Study of the use of Principal component analysis in order to estimate stellar parameters

V. Watson

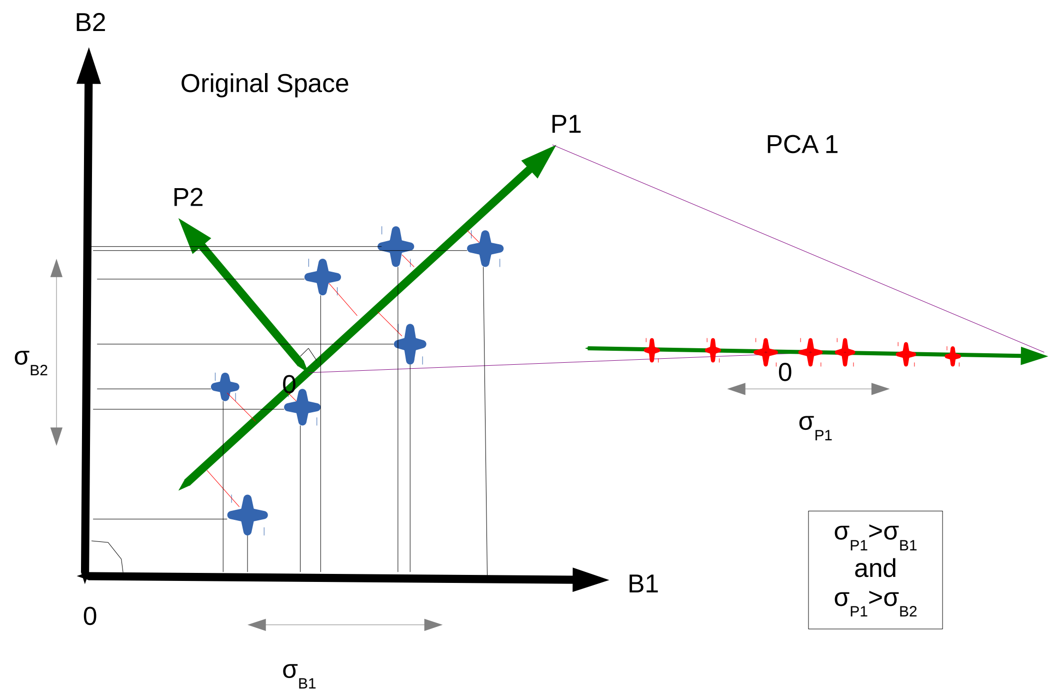
Rencontres d'Astrostatistiques 2014

J.F. Trouilhet - F. Paletou

November 13 2014 - Grenoble

Table of contents

- Dimensionality reduction
- Preserve the best inertia
- Projection of the individuals from the original data space to a data space defined by the first eigenvectors of the covariance matrix



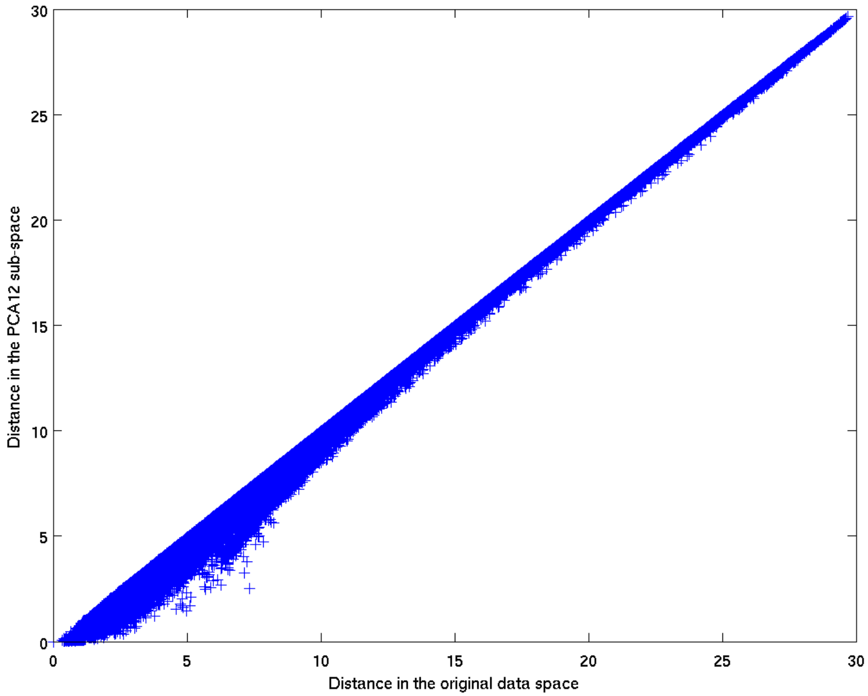
PCA principle

Principal Component Analysis

Information representation

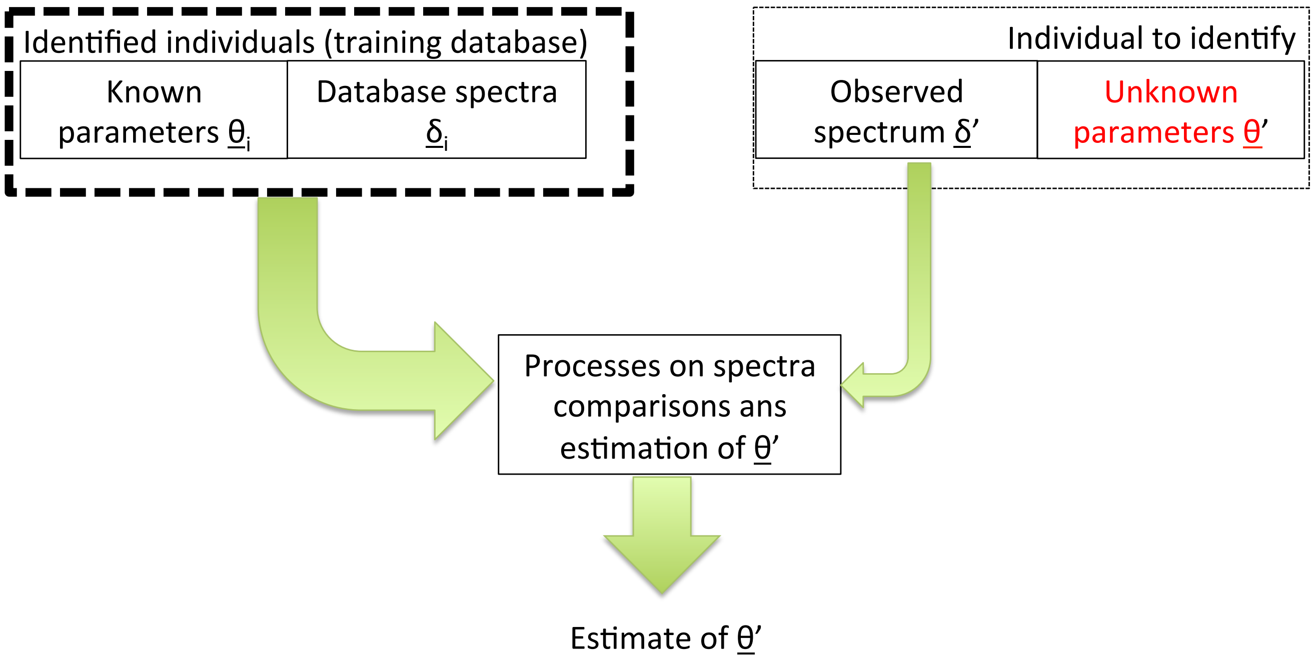
- Assumption 1 : Most similar spectra = ”Closest” stellar parameters
  - Assumption 1.1 : Distance between spectra shall be highly correlated to the distance between the associated parameters
  - Assumption 1.2 : The coordinates that spread the data the most are informative with respect to the considered parameters
- Assumption 2 : The PCA truncation preserves almost all the relevant information
  - Assumption 2.1 :There is no relevant **relative** spectral information (spectral independence)
  - Assumption 2.2 : The relevant information is on the first principal components that keeps the most of the variance in the data

PCA truncation gives a good preservation of the distance ordination in the data



Correlation of the inter-individuals euclidean distances between PCA sub-space and the original data space

Estimation of stellar parameters



Basic principle of our problem

Processes can involves

- Reduction of dimension
- Extraction of informative combination of the data
- Data conditioning to ease estimation process

Comparison is based on euclidean distance between the individuals

Estimation of stellar parameters

Advantages of PCA

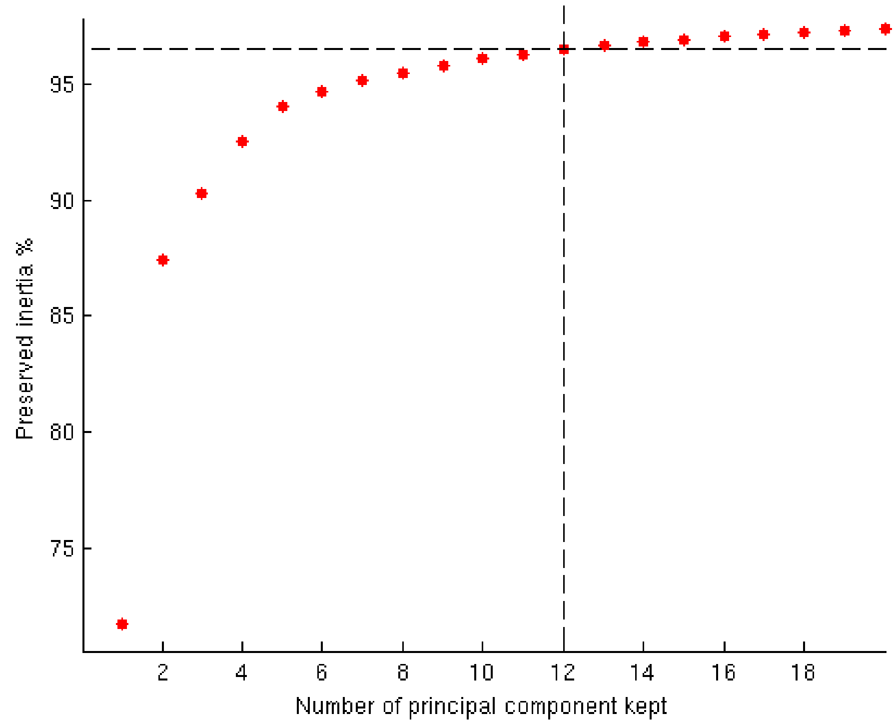
- Reduction of dimensionality
  - Less computational resources (eigenvectors of the covariance matrix only computed once)
  - Takes less space to store or transmit
- Keeping the individuals "ordered"
  - Changes the less the vicinity

Estimation of stellar parameters

Drawbacks of PCA

- PCA approach ignores the knowledge of the values and the physical meaning of the parameters in the training database
  - The goal is not here to classify objects
  - 4500K is closer to 5000K than to 5500K
- PCA does not take into account the spectral dependency in the data
  - All the components (flux values) in the original space are considered as independent one to another
  - Information enclosed in a datum relatively to another elsewhere in the spectra will be lost (*e.g.* $\frac{S(\lambda_i)}{S(\lambda_j)}$)

Effect of PCA on the original data space



Quantity of inertia (variance in the data) preserved in the data as a function of the number of components retained

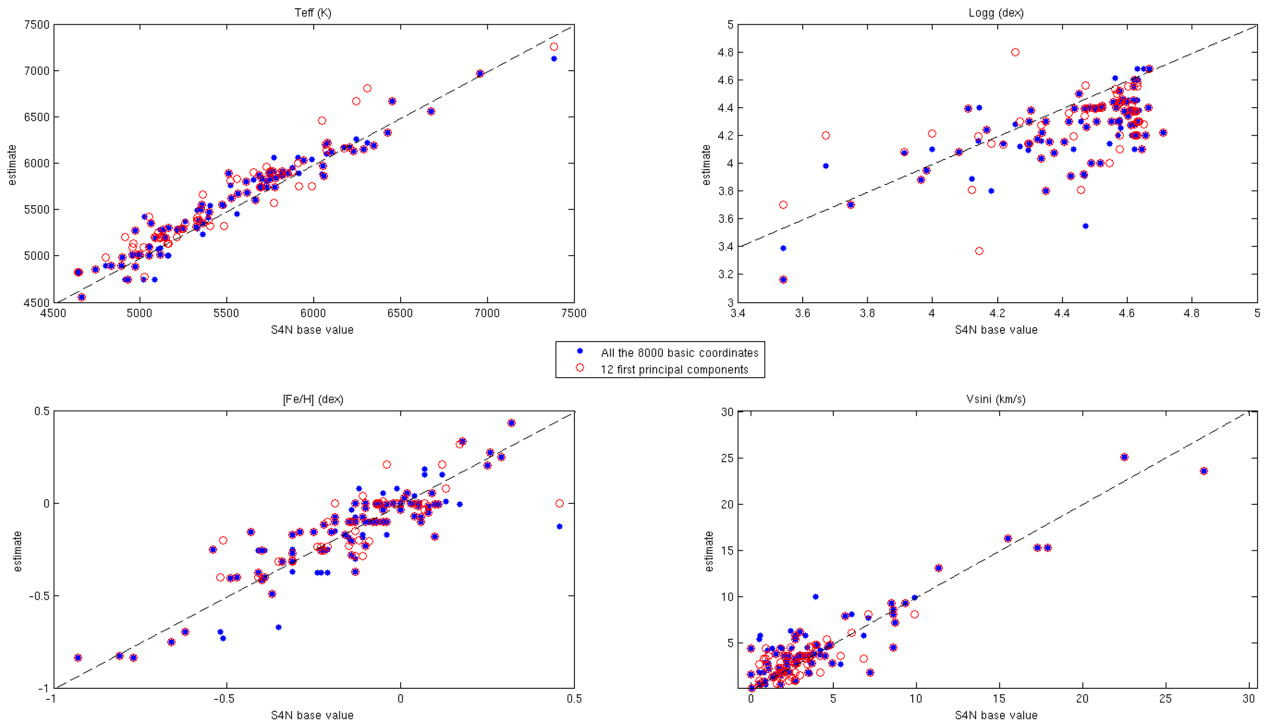
Effect of PCA on the original data space

Results with the 12 first principal components

Université de Toulouse

Teff
bias = 40 / 69
sigma= 131 / 145

Logg
bias = -0.192 / -0.175
sigma= 0.197 / 0.22



bias = -0.0137 / -0.0032
sigma= 0.123 / 0.115
[Fe/H]

bias = 0.53 / 0.30
sigma= 1.84 / 1.77
Vsin(i)

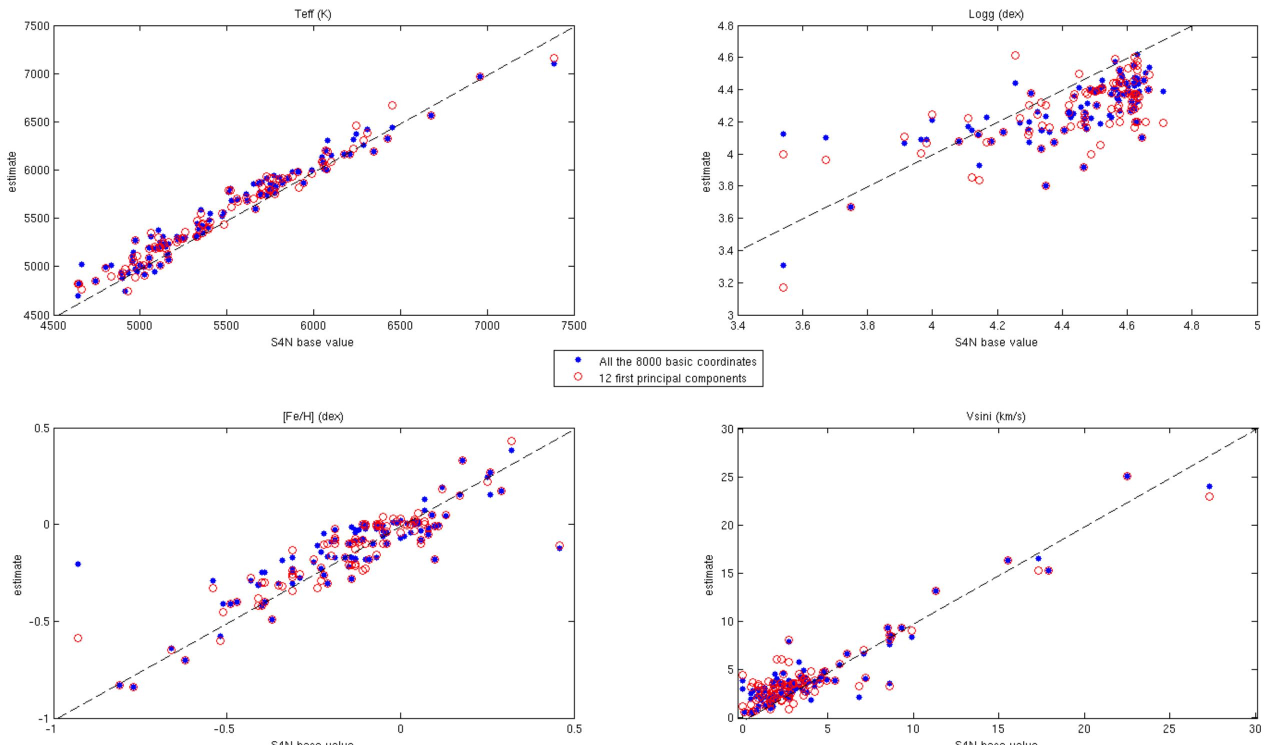
Estimation results with PCA12 considering the nearest neighbour

Effect of PCA on the original data space

Results with the 12 first principal components

Teff
bias = 65 / 55
sigma= 108 / 99

Logg
bias = -0.162 / -0.172
sigma= 0.169 / 0.183



bias = -0.0157 / -0.0024
sigma= 0.128 / 0.106
[Fe/H]

bias = 0.30 / 0.43
sigma= 1.50 / 1.68
Vsin(i)

Estimation results with PCA12 considering an area of vicinity

Effect of PCA on the original data space

Selection criterion

Select the most informative data in order to have less noisy effects on the estimation process.

The criterion we used was the absolute value of the linear correlation coefficient between the variation of each parameter and the variation of every spectra for each wavelength.

$$|C_{\theta_i,\delta_\lambda}| = \left| \frac{Cov(\theta_i, \delta_\lambda)}{\sqrt{\sigma^2_{\theta_i}\sigma^2_{\delta_\lambda}}} \right| \tag{1}$$

Effect of PCA on the original data space

Selection of relevant eigenvectors

Teff	Logg	[Fe/H]	Vsin(i)
1	9	3	1
8	5	4	2
7	14	1	5
3	8	8	6
4	1	5	3
5	6	16	20
21	11	12	4
10	3	6	13
6	16	21	21
12	42	17	7
14	23	9	16
17	21	14	24

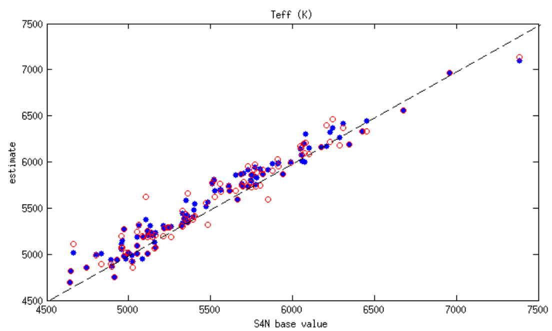
Indexes of eigenvector selected for each parameters

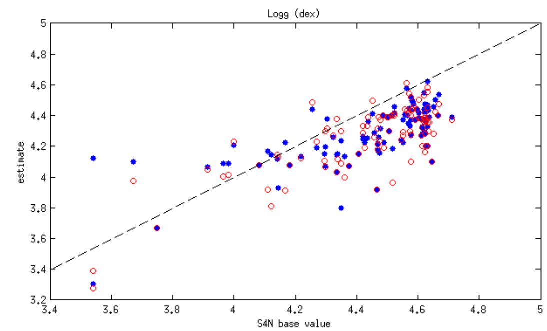
Effect of PCA on the original data space

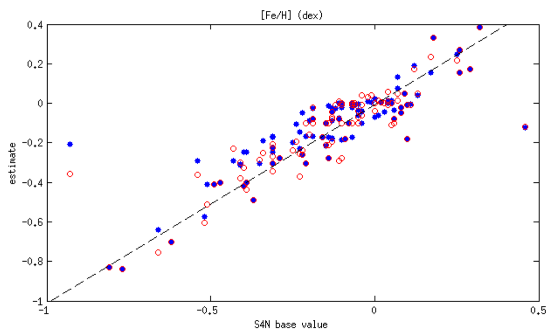
Selection of the most relevant eigenvectors

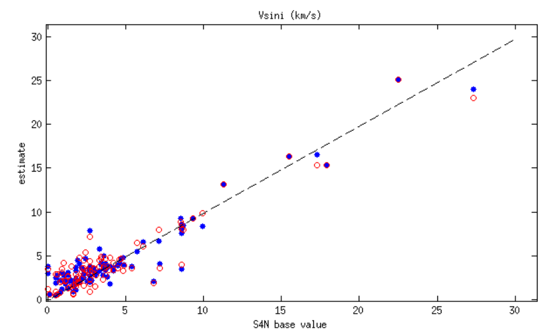
Teff
bias = 65 / 54
sigma= 108 / 127

Logg
bias = -0.162 / -0.183
sigma= 0.169 / 0.157









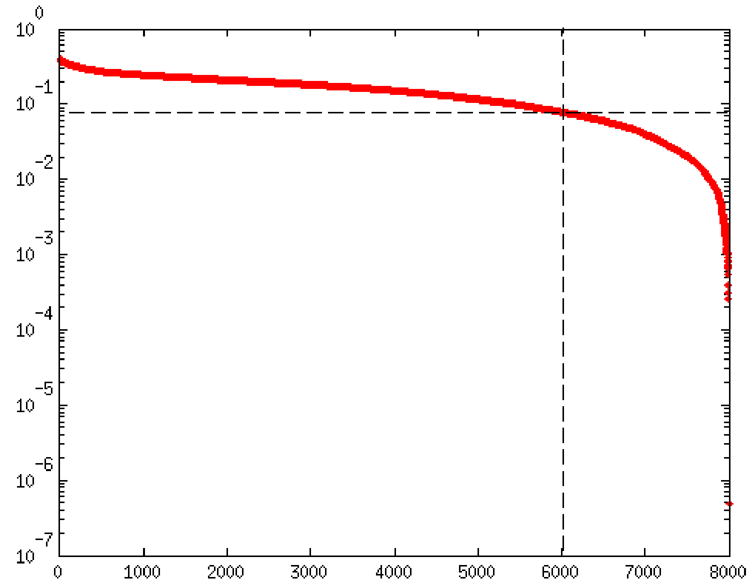
bias = -0.0157 / -0.0027
sigma= 0.128 / 0.118
[Fe/H]

bias = 0.30 / 0.30
sigma= 1.50 / 1.58
Vsin(i)

Estimation results with PCA12 considering an area of vicinity and the eigenvectors of the previous slide

Selection of relevant data

Selection threshold



Thresholds selection

	Teff	Logg	[Fe/H]	Vsin(i)
threshold	0.1	0.2	0.05	0.05
data kept	7267	5938	7155	7302

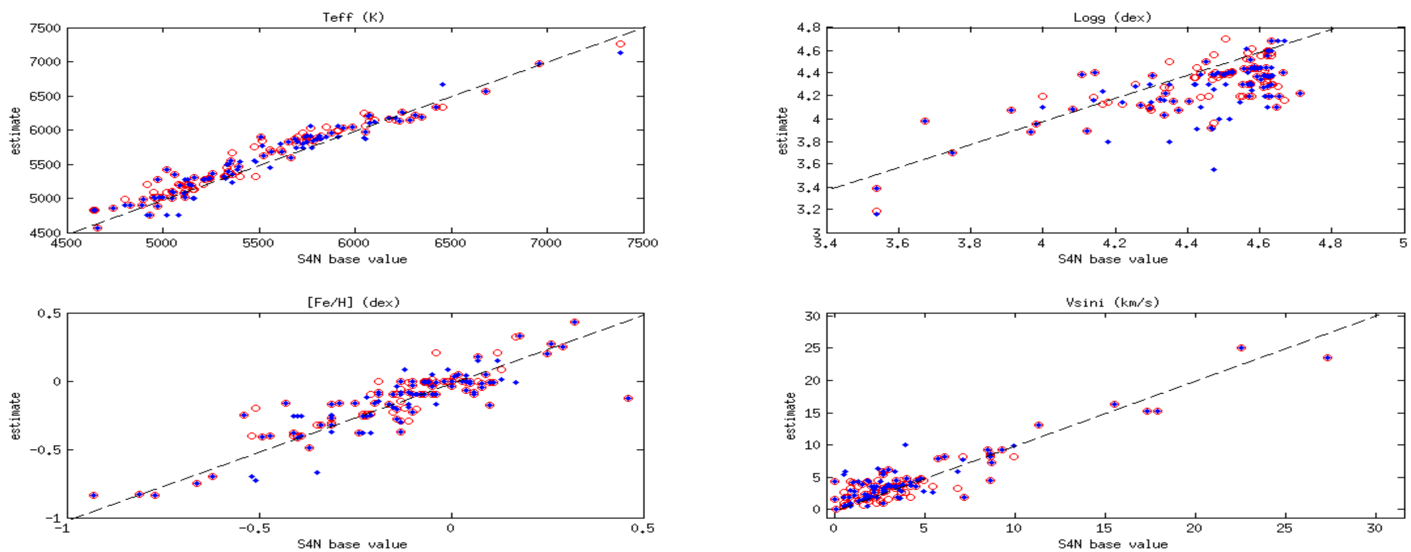
Relative threshold based on the most correlated value (arbitrarily evaluated)

Selection of relevant data

Results

Teff
bias = 40 / 69
sigma= 131 / 119

Logg
bias = -0.192 / -0.15
sigma= 0.197 / 0.177



bias = -0.0137 / 9.5 10^{-5}
sigma= 0.123 / 0.120
[Fe/H]

bias = 0.53 / 0.38
sigma= 1.84 / 1.76
Vsin(i)

Estimation result with PCA 12 after data selection considering 1 nearest neighbour

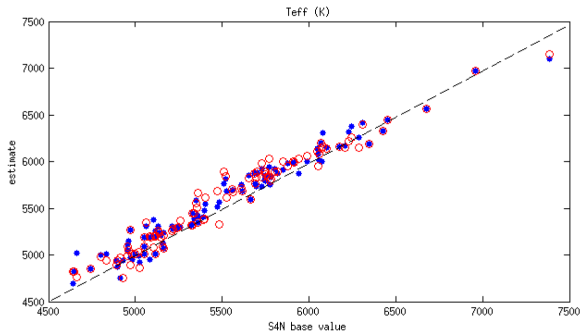
Université de Toulouse

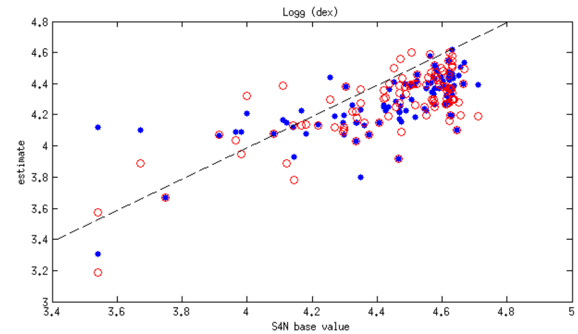
Selection of relevant data

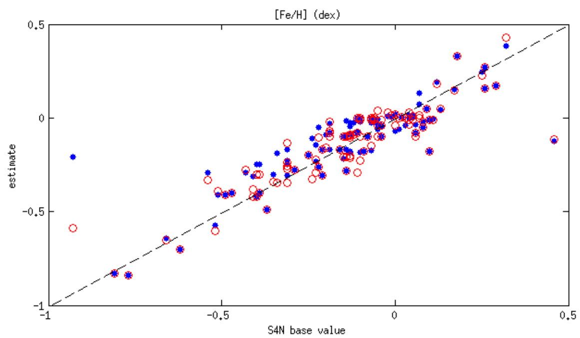
Results

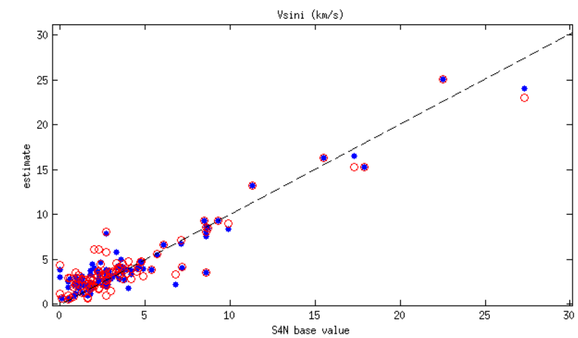
Teff
bias = 65 / 65
sigma= 108 / 112

Logg
bias = -0.162 / -0.16
sigma= 0.169 / 0.163









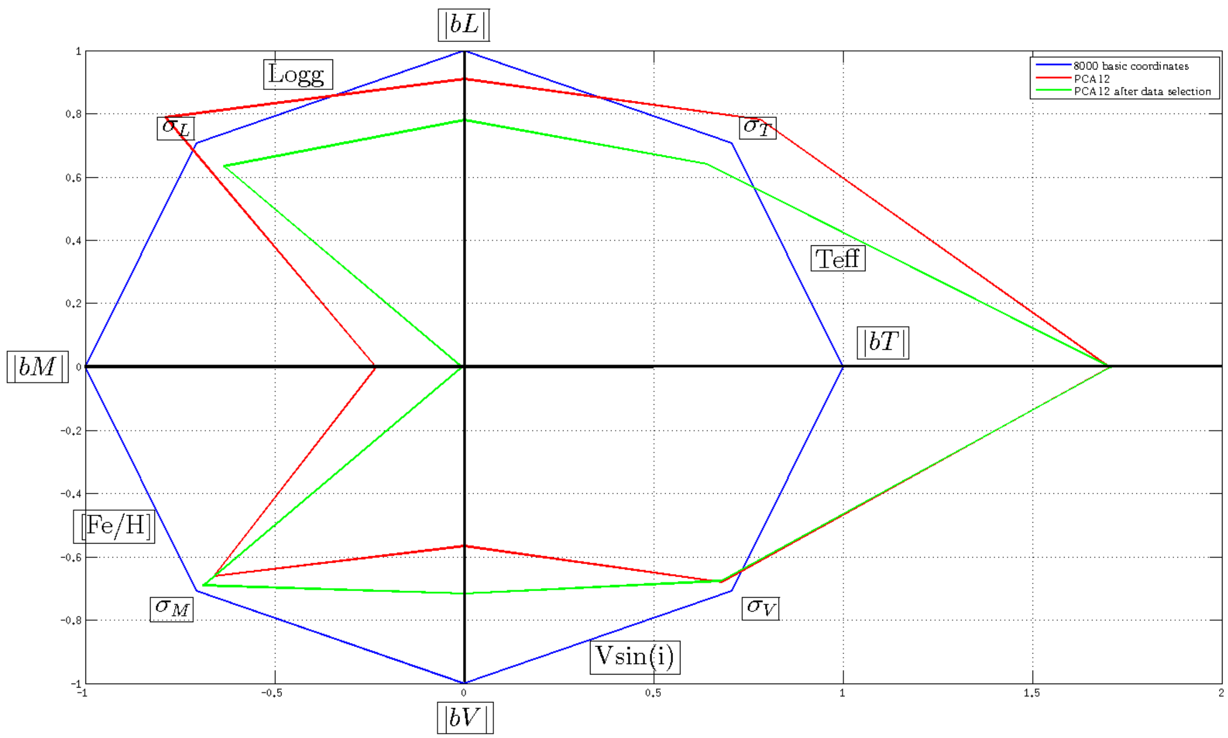
bias = -0.0157 / -0.0058
sigma= 0.128 / 0.108
[Fe/H]

bias = 0.30 / 0.38
sigma= 1.50 / 1.63
Vsin(i)

Estimation result with PCA 12 after data selection considering an area of vicinity

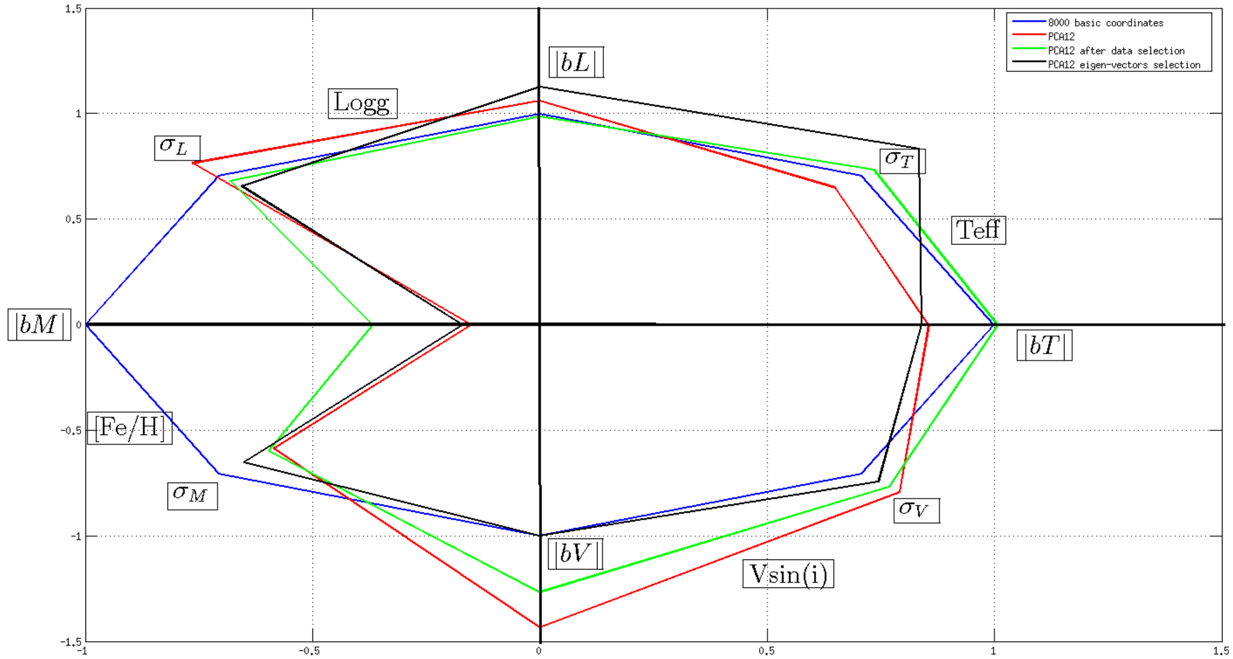
Conclusion

Comparative study - 1 neighbour



Comparative study of errors for 1 neighbour-based estimation

Data selection → improvement of surface gravity and metallicity estimation



Comparative study of errors for an estimation based on an area of vicinity

Data selection still improves PCA-based estimation regarding surface gravity, but no longer for metallicity.
Results regarding effective temperature are worse with such a selection.

Though results are not very conclusive about a selection of relevant data as it is done here, this study has shown that some data is to be considered as noise regarding some parameters.
This leads now to further investigate the following points :

- How is relevant information embedded in the data ? (non-linearity)
- How can we achieve the information relative to local spectral vicinities in the data ?
- Is there a more appropriate distance measurement to link the two spaces of representation of the individuals ?