_j^2$ is the most statistically efficient.

Properties of the Wavelet Variance Estimator

Lemma 5.5 (Consistency).

Let (X_t) be such that:

- $(X_t - X_{t-1})$ *is a (strongly) stationary process,*
- $(X_t - X_{t-1})^2$ *has absolutely summable covariance structure,*
- $\mathbb{E}\left[(X_t - X_{t-1})^4\right] < \infty$ *for some* $\delta > 0$*.*

Defining $\hat{\boldsymbol{\nu}} \equiv [\hat{\nu}_j^2]_{j=1,\dots,J}$*, with J being a bounded quantity, then we have*

$$\hat{\boldsymbol{\nu}} \xmapsto{\mathcal{P}} \boldsymbol{\nu}\,.$$

The proof of Lemma 5.5 can be found in Appendix O Go to Appendix O

Properties of the Wavelet Variance Estimator

Compare to consistency, asymptotic normality requires stronger conditions given in the following lemma. J is also bounded.

Lemma 5.6 (Asymptotic normality).

Let (X_t) be such that:

- $(X_t - X_{t-1})$ *is a (strongly) stationary process,*
- $(X_t - X_{t-1})$ *is a strong mixing process with mixing coefficient $\alpha(n)$ such that $\sum_{n=1}^{\infty} \alpha(n)^{\delta/2+\delta} < \infty$ for some $\delta > 0$,*
- $\mathbb{E}\left[(X_t - X_{t-1})^{4+2\delta}\right] < \infty$ *for some $\delta > 0$.*

Then we have,

$$\sqrt{T}\,(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\nu}}$ with elements $\sigma_{ij}^2 \equiv \sum_{h=-\infty}^{\infty} \text{cov}\left(W_{i,0} W_{j,0}, W_{i,h} W_{j,h}\right)$.

The proof of Lemma Appendix P can be found in Appendix P Go to Appendix P

Confidence Interval of the Wavelet Variance Estimator

Based on the asymptotic normality results (Lemma 5.6), we can construct the $(1-\alpha)$-confidence intervals for $\hat{\nu}_j$ as

$$\mathrm{CI}\left(\nu_j, \alpha\right) = \left[\hat{\nu}_j \pm z_{1-\frac{\alpha}{2}} \frac{\sigma_{jj}}{T}\right],$$

where $z_{1-\frac{\alpha}{2}} \equiv \mathbf{\Phi}^{-1}\left(1-\frac{\alpha}{2}\right)$ is the $\left(1-\frac{\alpha}{2}\right)$ quantile of a standard normal distribution. However, the Long-Run Variance σ_{jj} is usually known. Many methods has been proposed to consistently estimate under mild conditions (see e.g. Newey and West 1986). However, for finite T, CI based on the asymptotic normality would be problematic, because the lower limit of CI can be negative. Alternatively, to avoid this problem, we can use following asymptotic result

$$\eta \frac{\hat{\nu}_j}{\nu_j} \xmapsto{\mathcal{D}} \chi^2_\eta,$$

where η is a constant which can be estimated by $\max\left\{\frac{M_j}{2^j}, 1\right\}$. Moreover, this method can also avoid the estimation of Long-Run Variance σ_{jj}. Then the confidence interval is

$$\mathrm{CI}\left(\nu_j, \alpha\right) = \left[\frac{\eta\hat{\nu}_j}{F^{-1}_{\chi^2_\eta}(\alpha/2)}, \frac{\eta\hat{\nu}_j}{F^{-1}_{\chi^2_\eta}(1-\alpha/2)}\right].$$

Inverse Mapping

Find the parameters implied by the observed WV

In a similar way to the method based on the AV, the idea would be to find the model parameter values implied by the observed WV, i.e.

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\boldsymbol{\nu}}),$$

where $\boldsymbol{\theta}(\boldsymbol{\nu})$ is the inverse of the function $\boldsymbol{\nu}(\boldsymbol{\theta})$ (the theoretical WV).

Problems with inverse mapping

The inverse $\boldsymbol{\theta}(\boldsymbol{\nu})$ can often be complicated to find when considering the parametric model that characterizes the theoretical WV $\boldsymbol{\nu}(\boldsymbol{\theta})$ which are often complicated when considering stochastic errors coming from inertial sensors.

The Generalized Method of Wavelet Moments

Definition 5.7 (Generalized Method of Wavelet Moments).

The GMWM estimator is obtained as follows

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmin}} \, (\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta}))^T \boldsymbol{\Omega} (\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})),$$

where

- $\boldsymbol{\Omega}$ is a positive-definite weighting matrix
- $\boldsymbol{\nu}(\boldsymbol{\theta})$ is the theoretical WV implied by a model with parameter vector $\boldsymbol{\theta}$.

GMWM as an Extremum Estimator

The GMWM is part of the **class of extremum estimators** (see Definition 3.1), i.e.

$$\hat{Q}_n(\boldsymbol{\theta}) \equiv -(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta}))^T \boldsymbol{\Omega} (\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\boldsymbol{\theta})).$$

Consistency of the GMWM

Lemma 5.8.

Suppose that:

(C1) *The function* $\boldsymbol{\nu}(\boldsymbol{\theta})$ *is such that* $\boldsymbol{\nu}(\boldsymbol{\theta}) = \boldsymbol{\nu}(\boldsymbol{\theta}_0)$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

(C2) *The set* $\boldsymbol{\Theta}$ *is compact*

(C3) *The function* $\boldsymbol{\nu}(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$

(C4) $\hat{\boldsymbol{\nu}} \xrightarrow{\mathcal{P}} \boldsymbol{\nu}$

(C5) $\boldsymbol{\Omega}$ *is positive-definite and "consistent" (if estimated).*

Then, we have:

$$\hat{\boldsymbol{\theta}} \xmapsto{\mathcal{P}} \boldsymbol{\theta}_0.$$

Proof (Sketch): Consistency of the GMWM

Remark 30.

Identifiability Condition (C1) is equivalent to Condition 3.4 and, as seen earlier, it can be verified by checking the assumptions in Theorem 3.6. In the latter all assumptions are generally satisfied considering the most commonly employed time series models while showing that $J(\boldsymbol{\theta})$ is nonnegative (or nonpositive) is not necessarily straightforward, especially when dealing with latent models.

Study of $J(\boldsymbol{\theta})$

The nonnegative or nonpositive property of $J(\boldsymbol{\theta})$ has been verified for a wide class of time series models (see next section).

Proof (Sketch): Consistency of the GMWM

While Condition (C2) is always assumed and, as discussed for Condition (C1), Condition (C3) is generally verified for all times series considered previously and is therefore nearly always verified.

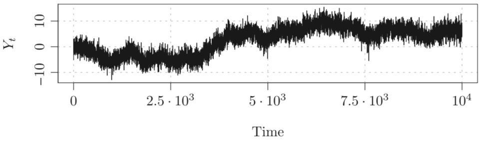
Condition (C4) is verified based on Lemma 5.5 and finally Condition (C5) is either assumed or can be verified using an appropriate estimator for $\boldsymbol{\Omega}$. Based on the latter two conditions, it is possible to verify that $\hat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$ (Condition 3.4).

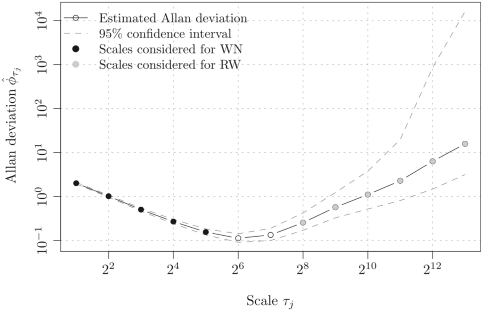
Based on these verifiable (verified) conditions, we have that

$$\hat{\boldsymbol{\theta}} \xmapsto{\mathcal{P}} \boldsymbol{\theta}_0,$$

thus concluding the proof. ■

Simulated Example: WN + RW





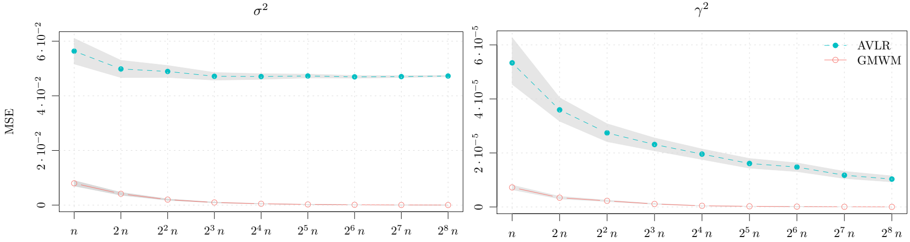
Simulation exercise

Consider the model "WN + RW" with parameters: $\sigma_0^2 = 4$ and $\gamma^2 = 0.01$. Simulation based on $B = 10^3$ Monte-carlo replications. (example coming from Guerrier, Molinari, and Stebler 2016)

Simulated Example: WN + RW

Consistency of AVLR and GMWM

If an estimator is consistent, then its Mean Squared Error (MSE) should tend to zero as the sample size (n) increases.



Conditions for Asymptotic Normality

Theorem 5.9.

In addition to Conditions (C1) to (C5), we assume:

(C6) $\boldsymbol{\theta}_0$ *is an interior point to* $\boldsymbol{\Theta}$

(C7) $\mathbf{H}(\boldsymbol{\theta}_0) \equiv \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \boldsymbol{\nu}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ *exists and is non-singular*

(C8) $\sqrt{T}(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}) \xrightarrow[T\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Then, we have

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow[T\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

where $\mathbf{V}$ *is the asymptotic covariance matrix of* $\hat{\boldsymbol{\theta}}$.

Proof (Sketch): Asymptotic Normality for the GMWM

The proof of asymptotic normality of the GMWM closely follows that of asymptotic normality for extremum estimators (see from Theorem 3.11 forwards).

Condition (C6) is another condition that inevitably needs to be assumed (as for Condition (C2)) while Condition (C7) is not necessary but guarantees the feasibility of Taylor expansion according to the approach used (we avoid technical details).

Knowing that Condition (C8) is verified (see Lemma 5.6), then we have that

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xmapsto[T\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{V})\,.$$

■

Model Selection using WVIC

In order to select a model among a set of candidate models for an observed time series, it is possible to estimate their out-of-sample prediction error and select the model that minimizes this error.

Definition 5.10 (Wavelet Variance Information Criterion (WVIC)).

The WVIC provides the out-of-sample prediction error between the WV implied by the selected model and the WV observed out-of-sample:

$$\mathrm{WVIC} = \mathbb{E}\left[\mathbb{E}_0\left[\left(\hat{\boldsymbol{\nu}}^0 - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right)^T \boldsymbol{\Omega}\left(\hat{\boldsymbol{\nu}}^0 - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right)\right]\right].$$

Estimation of WVIC

An estimator of the WVIC is given by

$$\widehat{\mathrm{WVIC}} = \left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right)^T \boldsymbol{\Omega}\left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right) + 2\,\mathrm{tr}\left\{\widehat{\mathrm{cov}}\left[\hat{\boldsymbol{\nu}}, \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right]\boldsymbol{\Omega}^T\right\}$$

Model Selection using WVIC

Remark 31 (How to compute $\widehat{\text{WVIC}}$).

While the first term in $\widehat{\text{WVIC}}$ is given by the value of the GMWM objective function at the solution $\hat{\boldsymbol{\theta}}$, it is not straightforward to compute the second term (optimism):

$$\text{tr}\left\{\widehat{\text{cov}}\left[\hat{\boldsymbol{\nu}}, \boldsymbol{\nu}(\hat{\boldsymbol{\theta}})\right]\boldsymbol{\Omega}^T\right\}.$$

Computation of optimism term

There are various approaches to compute the optimism term:

- Compute the analytic form of the covariance term (see Guerrier et al., 2018).
- Compute the term via parametric bootstrap (see Guerrier, Molinari, and Skaloud 2015).
- Compute it using independent replicates (see Radi et al. 2017).

Model Selection using WV

Properties of the WVIC

Let us assume we have a set of candidate models $\mathcal{M}$ where $\hat{m}$ denotes the model selected using $\widehat{\text{WVIC}}$ and m^* denotes the best model within $\mathcal{M}$. Then

- The WVIC overfits meaning that asymptotically it selects $m^* \subset \hat{m}$ (i.e. it always contains the best model)
- The WVIC is loss efficient:

$$\frac{\hat{Q}_n(\hat{\boldsymbol{\theta}}_{\hat{m}})}{\hat{Q}_n(\hat{\boldsymbol{\theta}}_{m^*})} \xrightarrow[T\to\infty]{\mathcal{P}} 1\,,$$

  i.e. the selected model performs as well as the best model asymptotically

Can we do better than the GMWM?

The GMWM generally has the following advantages:

- It can easily estimate many state-space and latent time series models
- Computationally efficient method
- It has convenient asymptotic properties
- It is a feasible and computationally efficient approach compared to alternative methods (especially for large sample sizes)
- It can easily be extended to more complex modelling settings (e.g. statistically robust, multivariate, non-stationary, etc.)

Can we do better than the GMWM?

IEEE SENSORS JOURNAL, VOL. 16, NO. 1, JANUARY 1, 2016 163

Maximum Likelihood Identification of Inertial Sensor Noise Model Parameters

Janosch Nikolic, *Student Member, IEEE*, Paul Furgale, *Member, IEEE*, Amir Melzer, *Member, IEEE*, and Roland Siegwart, *Fellow, IEEE*

***Abstract*—Accurate visual-inertial localization and mapping systems require accurate calibration and good sensor error models. To this end, we present a simple offline method to automatically determine the parameters of inertial sensor noise models. The proposed methodology identifies noise processes across a large range of strength and time-scales, for example, weak gyroscope bias fluctuations buried in broadband noise. This is accomplished with a classical maximum likelihood estimator, based on the integrated process (i.e., the angle, velocity, or position), rather than on the angular rate or acceleration as is standard in the literature. This trivial modification allows us to capture noise processes according to their effect on the integrated process, irrespective of their contribution to rate or acceleration noise. The cause of the noise is not discussed in this article. The method is tested on different classes of sensors by automatically identifying the parameters of a standard inertial sensor noise model. The results are analyzed qualitatively by comparing the model's Allan variance to the Allan variance computed directly from sensor data. A simulation that resembles one of the devices under test facilitates a quantitative analysis of the proposed estimator. Comparison with a competing, state-of-the-art method shows the advantages of the algorithm.**

***Index Terms*—Sensor phenomena and characterization, maximum likelihood estimation, gyroscope and accelerometer noise model.**

I. Introduction

INERTIAL sensors are employed in countless applications, ranging from consumer electronics to autonomous vehicles and unmanned aerial systems. Many different types of sensors are used, from low-cost, multi-axis microelectromechanical systems (MEMS) devices with a footprint below 20mm^2, to ring laser gyroscopes (RLGs). In order to understand the characteristics of these sensors, both stochastic ("noise") and deterministic (sensitivity, axes misalignment, etc.) errors have to be considered.

When inertial measurements are combined with data from other sensors, such as precise landmark observations from a camera system for example, an accurate model for the different sources of error in the inertial sensor data is vital. Such a model facilitates the optimal design of an estimator, and allows us to verify the proper operation of all software and hardware components.

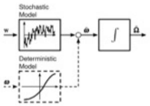


Fig. 1. Inertial sensor model with a deterministic and a random component. Here, the true angular rates ω are corrupted with deterministic errors, for example a scale factor that varies with temperature, as well as non-deterministic errors, such as additive broadband noise. This report presents a method to identify noise processes according to their contribution to the angular increments Ω̂.

A good sensor model often incorporates stochastic as well as deterministic components, as illustrated in Fig. 1. The stochastic part of the model includes errors such as broadband noise, or a slowly (randomly) varying bias, and describes them in a *probabilistic* sense. These errors can not be calibrated for, and limit the performance of the device fundamentally.

The instrumental noise figures are usually identified under static mechanical (i.e. non moving) and thermal conditions [1]. The stochastic model then provides an *upper bound* on the sensor performance, and crucial parameters for an estimator (i.e. noise densities).

This work focuses on how to identify such noise models automatically (for low-cost devices), since deriving a good noise model for an inertial sensor can be a difficult task for a non-expert. However, we also emphasise the importance of a good deterministic sensor model, since a stochastic model alone usually leads to over-confidence in the inertial data, especially when using uncalibrated sensors.

A. Existing Methods

Section II reviews the most commonly used stochastic models for inertial sensors. The Allan variance (AV) [2] and

Manuscript received August 2, 2015; revised September 1, 2015; accepted September 1, 2015. Date of publication September 3, 2015; date of current version December 10, 2015. This work was supported in part by the Armasuisse Science and Technology under Grant 8003501880 and in part by the Commission for Technology and Innovation under Grant 13394.1 PFFLENM. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amitava Chatterjee.

The authors are with the Autonomous Systems Laboratory, Swiss Federal Institute of Technology, Zurich CH-8092, Switzerland (e-mail: janosch.nikolic@mavt.ethz.ch; paul.furgale@mavt.ethz.ch; amir.melzer@mavt.ethz.ch; rsiegwart@ethz.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2015.2476668

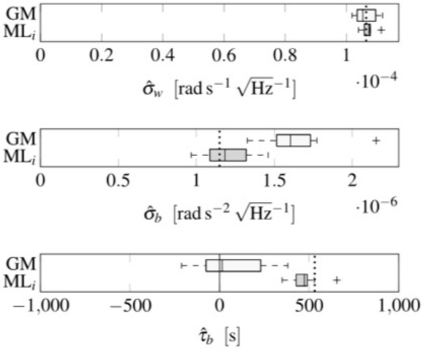


Fig. 11. Parameter estimates from two different methods: the generalized method of wavelet moments (GM) [7], [8], and integrated maximum likelihood (ML$_i$). The standard model parameters were estimated for 10 synthetic datasets, each 12 h long. Ground truth is indicated with a dotted line.

Can we do better than the GMWM?

Discussion on Maximum Likelihood-Based Methods for Inertial Sensor Calibration

Stéphane Guerrier, Roberto Molinari, and James Balamuta

***Abstract*—This letter highlights some issues which were overlooked in a recently published paper called maximum likelihood identification of inertial sensor noise model parameters. The latter paper does not consider existing alternative methods, which specifically tackle this issue in a possibly more direct manner and, although remaining a generally valid proposal, does not appear to improve on the earlier proposals. Finally, a simulation study rectifies the poor results of an estimator of reference in the same publication.**

***Index Terms*—State-space, calibration, time series, generalized method of wavelet moments.**

I. Introduction

The identification of a probabilistic model and the estimation of its relative parameters for the error signal issued from inertial sensors is a key challenge in many fields of engineering that has led to a great deal of research being produced. Recently, [1] (hereinafter *paper*) put forward a new algorithm for the automatic identification of the stochastic error model parameters. Considering other existing methods, the authors claim to improve the estimation of the classical maximum likelihood estimator by considering "the maximum likelihood estimator based on the integrated process rather than on the angular rate or acceleration as is standard in the literature". This proposal is indeed an interesting one but tackles a problem which can often be solved in a more straightforward and stable manner through a direct maximum likelihood approach. In this letter, we first of all discuss some critical points of the *paper* and then rectify some results which compare the algorithm proposed in the *paper* with the Generalized Method of Wavelet Moments (GMWM) introduced by [2]. For these discussions, we only consider the case of uniformly sampled observations but these analyses can be eventually extended to other types of sampling scenarios.

II. Discussion

A first observation which can be made on the *paper* is that the model put forward in it was already successfully estimated through the classical maximization of the likelihood function in [3], also applied within the framework of inertial sensors calibration. In the uniformly sampled setting, the simulation results in the *paper* do not seem to approach the results in [3] nor [4] which, moreover, are not cited nor considered for purposes of comparison. Indeed, the model considered in the *paper* is quite standard and can be defined as follows

$$\begin{aligned} X_t &= \phi X_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \\ Y_t &\sim \mathcal{N}(0, \gamma^2), \quad Z_t = X_t + Y_t \end{aligned} \tag{1}$$

where (X_t) is a first-order Gaussian autoregressive process (simply a reparametrization of a Gauss-Markov process), (Y_t) is a Gaussian white noise process and (Z_t) the observed process. Although limited in scope, this model can be considered as useful in the setting of inertial sensor calibration and for this reason was among those considered in [3] where it was shown that it can be estimated through classical maximum likelihood approaches which, however, diverge in even slightly more complex scenarios.

A second observation which can be made considering these results is that although the algorithm described in the *paper* proposes an innovative and effective approach for the estimation of this simple model, it does not appear to improve over existing methods (e.g. [3]) and can often be solved in a more straightforward manner. Indeed, in the *paper* (X_t) and (Y_t) are assumed to be independent and Z_t is multivariate Gaussian. In this setting, let $\mathbf{Z} = (Z_1, \ldots, Z_T)^T$, $\mathbf{Y} = (Y_1, \ldots, Y_T)^T$, $\mathbf{X} = (X_1, \ldots, X_T)^T$ and let us define $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{cov}(\mathbf{Z})$ where $\boldsymbol{\theta} = (\phi, \sigma^2, \gamma^2)^T$. We therefore have

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \gamma^2 \mathbf{I}_T + \frac{\sigma^2}{1-\phi^2}\left[\phi^{|i-j|}\right]_{i,j=1,\ldots,T}. \tag{2}$$

Considering this form, the log-likelihood function considered here, up to a constant, can be expressed as

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) - \frac{1}{2}\mathbf{Z}^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{Z}. \tag{3}$$

The maximization of (3) is straightforward in this case and does not require the use of a Kalman filter. Of course, with very large sample sizes such as those considered in the *paper*, this approach may not be as suitable as the inversion of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ entails a considerable computational burden. In this sense, the method proposed in the *paper* has a computational advantage over the maximization of (3) since it is based on a log sampling scheme. However, this implies a loss of efficiency which has yet to be quantified and thereby does not clarify the advantages of this proposal, for example, over a simple moment-based estimate which would deliver a faster and more reliable approach. An example of such moment-based estimator would be to consider the autocovariances at lags

Manuscript received February 5, 2016; accepted May 4, 2016. Date of publication May 9, 2016; date of current version June 16, 2016. The associate editor coordinating the review of this paper and approving it for publication was Prof. Julian C. C. Chan.

S. Guerrier and J. Balamuta are with the Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: stephane@illinois.edu; balamut2@illinois.edu).

R. Molinari is with the Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Geneva 1211, Switzerland (e-mail: roberto.molinari@unige.ch).

Digital Object Identifier 10.1109/JSEN.2016.2565389

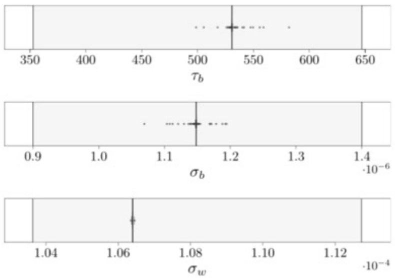


Fig. 1. Results for the simulation study described in Sec. III. Boxplots represent the results for the GMWM while the gray area represents the range of estimated values of the maximum likelihood approach presented in the *paper*.

Can we do better than the GMWM?

In short: Yes!

According to the type of model to be estimated, other estimators (if feasible) can estimate the parameters in a more statistically efficient manner.

Simulated Example

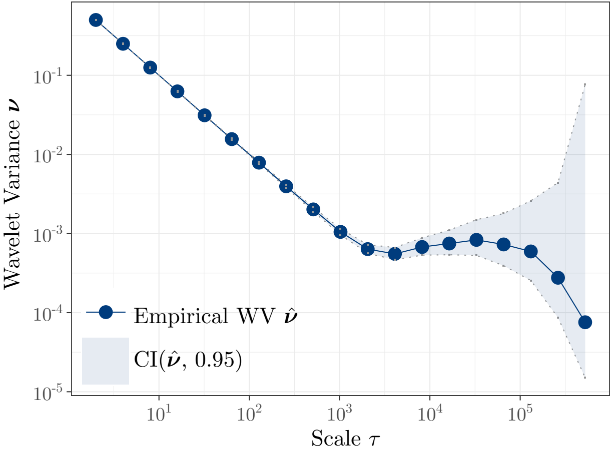
Let us consider a model made by a white noise (with parameter $\sigma^2 = 1$) and a first-order autoregressive model (with parameters $\phi = 0.9999$ and $\upsilon^2 = 10^{-6}$).

We simulate 100 time series of length $T = 1'000'000$ and estimate the parameters of the model using

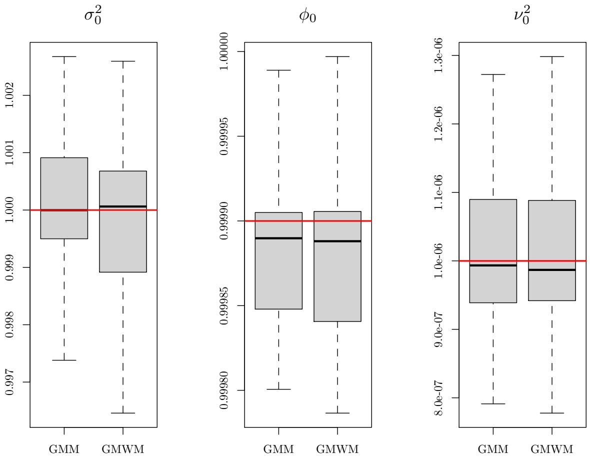
- GMWM
- GMM based on the first three lags of the autocovariance function ($\gamma(h)$ with $h = 0, \ldots, 3$)

Can we do better than the GMWM?

Below is the WV log-log plot of one of the simulated time series (NB it is similar to that of an IMU stochastic error signal)



Can we do better than the GMWM?



Although the GMM is slightly better for this model, the GMWM does not lose much efficiency and remains generally good and computationally efficient for many (complex) stationary and non-stationary models.

Conclusions

Summary

- The GMWM provides a **flexible modelling framework** for time series
  - A general, feasible and computationally efficient method with convenient asymptotic properties
  - The class of latent models covers frequently used models such as **structural time series and state-space models**
  - The results on **identifiability** allow to greatly weaken conditions for consistency of extremum estimators for a wide class of time series models
- Development and implementation of **the R `gmwm` package and online apps** that makes these tools available to researchers

Conclusions

Outlook

The GMWM framework provide sound statistical bases for a *computationally efficient inference* concerning

- **multivariate time series** and **mixed-effect models**
- **non-stationary time series** (i.e. time-varying parameters)
- **model selection** for time series models
- **robust tests and inference** for time series using the WV

Moreover, the GMWM computational platform will continue to be developed and updated in order to make all these results available.

References

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Guerrier, S., R. Molinari, and J. Skaloud (2015). "Automatic Identification and Calibration of Stochastic Parameters in Inertial Sensors". In: *Navigation* 62.4.

Guerrier, S., R. Molinari, and Y. Stebler (2016). "Theoretical Limitations of Allan Variance-based Regression for Time Series Model Estimation". In: *IEEE Signal Processing Letters* 23.5, pp. 597–601.

Newey, Whitney K and Kenneth D West (1986). *A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix*.

Percival, D. B. and P. Guttorp (1994). "Long-memory processes, the Allan variance and wavelets". In: *Wavelet Analysis and its Applications*. Vol. 4. Elsevier, pp. 325–344.

Percival, D. P. (1995). "On Estimation of the Wavelet Variance". In: *Biometrika* 82.3, pp. 619–631.

Radi, A et al. (2017). "An Automatic Calibration Approach for the Stochastic Parameters of Inertial Sensors". In: *Proceedings of the 30th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2017)*. IEEE, pp. 3053 –3060.

Contacts

Any questions?

More info...

SMAC-group.com
github.com/SMAC-Group
stephane@psu.edu

Appendix A: Examples on Strong and Weak Stationarity

Example 1 (Random walk): Consider a Gaussian random walk process X_t, as defined in 2.6 where initial value $X_0 = 0$. Since the model is "fixed in time" this process is clearly strongly stationary but not weakly stationary since

$$\text{var}(X_t) = \text{var}\left(\sum_{i=1}^{t} Z_t\right) = \sum_{i=1}^{t} \gamma^2 = t\gamma^2$$

and therefore $\mathbb{E}[X_t^2]$ does not exist. ●

Example 2 (mixtures): Let $X_t \overset{iid}{\sim} \exp(1)$ (i.e. exponential distribution with $\lambda = 1$) and $Y_t \overset{iid}{\sim} \mathcal{N}(1, 1)$. Then, let

$$Z_t = \begin{cases} X_t & \text{if } t \in \{2k|k \in \mathbb{N}\} \\ Y_t & \text{if } t \in \{2k+1|k \in \mathbb{N}\}. \end{cases}$$

Then, Z_t is weakly stationary but not strongly stationary. ●

▸ Return to Remark 11

Appendix B: Example: Weak Stationarity of an AR1

Consider an AR1 process (see Definition 2.7), defined as:

$$X_t = \phi X_{t-1} + Z_t, \quad Z_t \overset{iid}{\sim} \mathcal{N}(0, \nu^2),$$

with $|\phi| < 1$ and $\nu^2 < \infty$. Then, we have

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t = \phi\left[\phi X_{t-2} + Z_{t-1}\right] + Z_t = \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\ &\vdots \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j Z_{t-j}. \end{aligned}$$

By taking the limit in k (which is perfectly valid as we assume $t \in \mathbb{Z}$) and assuming $|\phi| < 1$, we obtain

$$X_t = \lim_{k\to\infty} X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}, \tag{29}$$

since $|\phi| < 1$. This shows that this process is linear (see Definition 2.17).

Appendix B: Example: Weak Stationarity of an AR1

Therefore, we have

$$\mathbb{E}\left[X_t\right] = \sum_{j=0}^{\infty} \phi^j \mathbb{E}[Z_{t-j}] = 0$$

$$\text{var}\left(X_t\right) = \text{var}\left(\sum_{j=0}^{\infty} \phi^j Z_{t-j}\right) = \sum_{j=0}^{\infty} \phi^{2j}\, \text{var}\left(Z_{t-j}\right) = \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} = \frac{\nu^2}{1-\phi^2}.$$

Moreover, we obtain (assuming for notational simplicity that $|h| > 1$

$$\text{cov}\left(X_t, X_{t+h}\right) = \phi\, \text{cov}\left(X_t, X_{t+h-1}\right) = \phi^2\, \text{cov}\left(X_t, X_{t+h-2}\right) = \phi^h\, \text{cov}(X_t, X_t).$$

When $h \in \mathbb{Z}$ we obtain

$$\text{cov}\left(X_t, X_{t+h}\right) = \phi^{|h|}\, \text{cov}(X_t, X_t) = \phi^{|h|} \frac{\nu^2}{1-\phi^2},$$

thus verifying the weak stationarity of the process.

Return to Remark 12

Appendix C: Derivation of Equation (7)

The derivation of (7) allows us to introduce a common technique used in time series analysis. Let $\mathbf{1}$ denotes a unit vector of dimension T and $\mathbf{X} = [X_1, \ldots, X_T]^T$, then we notice that $\bar{X}$ can be expressed as follows

$$\bar{X} = \frac{1}{T}\mathbf{1}^T\mathbf{X}.$$

Moreover, we remember that if $\mathbf{Y} \in \mathbb{R}^k$ and a random variable and $\mathbf{A} \in \mathbb{R}^{h \times k}$ a fixed matrix we have

$$\text{var}\left(\mathbf{AY}\right) = \mathbf{A}\,\text{var}\left(\mathbf{Y}\right)\mathbf{A}^T.$$

Therefore, we have

$$\begin{aligned}
\text{var}\left(\bar{X}\right) &= \frac{1}{T^2}\,\text{var}\left(\mathbf{1}^T\mathbf{X}\right) = \frac{1}{T^2}\mathbf{1}^T\,\text{var}\left(\mathbf{X}\right)\mathbf{1} \\
&= \frac{1}{T^2}\mathbf{1}^T \begin{bmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(1) & \gamma(0) & & \gamma(T-2) \\ \vdots & & \ddots & \vdots \\ \gamma(T-1) & \dots & \dots & \gamma(0) \end{bmatrix} \mathbf{1}.
\end{aligned}$$

Appendix C: Derivation of Equation (7)

By looking at the matrix $\text{var}\left(\mathbf{X}\right)$ one can notice that it contains T times the term $\gamma(0)$, $2(T-1)\,\gamma(1)$, $2(T-2)\,\gamma(2)$ and so on. Therefore, we have

$$\text{var}\left(\bar{X}\right) = \frac{1}{T^2}\left(T\gamma(0) + 2(T-1)\gamma(1) + 2(T-2)\gamma(2) + \ldots + \gamma(T-1)\right).$$

Since $\gamma(h)$ is symmetric we also have

$$\text{var}\left(\bar{X}\right) = \frac{1}{T^2}\sum_{i=T}^{T}\left(T - |i|\right)\gamma(i) = \frac{\gamma(0)}{T}\sum_{i=T}^{T}\left(1 - \frac{|i|}{T}\right)\rho(i).$$

It is worth noting that in the iid case we have that $\text{var}\left(\bar{X}\right) = \frac{1}{T}\,\text{var}(X_1)$. This result can naturally also be obtained using (7) since $\text{var}(X_1) = \gamma(0)$ and $\sum_{i=T}^{T}\left(1 - \frac{|i|}{T}\right)\rho(i) = 1$.

Return to Equation (7)

Appendix D: How to compute $\text{var}\left(\bar{X}\right)$ in practice?

As in the previous example, let us consider a stationary AR1 process, i.e.

$$X_t = \phi X_{t-1} + Z_t, \quad \text{where} \quad |\phi| < 1 \quad \text{and} \quad Z_t \overset{iid}{\sim} \mathcal{N}\left(0, \nu^2\right)$$

We already showed in Appendix B that $\gamma(h) = \phi^h \sigma^2 \left(1 - \phi^2\right)^{-1}$, therefore, we obtain (after some computations):

$$\text{var}\left(\bar{X}\right) = \frac{\nu^2 \left(T - 2\phi - T\phi^2 + 2\phi^{T+1}\right)}{T^2 \left(1 - \phi^2\right) \left(1 - \phi\right)^2}. \tag{30}$$

Unfortunately, deriving such an exact formula is often difficult when considering more complex models. However, asymptotic approximations are often employed to simplify the calculation. For example, in our case we have

$$\lim_{T \to \infty} T \, \text{var}\left(\bar{X}\right) = \frac{\nu^2}{\left(1 - \phi\right)^2},$$

providing the following approximate formula:

$$\text{var}\left(\bar{X}\right) \approx \frac{\nu^2}{T \left(1 - \phi\right)^2}.$$

Appendix D: How to compute $\text{var}\left(\bar{X}\right)$ in practice?

Alternatively, simulation methods can also be employed. For example, one could compute $\text{var}\left(\bar{X}\right)$ as follows:

Step 1: Simulate under the assumed model, i.e. $X_t^* \sim F_{\boldsymbol{\theta}_0}$, where $F_{\boldsymbol{\theta}_0}$ denotes the true model (in this case an AR1 process).

Step 2: Compute $\bar{X}^*$ (i.e. average based on (X_t^*)).

Step 3: Repeat Steps 1 and 2 B times.

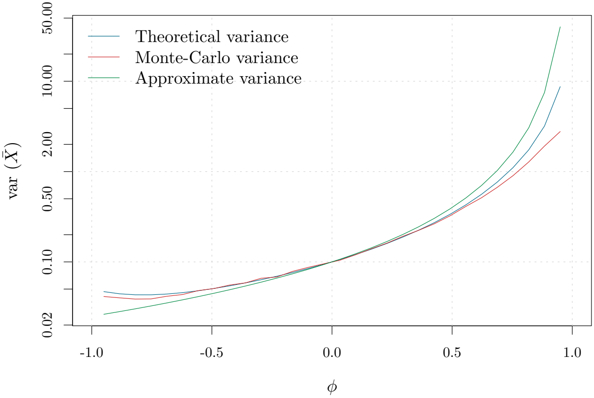
Step 4: Compute the empirical variance $\bar{X}^*$ (based on B independent replications).

The above procedure is known as Monte-carlo method (it is actually a Monte-carlo integral) and is closely related to the concept of parametric bootstrap (see Efron and Tibshirani 1994) which is a very popular tool in statistics.

Appendix D: How to compute $\text{var}\left(\bar{X}\right)$ in practice?

A numerical experiment

We consider $T = 10$, $B = 5000$ and grid of values for ϕ from -0.95 to 0.95.



Return to Equation (7)

Appendix E: Consistency - a simple example

Let $X_i \overset{iid}{\sim} \mathcal{E}(\lambda_0),\ \lambda_0 \in \mathbb{R}^+,\ i = 1, ..., n$. We wish to show that the MLE for λ_0 is consistent. Then, we have that the density of X is given by (assuming $X \geq 0$):

$$f(x|\lambda) = \lambda \exp\left(-\lambda x\right).$$

Therefore, the normalized log-likelihood function is given by

$$\mathcal{L}(\lambda|X_1, ..., X_n) = \log(\lambda) - \lambda \bar{X}_n,$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. By taking the first derivative we obtain:

$$\frac{\partial}{\partial \lambda}\,\mathcal{L}(\lambda|X_1, ..., X_n) = \frac{1}{\lambda} - \bar{X}_n,$$

which implies that MLE is such that $\frac{\partial}{\partial \lambda}\,\mathcal{L}(\hat{\lambda}|X_1, ..., X_n) = 0$. We obtain

$$\frac{1}{\hat{\lambda}} - \bar{X}_n = 0 \implies \hat{\lambda} = \bar{X}_n.$$

Appendix E: Consistency - a simple example

Finally, we verify that

$$\frac{\partial^2}{\partial \lambda^2} \mathcal{L}(\lambda | X_1, ..., X_n) = -\frac{1}{\lambda^2} < 0,$$

implying that $\hat{\lambda}$ is the maxima of $\mathcal{L}(\lambda | X_1, ..., X_n)$. Therefore, the MLE is a function of the sample mean $\bar{X}_n$. In this case, the consistency of $\hat{\lambda}$ is implied by Theorems 3.2 and 3.3. Indeed, it follows from the Weak Law of Large Number (i.e. Theorem 3.2) that $\bar{X}_n \overset{\mathcal{P}}{\longmapsto} \mu$, where μ is given by

$$\mu = \mathbb{E}[X_i] = \int_0^{\infty} x\lambda_0 \exp\left(-\lambda_0 x\right) dx = \frac{1}{\lambda_0}.$$

Since the function $f(x) = 1/x$ is continuous in $\mathbb{R}^+$ we obtain by the Continuous Mapping Theorem (i.e. Theorem 3.3) that $\hat{\lambda} \overset{\mathcal{P}}{\longmapsto} \lambda_0$, which concludes our example. ●

▸ Return to Theorem 3.2

Appendix F: Proof of Theorem 3.4

Let $\mathcal{G}$ be the ε-ball centered at $\boldsymbol{\theta}_0$ i.e. $\mathcal{G} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||_2 < \varepsilon\}$ for some $\varepsilon > 0$. Then $\hat{\boldsymbol{\theta}} \xmapsto{\mathcal{P}} \boldsymbol{\theta}_0$ is equivalent to

$$\lim_{n\to\infty} \Pr(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \geq \varepsilon) = \lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) = 0.$$

We define $\gamma = Q(\boldsymbol{\theta}_0) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\mathcal{G}} Q(\boldsymbol{\theta})$ which is strictly positive by 3.4. Then we have that $\hat{\boldsymbol{\theta}} \notin \mathcal{G}$ implies

$$Q(\hat{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\setminus\mathcal{G}} Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}_0) - \gamma$$

and therefore

$$\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) \leq \lim_{n\to\infty} \Pr(\mathcal{A})$$

where $\mathcal{A} = \left\{Q(\hat{\boldsymbol{\theta}}) \leq Q(\boldsymbol{\theta}_0) - \gamma\right\}$.

Appendix F: Proof of Theorem 3.4

Next, we define the following events:

$$\begin{aligned}
\mathcal{B} &= \left\{ \left| Q_n(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}}) \right| > \gamma/3 \right\} \\
\mathcal{C} &= \left\{ \left| Q_n(\boldsymbol{\theta}_0) - Q(\boldsymbol{\theta}_0) \right| > \gamma/3 \right\} \\
\mathcal{D} &= \left\{ \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}) \right| > \gamma/3 \right\}
\end{aligned}$$

and we have that

$$\begin{aligned}
\Pr(\mathcal{A}) \leq \Pr(\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C})) &= \Pr\left(\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C})^c\right) + \Pr(\mathcal{B} \cup \mathcal{C}) \\
&= \Pr(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c) + \Pr(\mathcal{B} \cup \mathcal{C}).
\end{aligned}$$

Appendix F: Proof of Theorem 3.4

It is easy to verify that $\Pr\left(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c\right) = \Pr(\varnothing)$ because if $\mathcal{A}$, $\mathcal{B}^c$ and $\mathcal{C}^c$ occur simultaneously we have that

$$
\begin{aligned}
Q_n(\hat{\boldsymbol{\theta}}) &= Q(\hat{\boldsymbol{\theta}}) + Q_n(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}}) \leq Q(\hat{\boldsymbol{\theta}}) + |Q_n(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}})| \\
&\leq Q(\hat{\boldsymbol{\theta}}) + \gamma/3 \leq Q(\boldsymbol{\theta}_0) - 2\gamma/3 = Q(\boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_0) + Q_n(\boldsymbol{\theta}_0) - 2\gamma/3 \\
&\leq |Q(\boldsymbol{\theta}_0) - Q_n(\boldsymbol{\theta}_0)| + Q_n(\boldsymbol{\theta}_0) - 2\gamma/3 \leq Q_n(\boldsymbol{\theta}_0) - \gamma/3 < Q_n(\boldsymbol{\theta}_0)
\end{aligned}
$$

which contradicts (8). Moreover, the probability $\Pr\left(\mathcal{B} \cup \mathcal{C}\right)$ can be bounded as follow

$$
\begin{aligned}
\Pr\left(\mathcal{B} \cup \mathcal{C}\right) &= \Pr\left(\sup_{\boldsymbol{\theta} \in \left\{\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}\right\}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \gamma/3\right) \\
&\leq \Pr\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \gamma/3\right) = \Pr\left(\mathcal{D}\right).
\end{aligned}
$$

Appendix F: Proof of Theorem 3.4

Finally, we have that

$$\begin{aligned}\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\theta}} \notin \mathcal{G}) \leq \lim_{n\to\infty} \Pr(\mathcal{A}) &\leq \lim_{n\to\infty} \Pr\left(\mathcal{A} \cap \mathcal{B}^c \cap \mathcal{C}^c\right) + \Pr\left(\mathcal{B} \cup \mathcal{C}\right)\\ &= \Pr(\varnothing) + \lim_{n\to\infty} \Pr\left(\mathcal{B} \cup \mathcal{C}\right) \leq \lim_{n\to\infty} \Pr(\mathcal{D}) = 0,\end{aligned}$$

which concludes the proof. ■

Return to Sketch of the proof

Appendix G: Proof of Lemma 3.5

Since $\mathbf{W} > 0$ is nonsingular and there exists a unique $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $\mathbf{g}_0(\boldsymbol{\theta}_0) = \mathbf{0}$, for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we have

$$Q_0(\boldsymbol{\theta}) = -\mathbf{g}_0(\boldsymbol{\theta})^T \mathbf{W} \mathbf{g}_0(\boldsymbol{\theta}) \leq -||\mathbf{g}_0(\boldsymbol{\theta}_0)^T \mathbf{W} \mathbf{g}_0(\boldsymbol{\theta}_0)|| = Q_0(\boldsymbol{\theta}_0).$$

Therefore, $Q_0(\boldsymbol{\theta})$ has a unique maximum (0) at $\boldsymbol{\theta}_0$. ■

Return to Lemma 3.5

Appendix H: Example - Proving Condition (C.4)

We return to the example presented in ▸ Example GMM, we define the following quantities:

$$\hat{Q}_n(\boldsymbol{\theta}) = -||\hat{\gamma} - \gamma(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}}.$$

where we assumed that $\hat{\mathbf{W}} \xmapsto{\mathcal{P}} \mathbf{W}$,

Since we have $Z_i \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$, we let

$$\mathbf{X}_i \equiv \begin{bmatrix} Z_i \\ Z_i^2 \\ Z_i^3 \end{bmatrix},$$

and therefore by the (weak) Law of Large Number (see Theorem 3.2), we have

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \xmapsto{\mathcal{P}} \mathbb{E}[\mathbf{X}_i] = \gamma(\boldsymbol{\theta}_0).$$

Next, we define $Q_0(\boldsymbol{\theta}) = -||\gamma_0 - \gamma(\boldsymbol{\theta})||^2_{\mathbf{W}}$ which the same of $\hat{Q}_n(\boldsymbol{\theta})$ when replacing all the random quantities by the limiting term.

Appendix H: Example - Proving Condition (C.4)

We are now interested in showing that the difference $\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})$ converge in probability to 0. For ease of notation, we will define $\gamma(\boldsymbol{\theta}_0)$ as γ_0:

$$\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}) \leq |\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| = \left|\underbrace{||\hat{\gamma} - \gamma(\boldsymbol{\theta})||^2_{\widehat{\mathbf{W}}} - ||\gamma_0 - \gamma(\boldsymbol{\theta})||^2_{\mathbf{W}}}_{\mathbf{A}}\right|.$$

Developing the above expression **A** inside the absolute value, and defining without loss of generality that $\mathbf{W}^* = \widehat{\mathbf{W}} - \mathbf{W}$ where $\mathbf{W}$ is a symmetric matrix, we get:

$$\begin{aligned}
\mathbf{A} =& \hat{\gamma}^T\widehat{\mathbf{W}}\hat{\gamma} + \gamma(\boldsymbol{\theta})^T\widehat{\mathbf{W}}\gamma(\boldsymbol{\theta}) - 2\hat{\gamma}^T\widehat{\mathbf{W}}\gamma(\boldsymbol{\theta}) - \gamma_0^T\mathbf{W}\gamma_0 - \gamma(\boldsymbol{\theta})^T\mathbf{W}\gamma(\boldsymbol{\theta}) + 2\gamma_0^T\mathbf{W}\gamma(\boldsymbol{\theta}) \\
=& \|\hat{\gamma} - \gamma_0\|^2_{\widehat{\mathbf{W}}} - \gamma_0^T\widehat{\mathbf{W}}\gamma_0 + 2\hat{\gamma}^T\widehat{\mathbf{W}}\gamma_0 - 2\hat{\gamma}^T\widehat{\mathbf{W}}\gamma(\boldsymbol{\theta}) + \gamma(\boldsymbol{\theta})^T\widehat{\mathbf{W}}\gamma(\boldsymbol{\theta}) \\
&- \gamma_0^T\mathbf{W}\gamma_0 - \gamma(\boldsymbol{\theta})^T\mathbf{W}\gamma(\boldsymbol{\theta}) + 2\gamma_0^T\mathbf{W}\gamma(\boldsymbol{\theta}) \\
=& \|\hat{\gamma} - \gamma_0\|^2_{\widehat{\mathbf{W}}} + \|\gamma_0 - \gamma(\boldsymbol{\theta})\|^2_{\mathbf{W}^*} + 2\gamma(\boldsymbol{\theta})\widehat{\mathbf{W}}\gamma_0 - 2\gamma_0^T\widehat{\mathbf{W}}\gamma_0 \\
&- 2\hat{\gamma}^T\widehat{\mathbf{W}}\gamma(\boldsymbol{\theta}) + 2\gamma_0^T\mathbf{W}\gamma(\boldsymbol{\theta}) \\
=& \|\hat{\gamma} - \gamma_0\|^2_{\widehat{\mathbf{W}}} + \|\gamma_0 - \gamma(\boldsymbol{\theta})\|^2_{\mathbf{W}^*} + 2\left(\gamma_0 - \gamma(\boldsymbol{\theta})\right)^T\widehat{\mathbf{W}}\left(\hat{\gamma} - \gamma_0\right).
\end{aligned}$$

Appendix H: Example - Proving Condition (C.4)

Therefore using triangular inequality, we can write the following:

$$|\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \leq \left|\underbrace{\|\gamma(\boldsymbol{\theta}_0) - \gamma(\boldsymbol{\theta})\|^2_{\mathbf{W}^*}}_{\equiv a_1}\right| + \left|\underbrace{\|\hat{\gamma} - \gamma(\boldsymbol{\theta}_0)\|^2_{\widehat{\mathbf{W}}}}_{\equiv a_2}\right|$$

$$+ \left|\underbrace{2\left(\gamma(\boldsymbol{\theta}_0) - \gamma(\boldsymbol{\theta})\right)^T \widehat{\mathbf{W}}\left(\hat{\gamma} - \gamma(\boldsymbol{\theta}_0)\right)}_{\equiv a_3}\right|.$$

Before continuing, suppose that $\boldsymbol{x}$ is a vector and $\mathbf{W}$ the above defined matrix, we have that $\boldsymbol{x}^T\mathbf{W}\boldsymbol{x} \leq \lambda_1\|\boldsymbol{x}\|^2$, where λ_1 is then largest eigenvalue of $\mathbf{W}$. Furthermore, lets also define the Frobnius norm of a matrix as $\|\mathbf{W}\| = (\sum_i^N \sum_j^J w_{i,j}^2)^{\frac{1}{2}}$ and $\|\mathbf{W}\|^2 = \sigma_{max} \leq \|\mathbf{W}\|$, where σ_{max} is the largest singular value of $\|\mathbf{W}\|$.

Example: Proving Condition (C.4)

Using those properties, we can investigate the terms in the above equation. Considering a_1 , we have that:

$$a_1 \leq \|\gamma_0 - \gamma(\boldsymbol{\theta})\|^2 \lambda_1 \leq \|\gamma_0 - \gamma(\boldsymbol{\theta})\|^2 \|\mathbf{W}^*\|$$
$$= \sum_{i=1}^{3} (\gamma_{0_i} - \gamma_i(\boldsymbol{\theta}))^2 \sqrt{\sum_{i=1}^{3} \sum_{j=1}^{3} (\hat{w}_{i,j} - w_{i,j})^2}.$$

By Conditions (C.1) to (C.3), $\sum_{i=1}^{3} (\gamma_{0_i} - \gamma_i(\boldsymbol{\theta}))^2 \leq 3 \max_i (\gamma_{0_i} - \gamma_i(\boldsymbol{\theta}))^2$ which is bounded.

By the previously mentioned on $\widehat{\mathbf{W}}$, we can also see that

$$\sqrt{\sum_{i=1}^{3} \sum_{j=1}^{3} (\hat{w}_{i,j} - w_{i,j})^2} \leq 3 \max_i \sqrt{(\hat{w}_{i,j} - w_{i,j})^2} \xrightarrow{\mathcal{P}} 0.$$

Appendix H: Example - Proving Condition (C.4)

Investigating the convergence of $\mathbf{a2} \leq \|\hat{\gamma} - \gamma_0\|^2 \|\widehat{\mathbf{W}}\|$, by the same conditions (C.1) to (C.3), and that the sample moments will converge to the population one, we have that:

$$a_2 \overset{\mathcal{P}}{\longmapsto} 0.$$

Finally, considering the previous results on a_1 and a_2, we can see that the last term a_3 tends also to 0, therefore, we can say that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$|\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \overset{\mathcal{P}}{\longmapsto} 0.$$

●

Return to the discussion on (C.4)

Appendix I: Proof of Theorem 3.8

Since $\bar{X}_n - \mu$ has mean zero, its variance is

$$\mathbb{E}[(\bar{X}_n - \mu)^2] = 1/n \sum_{k=-\infty}^{\infty} \zeta_k \leq 1/n \sum_{k=-\infty}^{\infty} |\zeta_k| \leq C/n.$$

By Chebychev's inequality, for all $\varepsilon > 0$

$$\Pr\left(|\bar{X}_n - \mu| \geq \varepsilon\right) \leq \frac{\mathbb{E}[(\bar{X}_n - \mu)^2]}{\varepsilon^2} \leq \frac{C}{\varepsilon^2 n},$$

so that

$$\lim_{n\to\infty} \Pr\left(|\bar{X}_n - \mu| \geq \varepsilon\right) \to 0,$$

which conclude the proof. $\blacksquare$

Return to Theorem 3.8

Appendix J: Proof of Corollary 3.9

The proof of the result is direct from Theorem 3.8. (X_t^2) has an absolutely summable covariance structure implies both (X_t) and $(X_t X_{t-h})$ for all $h \in \mathbb{Z}$ have absolutely summable covariance structure. Under the conditions that (X_t) is weakly stationary and has absolutely summable covariance structure, by Theorem 3.9, we have $\bar{X}_T \overset{\mathcal{P}}{\longmapsto} \mu$. Let

$$\tilde{\gamma}(h) = \frac{1}{T} \sum_{t=h+1}^{T} (X_t - \mu)(X_{t-h} - \mu).$$

Since $((X_t - \mu)(X_{t-h} - \mu))$ is stationary and has absolutely summable covariance structure, by Theorem 3.9 again, we have $\tilde{\gamma}(h) \overset{\mathcal{P}}{\longmapsto} \gamma(h)$. We can also show that $\sqrt{T}\,(\tilde{\gamma}(h) - \hat{\gamma}(h)) = o_p(1)$. Therefore, we obtain

$$\hat{\gamma}(h) \overset{\mathcal{P}}{\longmapsto} \gamma(h).$$

Appendix J: Proof of Corollary 3.9

Similarly, we can show that

$$\left(\hat{\gamma}(0), \hat{\gamma}(h)\right)^T \xmapsto{\mathcal{P}} \left(\gamma(0), \gamma(h)\right)^T.$$

Then by Theorem 3.3, we have

$$\hat{\rho}(h) \xmapsto{\mathcal{P}} \rho(h),$$

which conclude the proof. ∎

▸ Return to Corollary 3.9

Appendix K: Example on the Theoretical AV of an MA(1)

In this example we will derive the theoretical AV of an MA(1) process, which is defined as

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1}, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

where $|\theta| < 1$ and $\sigma^2 < \infty$. The variance of this process is given by

$$\sigma_X^2 = \text{var}\left(X_t\right) = \text{var}\left(\varepsilon_t\right) + \theta^2 \, \text{var}\left(\varepsilon_{t-1}\right) = \left(1 + \theta^2\right)\sigma^2.$$

Similarly, the autocovariance is given by

$$\gamma(h) = \text{cov}\left(X_t, X_{t-h}\right) = \begin{cases} \sigma_X^2 & \text{if } h = 0 \\ \theta\sigma^2 & \text{if } |h| = 1 \\ 0 & \text{if } |h| > 1. \end{cases}$$

Appendix K: Example on the Theoretical AV of an MA(1)

From the autocovariance we obtain

$$\rho(h) = \text{corr}\left(X_t, X_{t-h}\right) = \begin{cases} 1 & \text{if } h = 0 \\ \frac{\theta}{1+\theta^2} & \text{if } |h| = 1 \\ 0 & \text{if } |h| > 1. \end{cases}$$

We can now apply the formula given in Lemma 4.5, which leads to

$$\begin{aligned} \text{AVar}_j\left(X_t\right) &= \frac{\left(1+\theta^2\right)\sigma^2}{\tau_j^2}\left(\tau_j + \sum_{i=1}^{\tau_j - 1} i\left[2\rho(\tau_j - i) - \rho(i) - \rho(2\tau_j - i)\right]\right) \\ &= \frac{\left(1+\theta^2\right)\sigma^2}{\tau_j^2}\left(\tau_j + 2\sum_{i=1}^{\tau_j - 1} i\rho(\tau_j - i) - \sum_{i=1}^{\tau_j - 1} i\rho(i) - \sum_{i=1}^{\tau_j - 1} i\rho(2\tau_j - i)\right) \\ &= \frac{\left(1+\theta^2\right)\sigma^2}{\tau_j^2}\left(\tau_j + 2(\tau_j - 1)\rho(1) - \rho(1)\right) \\ &= \frac{\left(1+\theta^2\right)\sigma^2}{\tau_j^2}\left(\tau_j + (2\tau_j - 3)\frac{\theta}{1+\theta^2}\right). \end{aligned}$$

Return to the Properties of the AV

Appendix L: Proof of Lemma 4.7

The proof of the result is direct from Theorem 3.8. Let $Z_t = X_t - X_{t-1}$, then since Z_t is stationary with mean zero then so is $(X_t - X_{t-h})$ for all $h \in \mathbb{Z}$. This directly imply that $\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)}$ is also stationary (since it based on a linear combination of stationary processes) and so is $Y_t \equiv (\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)})^2$ (since it is based on a time invariant transformation of a stationary process). Moreover, there exist constant c_h such that

$$\sum_{h=-\infty}^{\infty} \gamma_Y(h) = \sum_{h=-\infty}^{\infty} c_{|h|}\gamma_{Z^2}(h).$$

Therefore, we obtain

$$\sum_{h=-\infty}^{\infty} |\gamma_Y(h)| = \sum_{h=-\infty}^{\infty} |c_{|h|}\gamma_{Z^2}(h)| \leq \sup_{k=1,\ldots,\infty} c_k \sum_{h=-\infty}^{\infty} |\gamma_{Z^2}(h)| < \infty,$$

since both terms are bounded. Using the same approach we have that $\mathbb{E}\left[Y_t^2\right]$ is bounded since $\mathbb{E}\left[Z_t^4\right]$ bounded. Thus, we can apply Theorem 3.8 on the process Y_t, i.e.

$$\widehat{\text{AVar}}_j\left(X_t\right) = \frac{1}{2}\bar{Z}_t \xrightarrow{\mathcal{P}} \frac{1}{2}\mathbb{E}[Z_t] = \text{AVar}_j\left(X_t\right),$$

which concludes the proof. ∎

▸ Return to Lemma 4.7

Appendix M: Proof of Lemma 4.8

The proof of the result is direct from Theorem 3.12. Let $Z_t = X_t - X_{t-1}$, then since Z_t is stationary with mean zero then so is $(X_t - X_{t-h})$ for all $h \in \mathbb{Z}$. This directly imply that $\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)}$ is also stationary (since it based on a linear combination of stationary processes) and so is $Y_t \equiv (\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)})^2$ (since it is based on a time invariant transformation of a stationary process). And since (Y_t) is a borel-measurable function of Z_t, we have (Y_t) is also a strong mixing process with mixing coefficient $\alpha^*(n) \leq \alpha(n)$, hence $\sum_{n=1}^{\infty} \alpha^*(n)^{\delta/2+\delta} < \infty$ for some $\delta > 0$. Moreover, since $(\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)})$ is a linear function of Z_t, and $\mathbb{E}\left[Z_t^{4+\delta}\right] < \infty$, by triangle inequality, we have $\mathbb{E}\left[(\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)})^{4+\delta}\right] < \infty$. Thus, we can apply Theorem 3.12 on the process Y_t, i.e.

$$\sqrt{T}\left(\widehat{\text{AVar}}_j\left(X_t\right) - \text{AVar}_j\left(X_t\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_T^2/T),$$

where $\sigma_T^2 \equiv \sum_{h=-\infty}^{\infty} \text{cov}\left(\left(\bar{X}_t^{(j)} - \bar{X}_{t-\tau_j}^{(j)}\right)^2, \left(\bar{X}_{t+h}^{(j)} - \bar{X}_{t+h-\tau_j}^{(j)}\right)^2\right)$, which concludes the proof. ■

▸ Return to Lemma 4.8

Appendix N: Derivation of Equation (26)

The derivation of the AV estimator closed form solution is pretty straightforward. From (25), we just first take the first derivative with respect to $\boldsymbol{\theta}$, which yields:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{\tau_j \in \boldsymbol{\eta}} \left[\log\left(\hat{\phi}_{\tau_j}\right) - g\left(\theta\right) - \lambda \log\left(\tau_j\right)\right]^2 = 2 \sum_{\tau_j \in \boldsymbol{\eta}} \left[\log\left(\hat{\phi}_{\tau_j}\right) - g\left(\theta\right) - \lambda \log\left(\tau_j\right)\right] g'\left(\theta\right) = 0$$

$$\Leftrightarrow \frac{1}{|\boldsymbol{\eta}|} g'\left(\theta\right) g\left(\theta\right) = g'\left(\theta\right) \sum_{\tau_j \in \boldsymbol{\eta}} \left[\log\left(\hat{\phi}_{\tau_j}\right) - \lambda \log\left(\tau_j\right)\right]$$

$$\Leftrightarrow \hat{\boldsymbol{\theta}}_{AV} = g^{-1} \left\{ \frac{1}{|\boldsymbol{\eta}|} \sum_{\tau_j \in \boldsymbol{\eta}} \left[\log\left(\hat{\phi}_{\tau_j}\right) - \lambda \log\left(\tau_j\right)\right] \right\}.$$

▸ Return to Equation (26)

Proof of Lemma 5.5

The proof of the element-wise consistency is direct from Theorem 3.8. Let $Z_t = X_t - X_{t-1}$, then since Z_t is stationary with mean zero then so is $(X_t - X_{t-h})$ for all $h \in \mathbb{Z}$. This directly imply that for all $j = 1, \dots, J$, $(W_{j,t})$ is also stationary (since it based on a linear combination of stationary processes) and so is $Y_t = W_{j,t}^2$ (since it is based on a time invariant transformation of a stationary process). Moreover, there exist constant c_h such that

$$\sum_{h=-\infty}^{\infty} \gamma_Y(h) = \sum_{h=-\infty}^{\infty} c_{|h|} \gamma_{Z^2}(h).$$

Therefore, we obtain

$$\sum_{h=-\infty}^{\infty} |\gamma_Y(h)| = \sum_{h=-\infty}^{\infty} |c_{|h|} \gamma_{Z^2}(h)| \leq \sup_{k=1,\dots,\infty} c_k \sum_{h=-\infty}^{\infty} |\gamma_{Z^2}(h)| < \infty,$$

since both terms are bounded. Using the same approach we have that $\mathbb{E}\left[Y_t^2\right]$ is bounded since $\mathbb{E}\left[Z_t^4\right]$ bounded. Thus, we can apply Theorem 3.8 on the process Y_t, i.e.

$$\hat{\nu}_j^2 = \frac{1}{M_j(T)} \sum_{t=L_j}^{T} W_{j,t}^2 \xrightarrow{\mathcal{P}} \nu_j^2.$$

Proof of Lemma 5.5

Since

$$\hat{\nu}_j^2 \xmapsto{\mathcal{P}} \nu_j^2 \,,$$

and given that J is a bounded quantity, we have that

$$\begin{aligned} \|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_2 &= \sum_{j=1}^{J} (\hat{\nu}_j^2 - \nu_j^2)^2 \\ &< J \max_j \left((\hat{\nu}_j^2 - \nu_j^2)^2 \right). \end{aligned}$$

Based on the element-wise consistency, we have that

$$J \max_j \left((\hat{\nu}_j^2 - \nu_j^2)^2 \right) \xmapsto{\mathcal{P}} 0,$$

and hence

$$\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_2 \xmapsto{\mathcal{P}} 0 \,.$$

which conclude the proof. ■

Return to Lemma 5.5

Proof of Lemma 5.6

The proof of the asymptotic normality of $\boldsymbol{a}^T\hat{\boldsymbol{\nu}}$, for any $\boldsymbol{a} \in \mathbb{R}^J$ is direct from Theorem 3.12. Let $Z_t = X_t - X_{t-1}$, then since Z_t is stationary with mean zero then so is $(X_t - X_{t-h})$ for all $h \in \mathbb{Z}$. This directly imply that for all $j = 1, \ldots, J$, $Y_t = \boldsymbol{a}^T(W_{j,t}^2)_{j=1,\ldots,J}$ is also stationary (since it based on a time invariant transformation of stationary processes). And since (Y_t) is a borel-measurable function of Z_t, we have (Y_t) is also a strong mixing process with mixing coefficient $\alpha^*(n) \leq \alpha(n)$, hence $\sum_{n=1}^{\infty} \alpha^*(n)^{\delta/2+\delta} < \infty$ for some $\delta > 0$. Moreover, since $(W_{j,t})$ is a linear function of Z_t, and $\mathbb{E}\left[Z_t^{4+2\delta}\right] < \infty$, by triangle inequality, we have $\mathbb{E}\left[(W_{j,t})^{4+2\delta}\right] < \infty$ for all $j = 1, \ldots, J$, which implies $\mathbb{E}\left[Y_t^{2+\delta}\right] < \infty$. Then Thus, we can apply Theorem 3.12 on the process Y_t, i.e.

$$\sqrt{T}\boldsymbol{a}^T\left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\right) \xmapsto{\mathcal{D}} \mathcal{N}(0, \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}),$$

then by the Cramér-Wold Device we have

$$\sqrt{T}\boldsymbol{s}^T\boldsymbol{\Sigma}^{-1/2}\left(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\right) \xmapsto{\mathcal{D}} \mathcal{N}(0, 1),$$

where $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\nu}}$, which concludes the proof. ■

▸ Return to Lemma 5.6

License