In-orbit test of the equivalence principle with MICROSCOPE : the missing data challenge

Quentin Baghi

PhD supervisor : Gilles Métris (OCA)
Lab supervisor : Bruno Christophe (ONERA)

Missing data in physics, May 11th 2015, Nice

cnes ONERA THE FRENCH AEROSPACE LAB Observatoire de la CÔTE d'AZUR

retour sur innovation

The problem

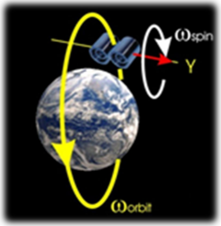
The simplified measurement equation reads :

$$\overrightarrow{s(t)} = \frac{1}{2}\delta\overrightarrow{g(O_{12})} + \frac{1}{2}\left([\mathbf{T}] - [\mathbf{In}]\right)\overrightarrow{O_1O_2} - [\mathbf{\Omega}]\dot{\overrightarrow{O_1O_2}} - \frac{1}{2}\ddot{\overrightarrow{O_1O_2}} + \overrightarrow{n(t)}$$

We want to detect and estimate the EP violation signal. In order to reject the bias of the perturbation terms, a linear regression analysis must be performed to estimate δ and all the instrumental parameters.

We are annoyed by :

- deterministic perturbations : Earth gravity gradient, inertial forces, instrument defects...
- random perturbations : noise, unpredicted accelerations peaks $\Rightarrow$ corrupted or unavailable information



ONERA
THE FRENCH AEROSPACE LAB

Formulation

The problem of the regression analysis (calibration or EP session) can be formalized in a simple manner :

$$y = M\left(A\beta + n\right)$$

- y observed time series vector $(N \times 1)$
- M mask matrix (diagonal) : $M_{ii} = 1$ if y_i is observed, $M_{ii} = 0$ otherwise.
- A model matrix $(N \times K)$
- β vector of parameters to be estimated $(K \times 1)$
- n noise vector of unknown power spectral density $S_n(f)$ $(N \times 1)$

The least squares solution is :

$$\hat{\beta} = \left(A^* M^* M A\right)^{-1} A^* M^* y$$

Uncertainty and missing data

It can be shown that in the case of a harmonic signal at f_{EP} ($A_i = \sin(2\pi f_{EP} i \tau_s)$) the least squares standard error is proportional to the expectation of the masked noise periodogram in the Fourier space :

$$\mathrm{Var}[\beta] \approx \frac{2}{N(1-\alpha)^2} \mathrm{E}\left[P_{Mn,N}(f_{\mathrm{EP}})\right]$$

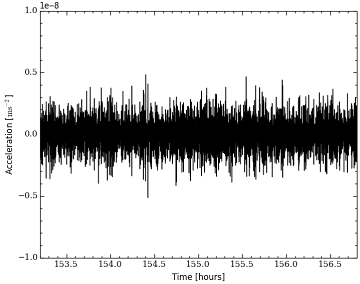
where α is the fraction of missing data.
The periodogram $P_{Mn,N}(f)$ of the masked noise Mn has been defined as :

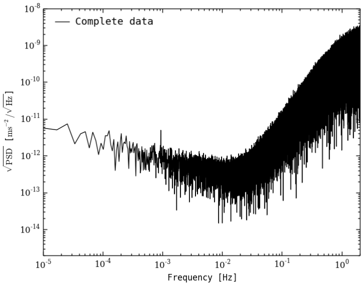
$$P_{Mn,N}(f) = \frac{1}{N} \left| \sum_{i=0}^{N-1} M_{ii} n_i e^{-2j\pi f i \tau_s} \right|^2$$

An example

The missing data induce a convolution effect between the noise and the observation window.

$$\mathrm{E}\left[P_{Mn,N}(f)\right] = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} P_{M,N}(f-f')S_n(f')df'$$



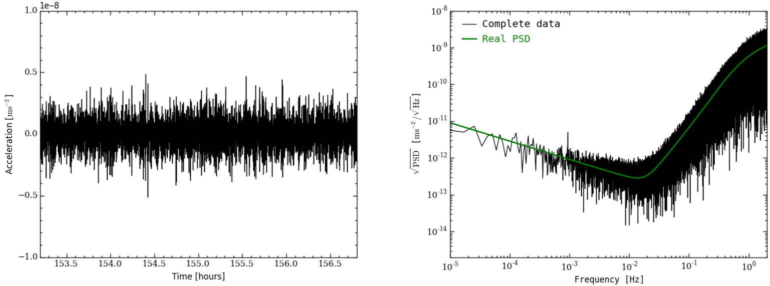


With a complete data set : $\sigma_\delta = 1 \times 10^{-15}$

An example

The missing data induce a convolution effect between the noise and the observation window.

$$\mathrm{E}\left[P_{Mn,N}(f)\right] = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} P_{M,N}(f - f')S_n(f')df'$$

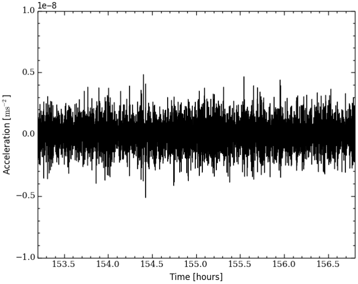


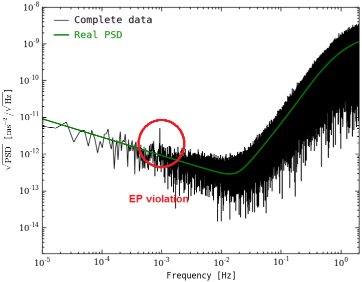
With a complete data set : $\sigma_\delta = 1 \times 10^{-15}$

An example

The missing data induce a convolution effect between the noise and the observation window.

$$\mathrm{E}\left[P_{Mn,N}(f)\right] = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} P_{M,N}(f-f')\color{green}{S_n(f')df'}$$



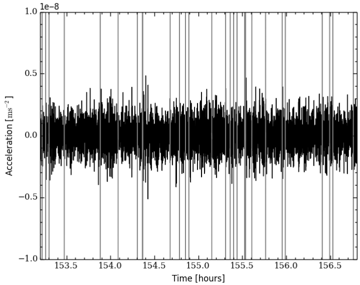


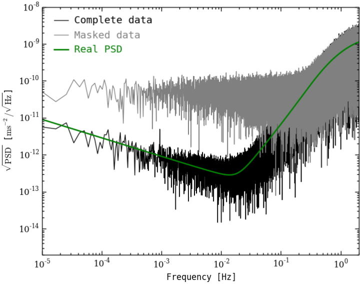
With a complete data set : $\sigma_\delta = 1 \times 10^{-15}$

An example

The missing data induce a convolution effect between the noise and the observation window.

$$\mathrm{E}\left[P_{Mn,N}(f)\right] = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} P_{M,N}(f-f')S_n(f')df'$$



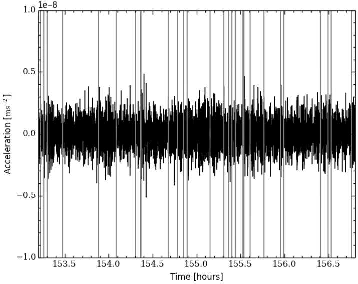


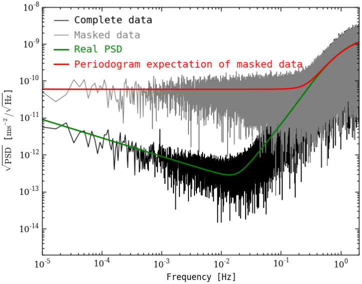
With missing data (2% losses only) : $\sigma_\delta = 65 \times 10^{-15}$

An example

The missing data induce a convolution effect between the noise and the observation window.

$$\mathrm{E}\left[P_{Mn,N}(f)\right] = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} P_{M,N}(f-f')S_n(f')df'$$

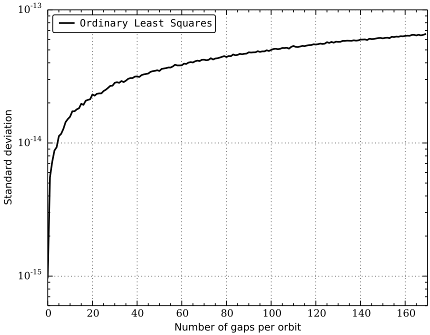




With missing data (2% losses only) : $\sigma_\delta = 65 \times 10^{-15}$

Performance of ordinary least squares

In the presence of colored noise, the variance of the ordinary least squares estimate is highly sensitive to the loss of data.



Problem analysis

So we want to perform a linear regression with :

- unknown colored noise
- frequent and short data gaps
- large data samples ($N > 10^6$)

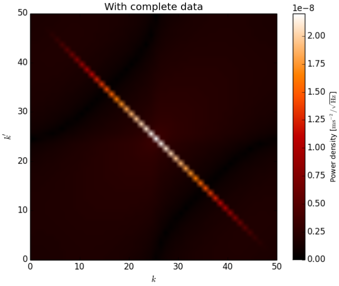
The classical Fourier analysis or ordinary least squares fail in estimating the parameters with a good precision. The problem of such methods is that they are not optimal with respect to the variance.
$\Rightarrow$ Solution : perform a general least squares-like estimate on the observed data y_o (where $M_{i,i} = 1$) :

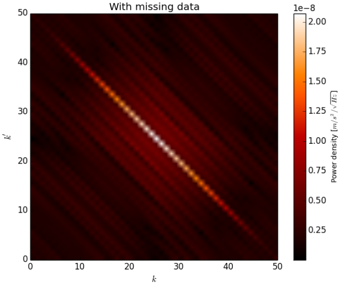
$$\hat{\beta} = ({A_o}^* {\Sigma_o}^{-1} A_o)^{-1} \cdot {A_o}^* {\Sigma_o}^{-1} y_o$$

$\Rightarrow$ The noise covariance matrix ($\Leftrightarrow$ the PSD) must be estimated

Problem analysis

Pb. 1 A spectral method to estimate the PSD is difficult in the presence of irregularly sampled data
Pb. 2 The matrix Σ_o is not diagonal in Fourier space. For large samples, it cannot be stored nor inverted directly





Implemented solution

To solve these problems, we implement a data analysis method with the following steps :

1. Estimation of the noise PSD with an autoregressive (AR) model : temporal model, Pb. 1 solved.
2. Whitening of the data using an orthogonalization process with a Kalman filter, no matrix storage, Pb. 2 solved.
3. Estimation of the parameters with an approximate generalized least squares estimator constructed with the orthogonal vector

Step 1 : estimation of AR parameters

$$n(t) + a_1 n(t-1) + ... + a_p n(t-p) = \epsilon(t)$$

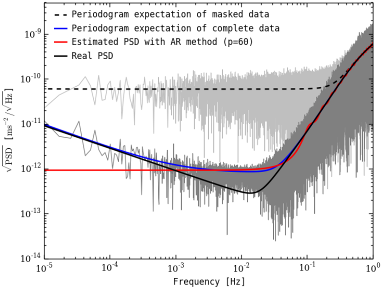
With :

$$\epsilon(t) \sim \mathcal{N}\left(0, \sigma^2\right)$$

Estimation of $a_1, ..., a_p, \sigma^2$ with Burg's algorithm adapted to missing data.

The idea is to fit any arbitrary power spectral density by a rational function in $\exp(-2i\pi f/f_s)$ of the form :

$$S(f) = \frac{\sigma^2/f_s}{\left|1 + a_1 e^{-2i\pi f/f_s} + ... + a_p e^{-2i\pi p f/f_s}\right|^2}$$



Whitening

Step 2 : data orthogonalization

We want to calculate the whitened vectors $e_o = L^{-1} y_o$ et $E_o = L^{-1} A_o$ without store nor invert L, where L is the Cholesky decomposition of Σ_o :

$$\Sigma_o = LL^*$$

To do this we use a Kalman filter.

Step 3 : estimation of regression parameters
From the previous calculations one can construct an estimator with a quasi minimal variance :

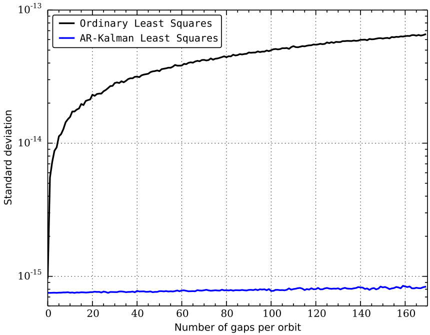
$$\begin{aligned}\hat{\beta} &= \left(E_o^* E_o\right)^{-1} E_o^* e_o \\ &\approx \left(A_o{}^* \Sigma_o{}^{-1} A_o\right)^{-1} \cdot A_o{}^* \Sigma_o{}^{-1} y_o\end{aligned}$$

$\Rightarrow E_o, e_o$ calculated with Kalman outputs : minimize the variance without computing Σ_o

Results

Standard deviation of the estimation the EP violation parameter δ :

Mask	Ordinary least squares	KARMA
Complete data	1.0×10^{-15}	9.6×10^{-16}
Tank crackles	6.5×10^{-14}	1.1×10^{-15}



Data reconstruction

Once one has estimated the noise PSD and the regression parameters β, it is possible to impute missing data by calculating their conditional expectations.

y_o : observed data vector

y_m : missing data vector

$$S_n(f)$$

Data reconstruction

Once one has estimated the noise PSD and the regression parameters β, it is possible to impute missing data by calculating their conditional expectations.

y_o : observed data vector

y_m : missing data vector

$$S_n(f) \Rightarrow R(t)$$

Data reconstruction

Once one has estimated the noise PSD and the regression parameters β, it is possible to impute missing data by calculating their conditional expectations.

y_o : observed data vector

y_m : missing data vector

$$S_n(f) \Rightarrow R(t) \Rightarrow \Sigma_{i,j} = R(i-j)$$

Data reconstruction

Once one has estimated the noise PSD and the regression parameters β, it is possible to impute missing data by calculating their conditional expectations.

y_o : observed data vector

y_m : missing data vector

Conditional expectation of the missing data given the observed data :

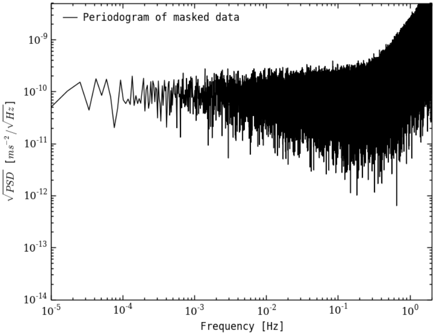
$$\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)$$

ONERA THE FRENCH AEROSPACE LAB

Data reconstruction

Conditional expectation of the missing data given the observed data :

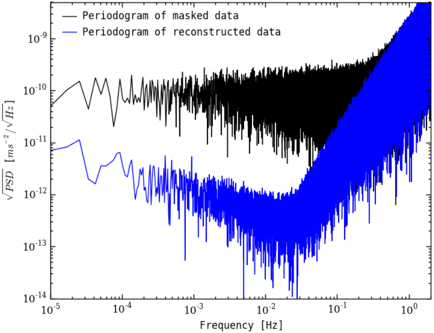
$$\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)$$



Data reconstruction

Conditional expectation of the missing data given the observed data :

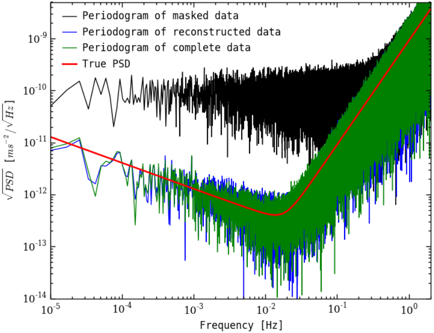
$$\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)$$



Data reconstruction

Conditional expectation of the missing data given the observed data :

$$\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)$$



ONERA THE FRENCH AEROSPACE LAB

Conclusion

- For linear regression analysis with highly correlated noise and random missing data, ignoring the missing data or basic interpolation can lead to significant increase of the uncertainty in an ordinary least square fitting approach.
- To construct an estimator with a variance close to the minimal bound, the noise covariance must be estimated.
- We implemented a method based on a high order AR fit of the noise, which shows good results (reduction of the standard error by a factor 60).
- The method provides outputs for data reconstruction : this can be useful for “visual convenience” or to improve parameter estimation (e.g. EM algorithm).

Questions

- R. H. Jones (1980), Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics* 22, 389
- V. Gómez and A. Maravall (1994), Estimation, prediction, and interpolation for nonstationary series with the Kalman filter, *Journal of the American Statistical Association* 89, 426
- Q. Baghi, G. Métris, J. Bergé, B. Christophe, P. Touboul, and M. Rodrigues (2015), Regression analysis with missing data and unknown colored noise : Application to the MICROSCOPE space mission, *Phys. Rev. D* 91, 062003