Quasar redshift determination through weighted PCA

Ludovic Delchambre

Extragalactic Astrophysics and Space Observations (AEOS)
Institute of Astrophysics and Geophysics
University of Liège, Belgium

July 9, 2015
GAGNES 2015

Liège team: Delchambre L., Surdej J.

Outline

Outline

Outline

Outline

Outline

QSO Classifier module

Goal

For each object classified as QSO by DSC, find:

- Redshift
- QSO type (type I/II or BAL)
- Continuum slope
- Emission lines EW
- A_0 extinction parameter

Implementation

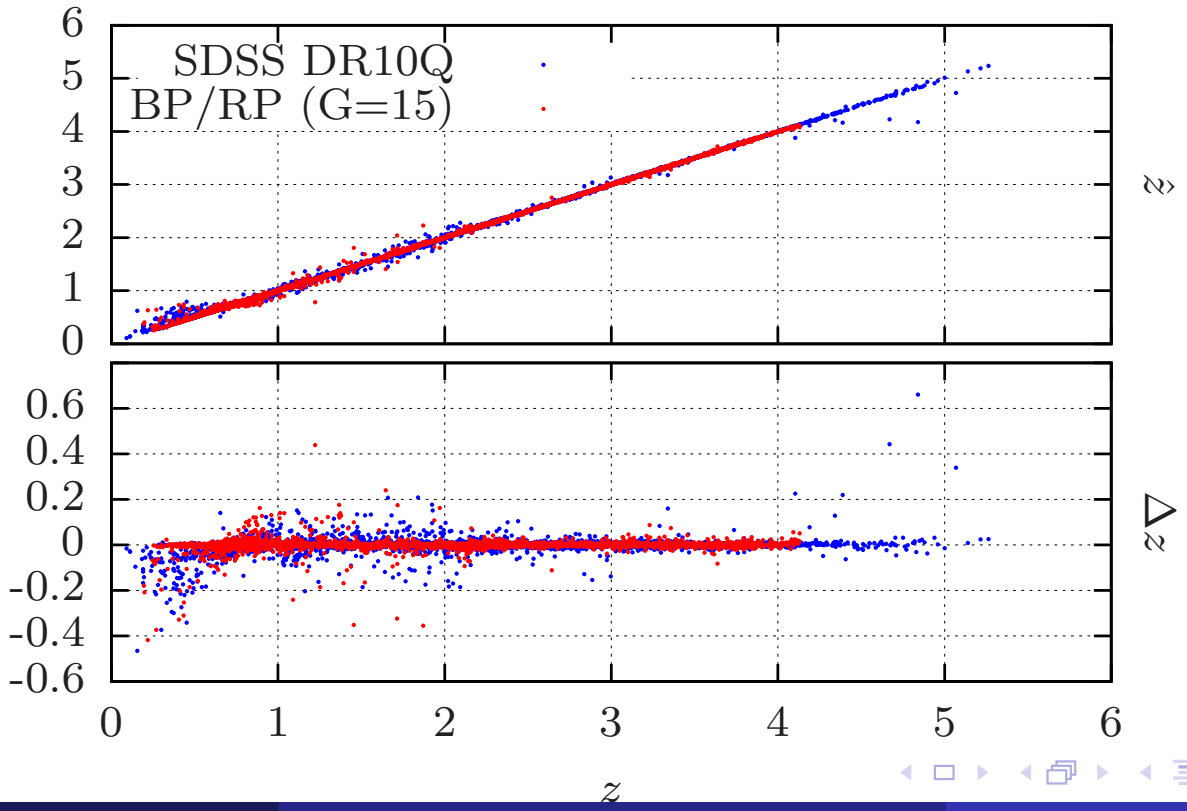
- KNN & ERT supervised learning methods
- Learning library: Semi-empirical Gaia BP/RP spectra based on SDSS DR10Q.

Redshift K-NN results

Limited G-mag

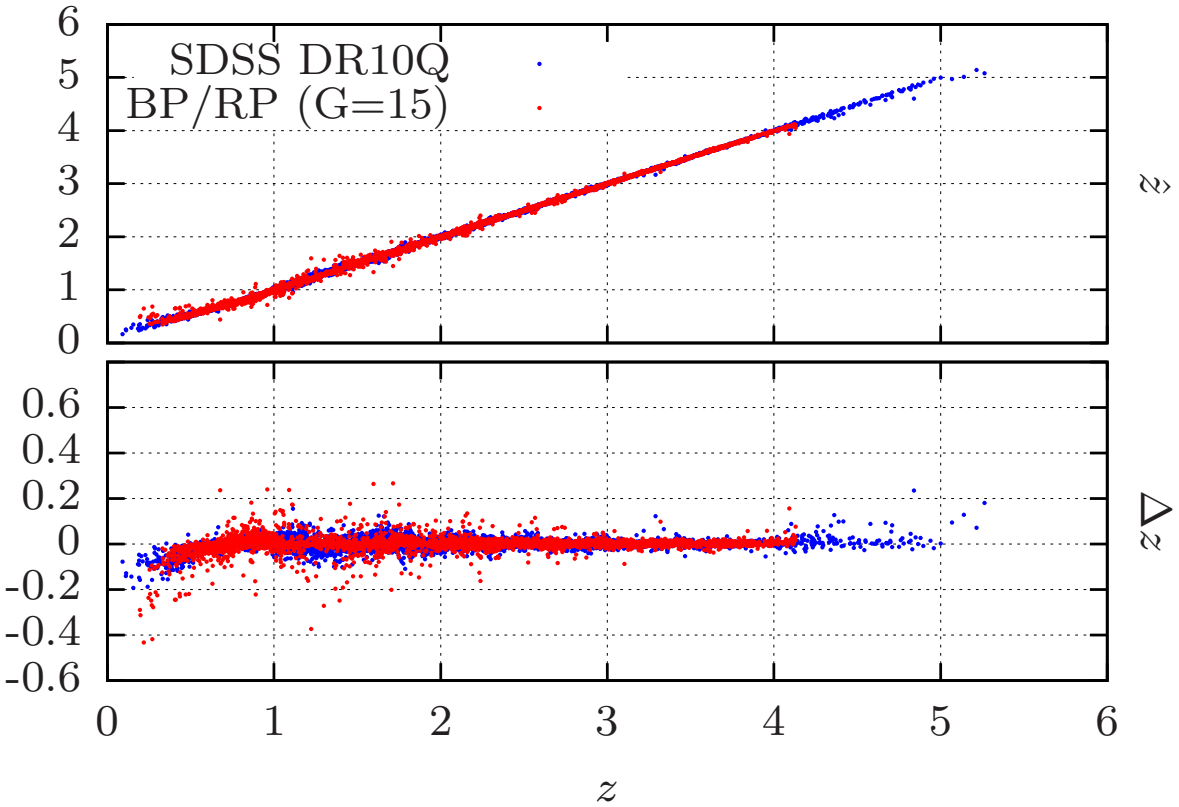
Only G=15 due to huge CPU ressources needed for simulations

	$\bar{z}$	σ_z
SDSS DR10Q	$6 \cdot 10^{-4}$	0.0151
BP/RP (G=15)	$3 \cdot 10^{-4}$	0.0154



Redshift ERT results

	$\bar{z}$	σ_z
SDSS DR10Q	$1 \cdot 10^{-4}$	0.0121
BP/RP (G=15)	$1 \cdot 10^{-4}$	0.0209



State of the art & end of the story?

Supervised learning methods approach

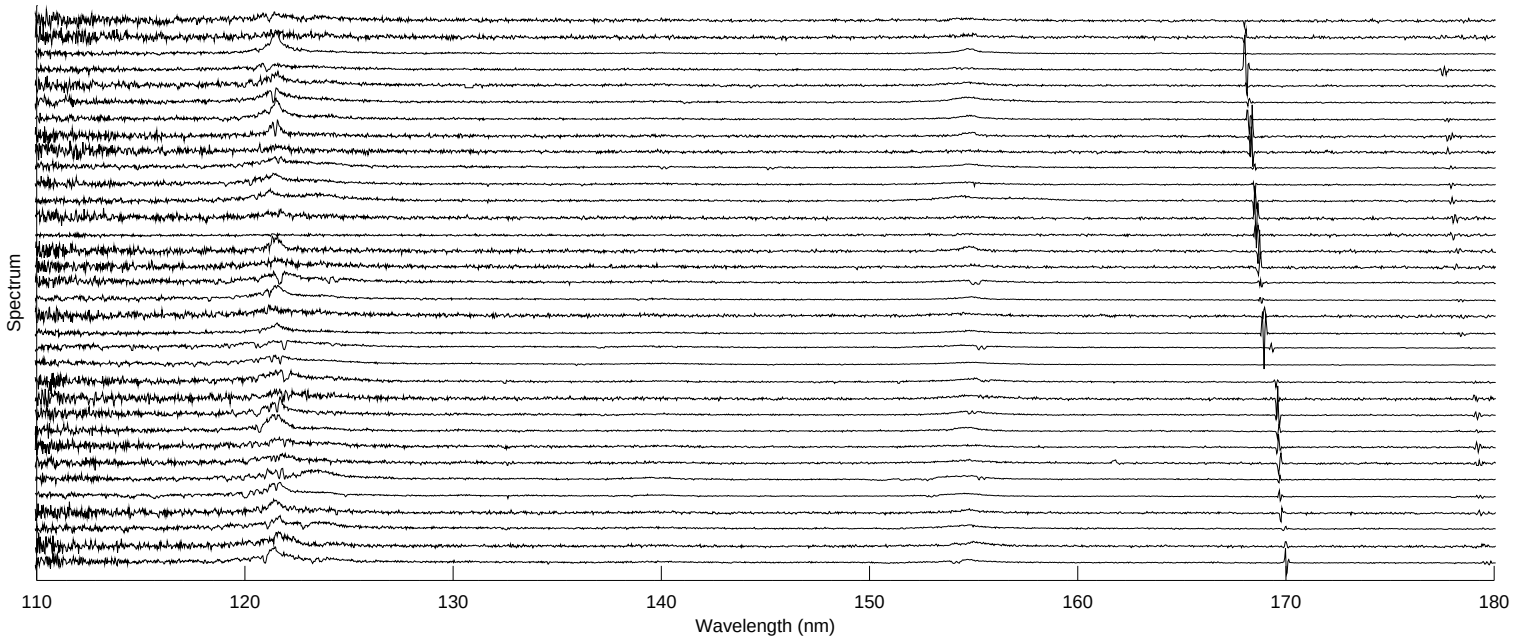
- **Pros**
  - Fast
  - Fairly good prediction (mainly QSO type)
  - Well supported and extensively used within DPAC
- **Cons**
  - Black box algorithms
  - Unavoidable bias/variance trade-off
  - Provide only a near-optimal solution (eg. in a χ^2 sense)

Redshift considerations

- APs strongly depend on z
- Line mismatch problem $\rightarrow$ interesting to have multiple estimates
- Not optimal using supervised learning methods

Outline

Principal Component Analysis



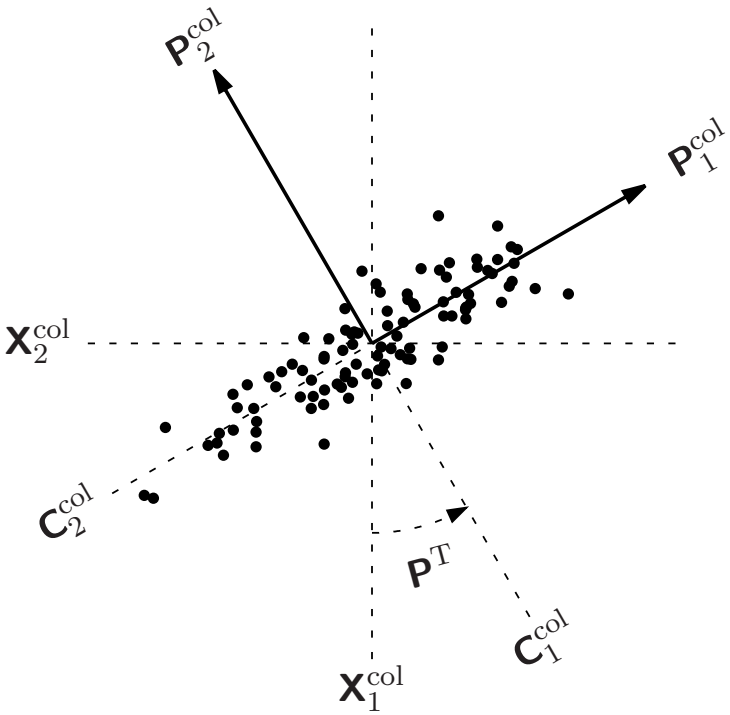
Question...

How can we extract a small set of templates from these data such that their linear combination explains at best the observed variance?

Principal component analysis

Goal

Find an orthogonal matrix $\mathbf{P}$ in $\mathbf{X} = \mathbf{PC}$ such that $\sigma^2 = \mathbf{CC}^T$ is diagonal and for which $\sigma_j^2 \leq \sigma_i^2$; $\forall i < j$.



Solution using SVD

Given the SVD of $\mathbf{X} \equiv \mathbf{U\Sigma V}^T$
We have $\mathbf{P} = \mathbf{U}$ and $\mathbf{C} = \mathbf{\Sigma V}^T$

PCA for spectra

$\mathbf{X}$: (Mean-subtracted) Spectral library
$\mathbf{P}$: Spectral (Principal) Components
$\mathbf{C}$: Spectral Coefficients

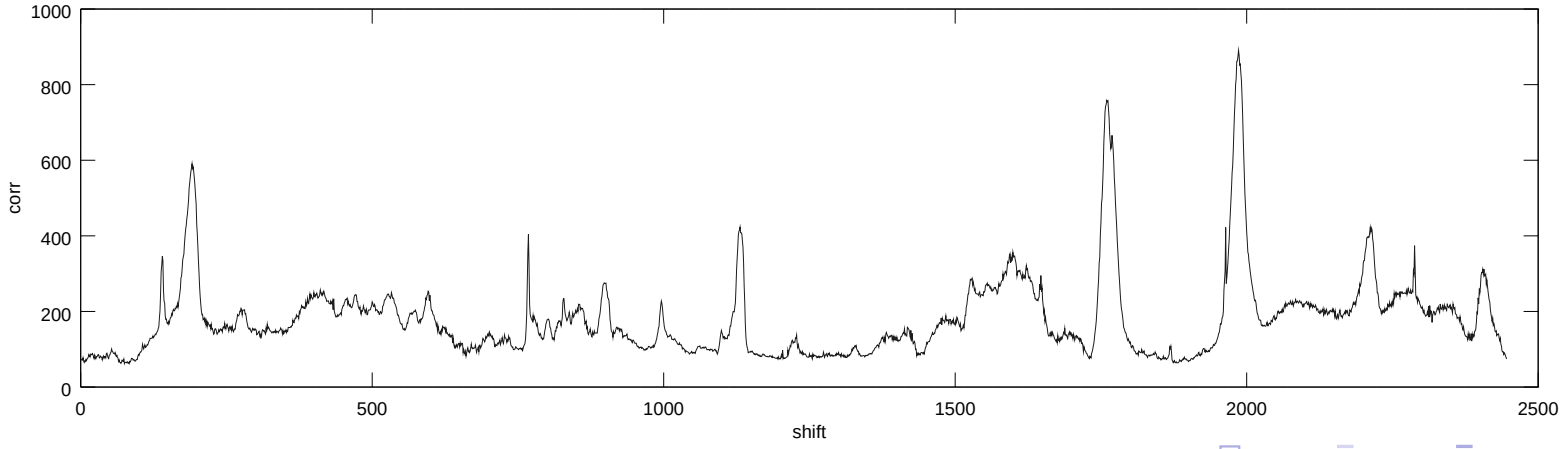
Phase correlation

Algorithm

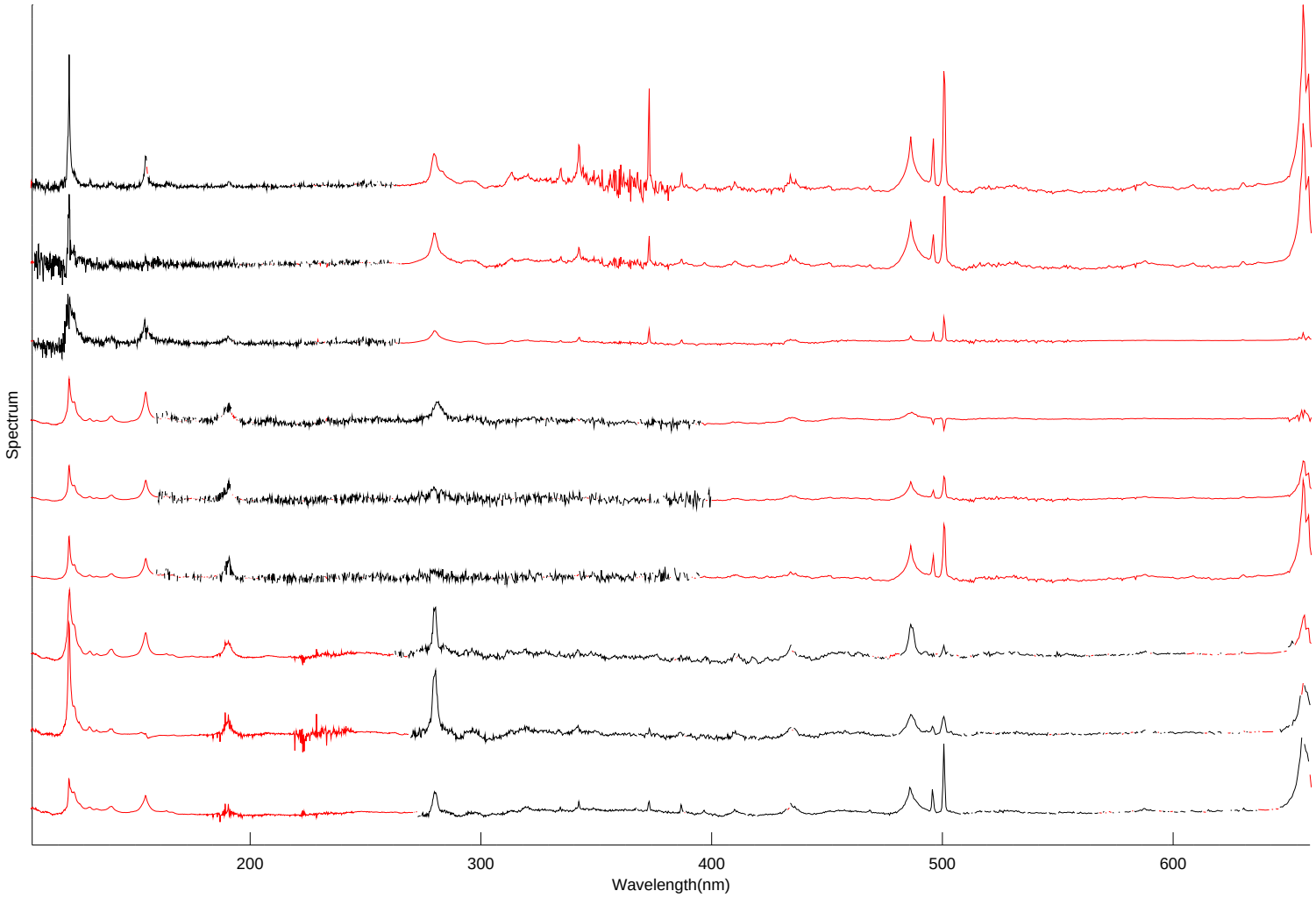
Find $\chi^2(z) = \|\vec{y}(z) - \mathbf{P}\vec{a}(z)\|^2$; $\forall z$ with $\vec{y}(z) \equiv$ shifted observation
Since $\mathbf{P}$ is orthogonal, we have $\chi^2(z) = \|\vec{y}(z)\|^2 - \|\vec{a}(z)\|^2$.
$\Rightarrow$ We seek to maximize $\|\vec{a}(z)\|^2$ with $\vec{a}(z) = \mathbf{P}^T\vec{y}(z)$.
In more details: $a_i(z) = \sum_j \mathbf{P}_{j,i} y_{j+z} \Leftrightarrow \mathcal{F}\vec{a} = \mathcal{F}\mathbf{P}^T\mathcal{F}\vec{y}^\star$

Practicalities

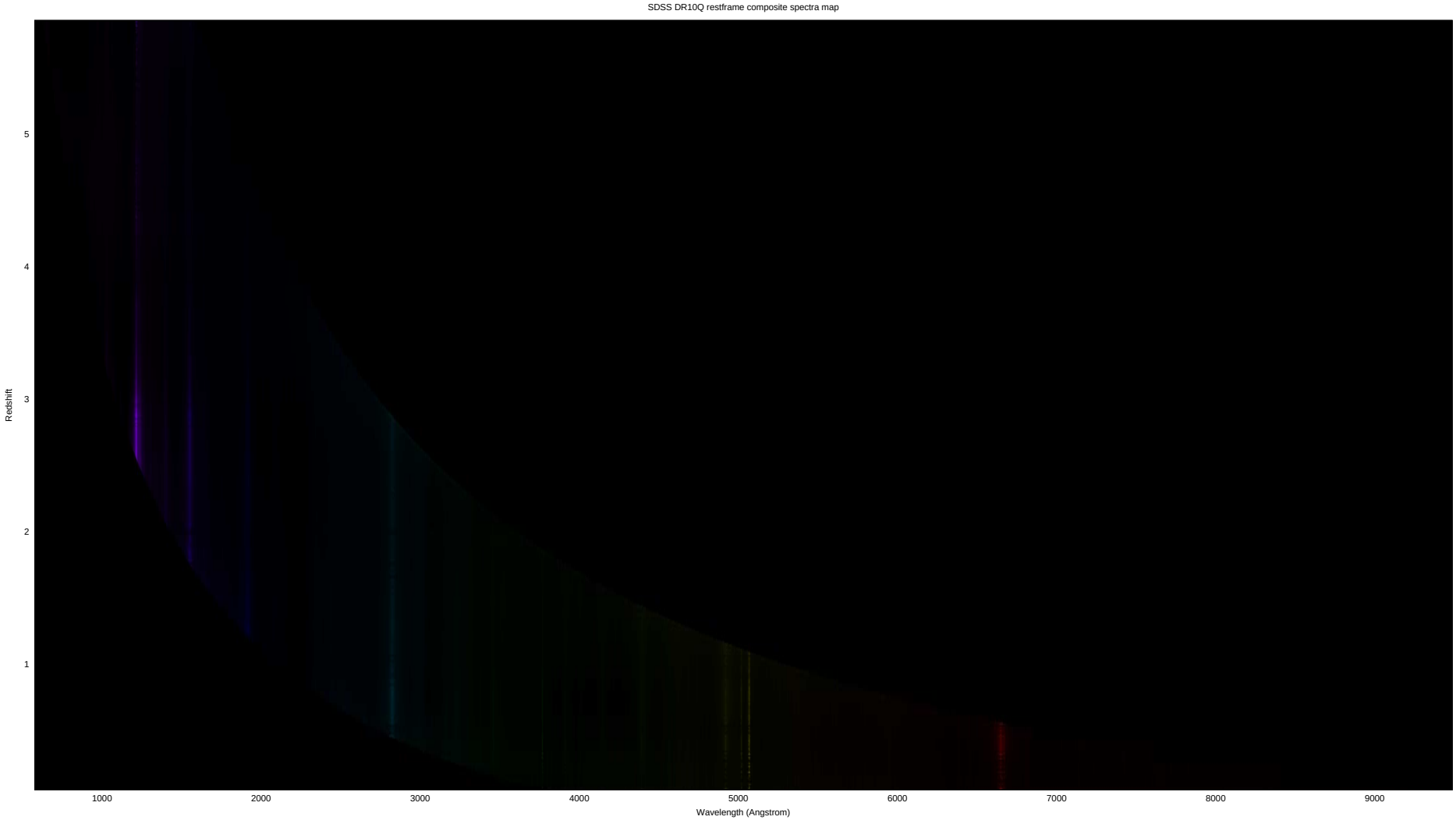
Continuum & mean spectrum subtraction



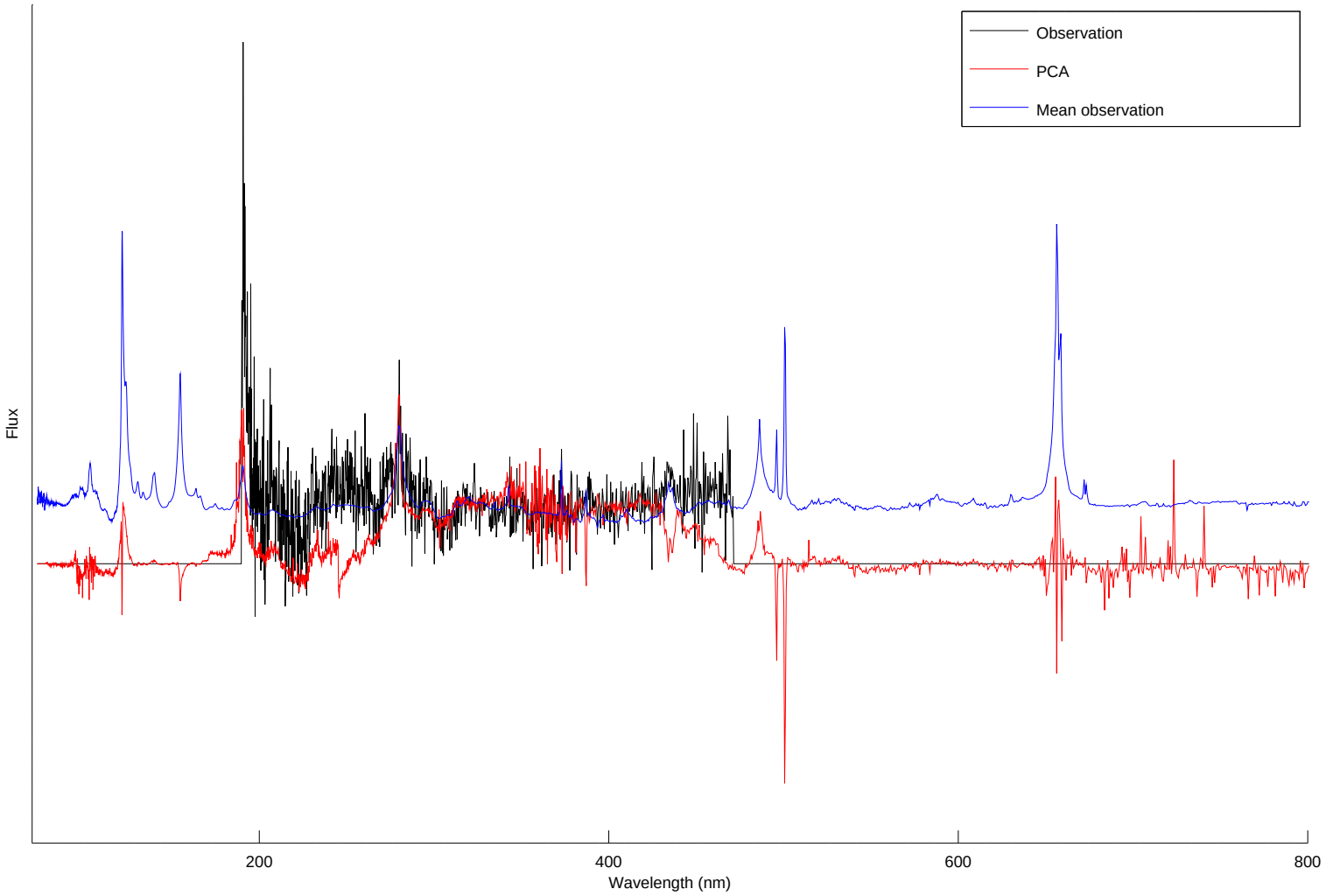
Principal components extrapolation



Principal components extrapolation



Windowed observations



Outline

Bailey implementation

Goal

Minimize $\chi^2 = \sum_{obs\ j} \left\| \mathbf{W}_j \mathbf{X}_j^{col} - \mathbf{W}_j \mathbf{P} \mathbf{C}_j^{col} \right\|^2$

EM Algorithm

(E-step) $\mathbf{C}_j^{col} = \left(\mathbf{P}^T \mathbf{W}_j^2 \mathbf{P} \right)^{-1} \mathbf{P}^T \mathbf{W}_j^2 \mathbf{X}_j^{col}$

(M-step) $\mathbf{P}_{ik} = \dfrac{\sum_j \mathbf{C}_{ik} \mathbf{W}_{ij}^2 \mathbf{X}_{ij}}{\sum_j \mathbf{C}_{ik} \mathbf{W}_{ij}^2 \mathbf{C}_{ik}}; \forall k$

Drawbacks

- Bad convergence and numerical stability problems
- Spectra have "negative emission lines" (eg. reversed Lyα)

New implementation[1]

Goal

Find $\mathbf{P}$ such that $\mathbf{P}^T \sigma^2 \mathbf{P}$ is diagonal.

where $\sigma^2 = \dfrac{(\mathbf{X} \circ \mathbf{W})(\mathbf{X} \circ \mathbf{W})^T}{\mathbf{W}\mathbf{W}^T}$

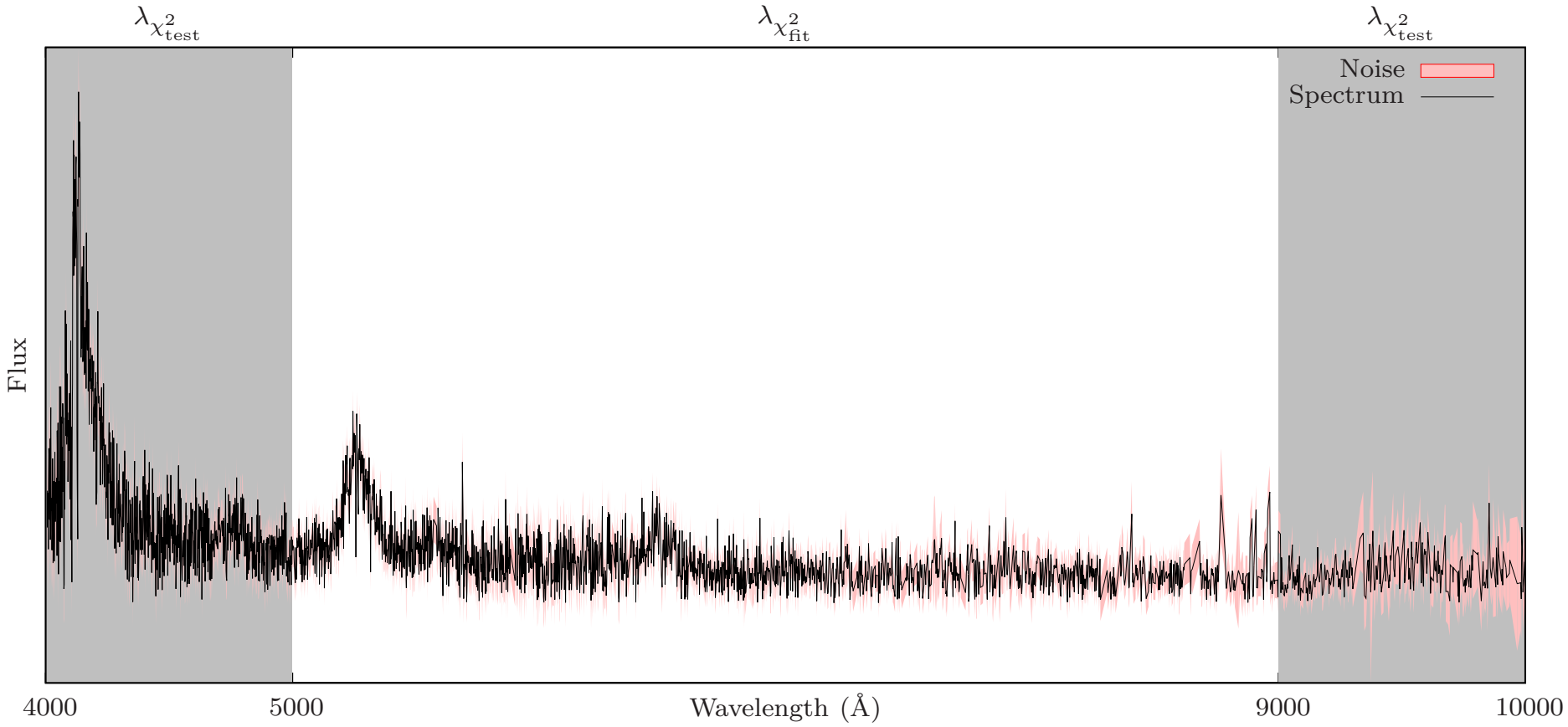
Power iteration algorithm

(1) Find dominant eigenvector (the one with the highest eigenvalue)
$\mathbf{u}^{(k)} = \sigma^2 \mathbf{u}^{(k-1)} = \sigma^{2^k} \mathbf{u}^{(0)}$, where $\mathbf{u}^{(0)} = rand()$

(2) Restart algorithm with $\sigma^{2'} = \sigma^2 - \lambda \mathbf{u}^{(k)} \otimes \mathbf{u}^{(k)}$
where $\lambda = \mathbf{u}^{(k)} \cdot \sigma^2 \mathbf{u}^{(k)}$

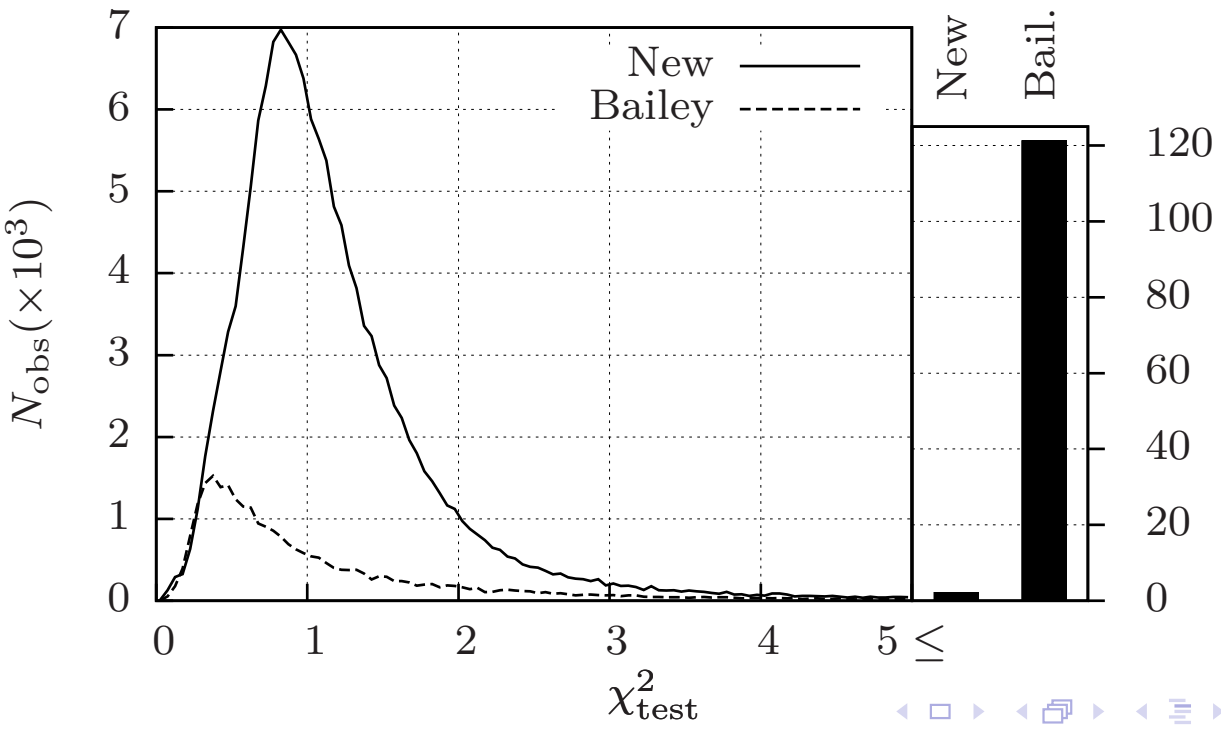
[1] Delchambre L.(2015), MNRAS, 446, 3545-3555

Comparison of PCA methods



Comparison of PCA methods

Stats over N=148,050 spectra	New	Bailey
Dataset χ^2_{fit}	0.107	0.094
Dataset χ^2_{test}	1.064	$8 \cdot 10^{12}$
Median χ^2_{test}	1.021	$8 \cdot 10^{4}$
Ratio of observations having $\chi^2_{\text{test}} \geq 5$	0.014	0.81



Outline

Weighted phase retrieval

Algorithm

Minimize $\chi^2(z) = \|\mathbf{W}\vec{y} - \mathbf{W}\mathbf{T}(z)\vec{a}(z)\|^2$; $\forall z$

with

$\mathbf{T}$, the (not necessary orthogonal) templates

$\mathbf{W}$, the observation weights.

Normal equations solution:

$$\vec{a}(z) = \left(\mathbf{T}^T(z)\mathbf{W}^2\mathbf{T}(z)\right)^{-1}\mathbf{T}^T(z)\mathbf{W}^2\vec{y}$$

$\Rightarrow$ Safest way to retrieve z

$\Rightarrow$ Solution used within SDSS-III using SVD

Drawback

Time complexity of $\mathcal{O}(N^2) \Rightarrow$ Too slow to be used within Gaia pipeline

Weighted phase retrieval

The good new

An $\mathcal{O}(N \log N)$ algorithm exist that is stable & highly-threadable.

N	N_T	Old	New
10^3	10	0.243 s	2.049 ms
10^4	**10**	**24.3 s**	**0.02 s**
10^6	10	67.5 h	2.49 s
10^9	10	7705 y	48 m
10^3	10^2	20.7 s	1.09 s
10^6	10^2	236 d	18 m
10^9	10^2	64697 y	13 d

Table : Time complexity regarding the various algorithm for various values of the parameters N and N_T on a 2.5Ghz CPU (ms=millisecond; s=second; m=minute; h=hour; d=day; y=year).

Weighted phase retrieval

The bad new

I'm a bit late in submitting the associated paper...

References

1. Bailer-Jones C. A. L. (2013), The Gaia astrophysical parameters inference system (Apsis). Pre-launch description, A&A, 559, A74
2. Bailey S.(2012), Principal component analysis with Noisy and/or missing data, PASP, 124, 919, 1015-1023
3. Delchambre L.(2015), Weighted principal component analysis: a weighted covariance eigendecomposition approach, MNRAS, 446, 3545-3555
4. Glazebrook K.(1997), Automatic redshift determination by use of principal component analysis. I. Fundamentals, ApJ, 492, 98-109
5. Pâris I.(2014), The Sloan Digital Sky Survey quasar catalog: tenth data release, A&A, 563, A54