The performance of the data processing

Quentin Baghi[1], Joël Bergé[1], Gilles Métris[2], Sandrine Pires[3], Pierre Touboul[1], Manuel Rogrigues[1], Bruno Christophe[1]

[1]ONERA - The French Aerospace Lab
[2]OCA - Observatoire de la côte d'Azur
[3]Laboratoire AIM, CEA/DSM-CNRS Université Paris Diderot IRFU/SAp

MICROSCOPE Colloquium IV, 17 November 2015, Palaiseau, France

retour sur innovation

The measurement equation

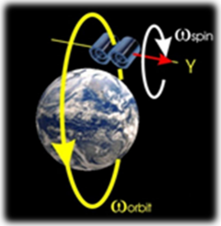
We measure the difference s between the accelerations of the two test-masses w.r.t. time:

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

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



The measurement equation

We then obtain time series:

$$s(t) = \delta g(t) + \sum_{i=1}^{q} \alpha_i p_i(t) + n(t)$$

Equivalently in matrix formulation:

$$s = A\beta + n$$

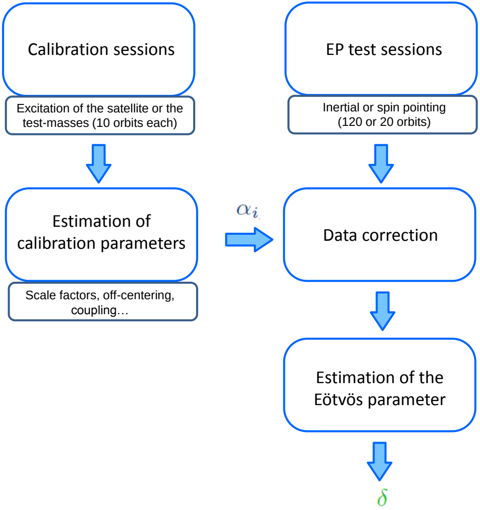
$$A = \begin{pmatrix} g(t) & p_1(t) & \dots & p_q(t) \end{pmatrix}$$
$$\beta = \begin{pmatrix} \delta & \alpha_1 \dots \alpha_q \end{pmatrix}^T$$

- δ : EP violation signal parameter, to be estimated
- α_i : perturbations parameters, to be estimated
- $n(t)$: stationary noise of unknown power spectral density $S(f)$

Ordinary least squares solution:

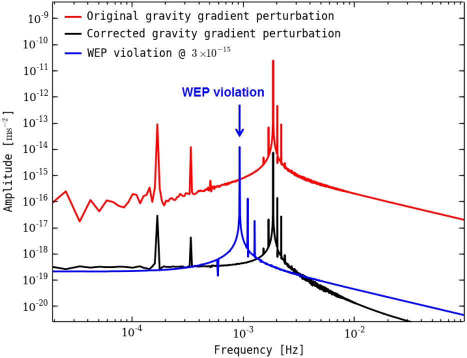
$$\hat{\beta} = (A^* A)^{-1} \cdot A^* s$$

The data process



Treatment of modeled perturbations

We have physical models or measurement of the perturbations involved in the equation of the differential acceleration. The strategy is to fit their amplitudes via least square linear regression and remove them from the signal.
Example: removing the gradient terms due to off-centerings



Treatment of unmodeled perturbations

Outside the measurement frequency band (centered on f_{EP}), the signal may contain perturbations that have not been modeled. Proposed strategy:

- Most of the perturbations signals are expected to be harmonic, with frequency f_p and phase ϕ_p: $A_p \sin(2\pi f_p t + \phi_p)$
- These perturbations induce a bias at f_{EP}, depending on f_p and ϕ_p

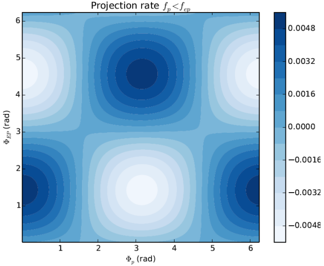


Fig.1 Normalized projection rate as a function of phases

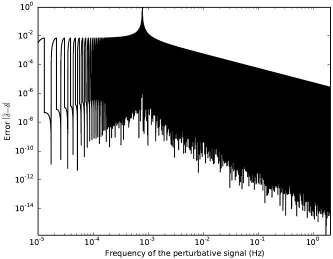


Fig.2 Normalized projection rate as a function of frequency

Worst case: maximize the projection rate w.r.t the phase

Treatment of unmodeled perturbations

Maximum bias w.r.t. the phase (Hardy et al, 2013):

$$\tau_{\mathrm{max}} = \frac{4}{T}\frac{1}{\left|\omega_{\mathrm{EP}}^2 - \omega_{\mathrm{p}}^2\right|}\left|\sin\left(\omega_{\mathrm{p}}^2 T/2\right)\right|^2 \sqrt{\omega_{\mathrm{EP}}^2 \cos^2\phi_{\mathrm{EP}} - \omega_{\mathrm{p}}^2 \sin^2\phi_{\mathrm{EP}}} \qquad (1)$$

Most of the perturbations are expected to be at frequencies multiple of the orbital frequency f_{orb} and the spin frequency f_{spin}. For these frequencies, the bias is nullified if one chooses a spin frequency f_{spin} and an integration time T such that:

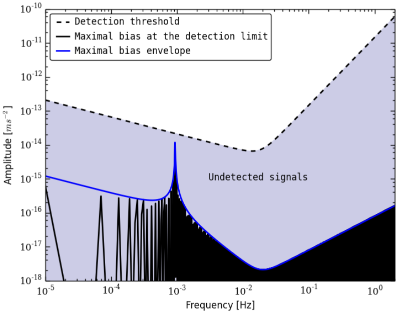
$$T = \frac{k_1}{f_{\mathrm{orb}}},\ k_1 \in \mathbb{N} \qquad (2)$$

$$f_{\mathrm{spin}} = \frac{k_2}{T},\ k_2 \in \mathbb{N} \qquad (3)$$

If $f_{\mathrm{p}} = n_1 f_{\mathrm{orb}} + n_2 f_{\mathrm{spin}}$, then $\tau_{\mathrm{max}} = 0$

Treatment of unmodeled perturbations

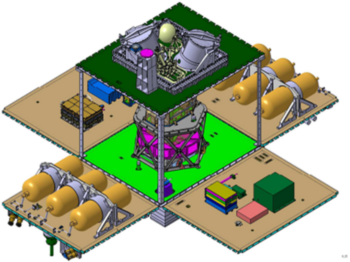
And what if the frequency of the perturbation signal is arbitrary? What is left is statistical detection of unknown periodicities. Some simulations using maximum likelihood algorithm by MCMC show quite good results. Can be useful also for search for other physics.

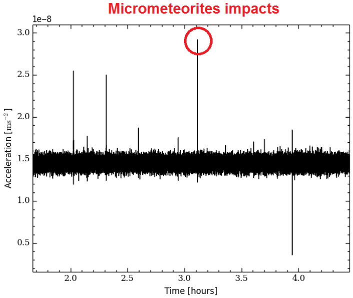


Amplitude detection threshold and corresponding bias onto the EP frequency as a function of the frequency of the perturbation signal

The missing data problem

Various physical phenomena occur in flight (crackles of the cold gas tank, telemetry losses, micrometeorites) which perturb or saturate the very sensitive accelerometer $\Rightarrow$ some intervals in the time series are not usable

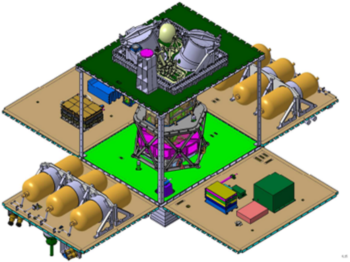


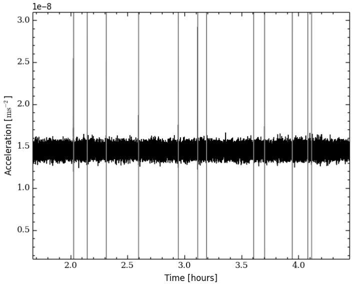


$\Rightarrow$ We have to **mask some portions of data** with a mask matrix M (diagonal matrix s.t. $M_{ii} = 1$ if the data is available, $M_{ii} = 0$ if the data is missing or corrupted)

The missing data problem

Various physical phenomena occur in flight (crackles of the cold gas tank, telemetry losses, micrometeorites) which perturb or saturate the very sensitive accelerometer ⇒ some intervals in the time series are not usable





⇒ We have to **mask some portions of data** with a mask matrix M (diagonal matrix s.t. $M_{ii} = 1$ if the data is available, $M_{ii} = 0$ if the data is missing or corrupted)

The missing data problem

In the presence of missing data, the measurement equation is rewritten by taking into account the mask vector M:

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

$$A = \begin{pmatrix} g(t) & p_1(t) & \dots & p_q(t) \end{pmatrix}$$
$$\beta = \begin{pmatrix} \delta & \alpha_1 \dots \alpha_q \end{pmatrix}^T$$

- δ : EP violation signal parameter, to be estimated
- α_i : perturbations parameters, to be estimated
- $n(t)$: stationary noise of unknown power spectral density $S(f)$

Ordinary least squares solution:

$$\hat{\beta} = (A^* M^* M A)^{-1} \cdot A^* M^* M s$$

Thus we just fit our model where data is available.

The missing data problem

In the presence of missing data, the measurement equation is rewritten by taking into account the mask vector M:

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

$$A = \begin{pmatrix} g(t) & p_1(t) & \dots & p_q(t) \end{pmatrix}$$
$$\beta = \begin{pmatrix} \delta & \alpha_1 \dots \alpha_q & \end{pmatrix}^T$$

- δ : EP violation signal parameter, to be estimated
- α_i : perturbations parameters, to be estimated
- $n(t)$: stationary noise of unknown power spectral density $S(f)$

Ordinary least squares solution:

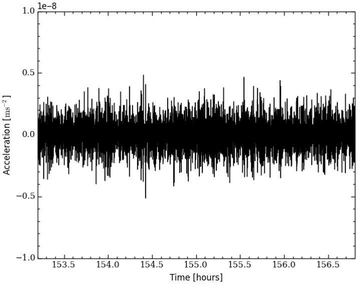
$$\hat{\beta} = \left(A^* M^* M A\right)^{-1} \cdot A^* M^* M s$$

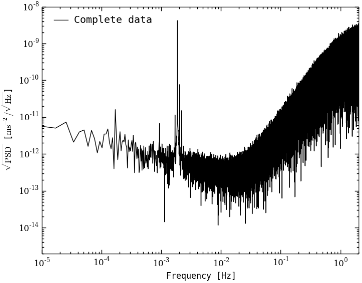
What is the consequence of this?

An example

An example: simulated acceleration of a 20 orbits (1 orbit = 1.6h) spin session sampled at 4 Hz, and its Fourier representation

$$\text{Periodogram: } P_{s,N}(f) \equiv \frac{1}{N}\left|\sum_{i=0}^{N-1} s_i e^{-2j\pi f i \tau_s}\right|^2$$



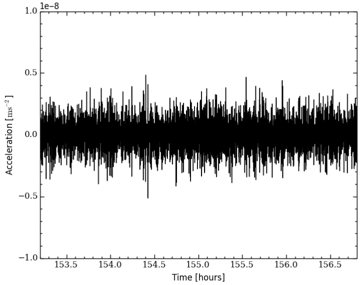


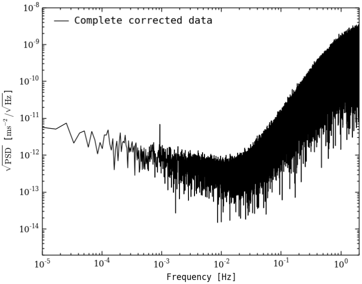
With a complete data set: $\sigma_\delta = 1 \times 10^{-15}$

An example

An example: simulated acceleration of a 20 orbits (1 orbit = 1.6h) spin session sampled at 4 Hz, and its Fourier representation

$$\text{Periodogram: } P_{s,N}(f) \equiv \frac{1}{N}\left|\sum_{i=0}^{N-1} s_i e^{-2j\pi f i \tau_s}\right|^2$$



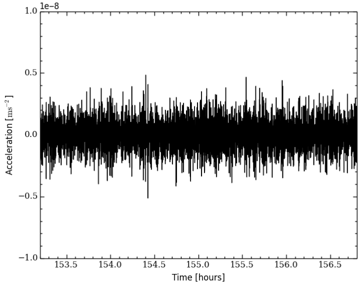


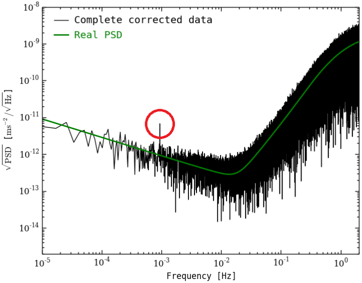
With a complete data set: $\sigma_\delta = 1 \times 10^{-15}$

An example

An example: simulated acceleration of a 20 orbits (1 orbit = 1.6h) spin session sampled at 4 Hz, and its Fourier representation

$$\text{Periodogram: } P_{s,N}(f) \equiv \frac{1}{N}\left|\sum_{i=0}^{N-1} s_i e^{-2j\pi f i \tau_s}\right|^2$$



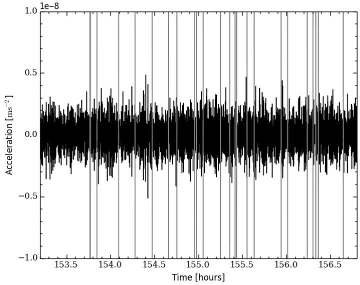


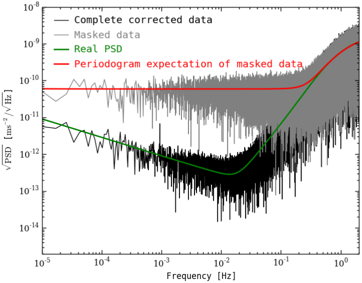
With a complete data set: $\sigma_\delta = 1 \times 10^{-15}$

An example

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

$$\underbrace{\mathrm{E}\left[P_{Mn,N}(f)\right]}_{\text{Masked noise periodogram}} = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} \underbrace{P_{M,N}(f-f')}_{\text{Mask periodogram}} \underbrace{S(f')}_{\text{Noise PSD}} \, df'$$





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

Performance of ordinary least squares

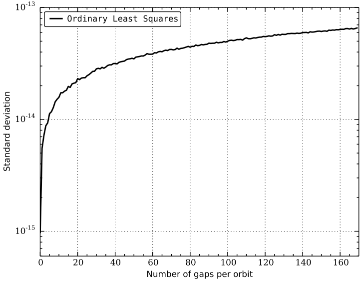
In the presence of colored noise, the variance of the ordinary least squares (OLS) estimate is highly sensitive to the loss of data and directly proportional to the leakage effect:

$$\mathrm{Cov}\left(\hat{\beta}\right) = Q^{-1}A^*M^*{\color{red}M\Gamma M^*}MAQ^{-1}$$

with $Q = A^*M^*MA$. Γ is diagonal in the Fourier basis (if we apply the Fourier transformation matrix F), but ${\color{red}\Gamma_M = M\Gamma M^*}$ is not. We additionaly have:

$$\mathrm{diag}\left(F\Gamma_M F^*\right)_k \propto \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} P_{M,N}(f_k - f')S(f')df'$$

$\rightarrow$ Leakage effect !



Proposed methods

To cope with this problem, 2 solutions with different philosophies have been developed:

1. Inpainting: fills the data gaps to restore the noise statistics, using a discrete cosine transform (DCT) dictionary.
    - completely independent of any physical model
    - uses a technique originally suitable to signals that are sparse in the dictionary
    - applied prior to any regression analysis
2. KARMA: does not fill the data gaps. Estimates instead the noise PSD from the available data and use it to weight the data to obtain a nearly optimal weighting (noise whitening).
    - depends on the physical model
    - uses a change of basis where the observed data is not correlated
    - designed to directly perform the linear regression

1. Inpainting

Inpainting objective: recover X from observed Y:

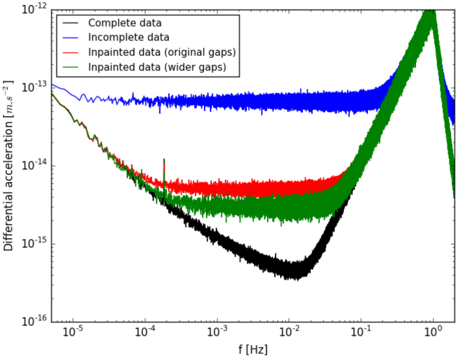
$$Y = MX$$

Strategy: assume that in a given dictionary Φ, X has a sparse representation $\alpha = \Phi^* X$ (*i.e.* most of the coefficients of α are zero), and find $\hat{X} = \Phi\hat{\alpha}$ such as:

$$\hat{\alpha} = \min_{\alpha} \left\{ \|\alpha\|_1 \text{ subject to } \|Y - MX\|_2^2 \leq \sigma^2 \right\}$$

1. Inpainting

Example: simulation of the MICROSCOPE differential acceleration performed on a 120 orbits inertial session (averaged over 40 simulations). There are about 300 gaps per orbit of various durations, corresponding to a 3% data loss.



J. Bergé, S. Pires, Q. Baghi, P. Touboul, G. Métris, Dealing with missing data: An inpainting application to the MICROSCOPE space mission, submitted to *Phys. Rev. D*

2. KARMA

KARMA objective: optimally recover β from observed Y:

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

Stragegy in 3 steps:

- Estimation of the noise covariance by approximating it wich an autoregressive (AR) model:

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

  temporal model $\Rightarrow$ avoids distortion in the Fourier domain

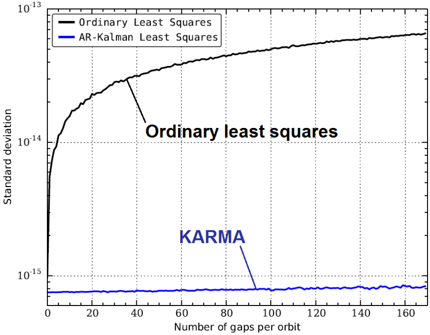
- Whitening of the data using with a Kalman filter $\Rightarrow$ no need to store and invert the large covariance matrix Σ_o of the observed data $Y_o = Y$ where $M_{ii} = 1$. It is equivalent to compute $L^{-1}A_o$ and $L^{-1}Y_o$, where L is the Cholesky decomposition of Σ_o: $\Sigma_o = LL^*$

- Estimation of the parameters with an approximate generalized least squares estimator constructed with the orthogonal vector; equivalent to computing:

  $$\hat{\beta} = \left(A_o{}^* \hat{\Sigma}_o^{-1} A_o\right)^{-1} \cdot A_o{}^* \hat{\Sigma}_o^{-1} Y_o$$

Results

Standard deviation of the estimation the EP violation parameter δ :



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

Comparisons

Comparison of sample standard deviation [in 10^{-15} units] calculated from 300 inertial simulation draws:

Data window	OLS	OLS after inpainting	KARMA
Complete	0.76	-	-
300 gaps/orbit	13.5	1.10	0.79

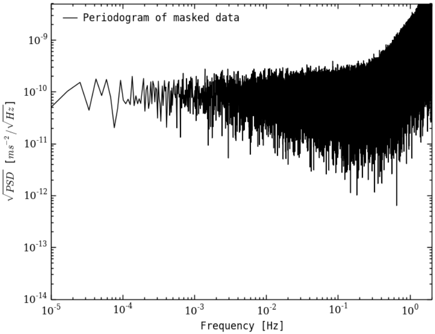
$\rightarrow$ Both Inpainting and KARMA allow us to reduce the statistical uncertainty by more than 1 order of magnitude with respect to ordinary least squares

Further improvements of the inpainting algorithm are underway to even more reduce the reconstruction noise, using a data-driven constraint on the variance in different spectral bands (with wavelets transforms).

Aside: data reconstruction from KARMA

Even if this is not necessary in a linear regression purpose, the data can be reconstructed from KARMA outputs using conditional expectation of the missing data given the observed data :

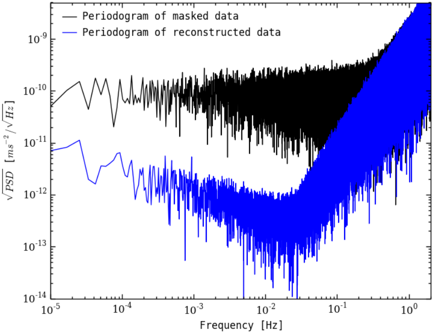
$$\boxed{\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)}$$



Aside: data reconstruction from KARMA

Even if this is not necessary in a linear regression purpose, the data can be reconstructed from KARMA outputs using conditional expectation of the missing data given the observed data :

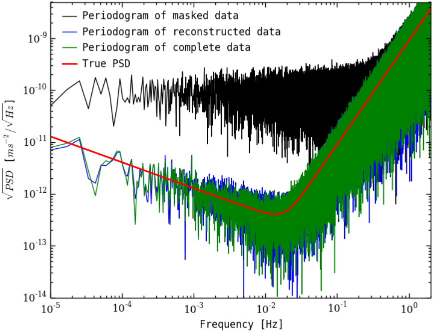
$$\boxed{\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)}$$



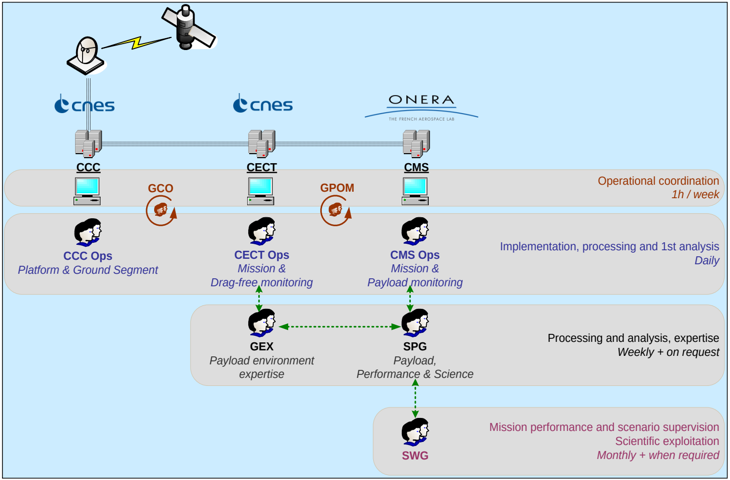
Aside: data reconstruction from KARMA

Even if this is not necessary in a linear regression purpose, the data can be reconstructed from KARMA outputs using conditional expectation of the missing data given the observed data :

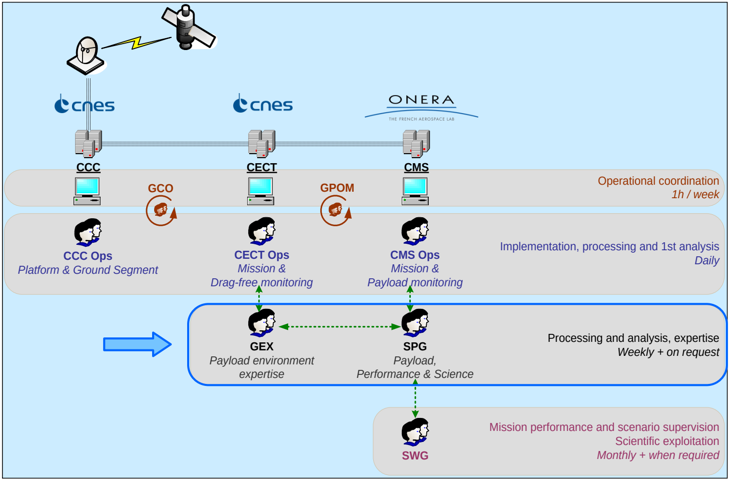
$$\boxed{\mu_{m|o} = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}\left(y_o - \mu_o\right)}$$



Implementation in the MICROSCOPE Scientific Mission Center (SMC)



Implementation in the MICROSCOPE Scientific Mission Center (SMC)



Conclusion

- Data processing tools have been implemented to perform various analyzes such as estimation of calibration parameters of the instruments during dedicated sessions in flight, data correction, noise PSD estimation, detection of unknown harmonic signals
- The problem of missing data has been addressed: linear regression in the presence of missing data and data reconstruction algorithms have been developed $\rightarrow$ useful study for cooperation with other scientific space missions such that LISA Pathfinder
- Two methods with different theoretical basis have been implemented and give similar result: Inpainting and KARMA
- Hopefully the precision is still 10^{-15} in spite of the missing data!
- We are able to produce consistent sets of complete data to the scientific community
- Among other data analysis tools, the presented method is currently implemented in the Scientific Mission Center in ONERA facilities near Paris


THANK YOU !