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

1. QSOC status
1. QSOC status

2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses
Outline

1. QSOC status

2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses

3. Weighted Principal Component Analysis
Outline

1. QSOC status
2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses
3. Weighted Principal Component Analysis
4. Weighted phase retrieval
Outline

1. QSOC status
2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses
3. Weighted Principal Component Analysis
4. Weighted phase retrieval
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.
Limited G-mag

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

<table>
<thead>
<tr>
<th></th>
<th>$\bar{z}$</th>
<th>$\sigma_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDSS DR10Q</td>
<td>$6 \cdot 10^{-4}$</td>
<td>0.0151</td>
</tr>
<tr>
<td>BP/RP (G=15)</td>
<td>$3 \cdot 10^{-4}$</td>
<td>0.0154</td>
</tr>
</tbody>
</table>
Redshift ERT results

<table>
<thead>
<tr>
<th></th>
<th>$\bar{z}$</th>
<th>$\sigma_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDSS DR10Q</td>
<td>$1 \cdot 10^{-4}$</td>
<td>0.0121</td>
</tr>
<tr>
<td>BP/RP (G=15)</td>
<td>$1 \cdot 10^{-4}$</td>
<td>0.0209</td>
</tr>
</tbody>
</table>

![Graph showing redshift distribution]
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 $\chi^2$ sense)

Redshift considerations

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

1. QSOC status

2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses

3. Weighted Principal Component Analysis

4. Weighted phase retrieval
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 \( P \) in \( X = PC \) such that \( \sigma^2 = CC^T \) is diagonal and for which \( \sigma_j^2 \leq \sigma_i^2 \); \( \forall i < j \).

**Solution using SVD**

Given the SVD of \( X \equiv U\Sigma V^T \)
We have \( P = U \) and \( C = \Sigma V^T \)

**PCA for spectra**

\( X \): (Mean-subtracted) Spectral library
\( P \): Spectral (Principal) Components
\( C \): Spectral Coefficients
Phase correlation

Algorithm

Find $\chi^2(z) = \|\vec{y}(z) - \mathbf{P}\vec{a}(z)\|^2$; $\forall z$ with $\vec{y}(z)$ ≡ 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: $a_i(z) = \sum_j P_{j,i} y_{j+z} \Leftrightarrow \mathcal{F}\vec{a} = \mathcal{F}\mathbf{P}^T\mathcal{F}\vec{y}^*$

Practicalities

Continuum & mean spectrum subtraction
Principal components extrapolation
Principal components extrapolation
Windowed observations

Flux

Wavelength (nm)

Observation
PCA
Mean observation
Outline

1. QSOC status

2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses

3. Weighted Principal Component Analysis

4. Weighted phase retrieval

Delchambre L. (Ulg)

Redshift determination using WPCA

GAGNES 2015
Weighted Principal Component Analysis

Bailey implementation

**Goal**

Minimize \( \chi^2 = \sum_{\text{obs } j} \left\| W_j X_{j}^{\text{col}} - W_j PC_{j}^{\text{col}} \right\|^2 \)

**EM Algorithm**

(E-step) \( C_{j}^{\text{col}} = \left( P^T W_{j}^2 P \right)^{-1} P^T W_{j}^2 X_{j}^{\text{col}} \)

(M-step) \( P_{ik} = \frac{\sum_{j} C_{ik} W_{ij}^2 X_{ij}}{\sum_{j} C_{ik} W_{ij}^2 C_{ik}} \); \( \forall k \)

**Drawbacks**

- Bad convergence and numerical stability problems
- Spectra have "negative emission lines" (eg. reversed Ly\(\alpha\))
Goal

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

where $\sigma^2 = \frac{(X \circ W)(X \circ W)^T}{WW^T}$

Power iteration algorithm

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

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

---

1Delchambre L.(2015), MNRAS, 446, 3545-3555
Comparison of PCA methods

Flux

Wavelength (Å)

\[ \chi^2_{\text{fit}} \]

\[ \chi^2_{\text{test}} \]

\[ \lambda \]

Delchambre L. (Ulg)

Redshift determination using WPCA

GAGNES 2015
Comparison of PCA methods

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

![Graph showing distribution of $\chi^2_{\text{test}}$]
Outline

1. QSOC status

2. Redshift determination using Principal Component Analysis
   - Principal Component Analysis
   - Phase correlation
   - Method weaknesses

3. Weighted Principal Component Analysis

4. Weighted phase retrieval
## Weighted phase retrieval

**Algorithm**

Minimize $\chi^2(z) = \| W\tilde{y} - WT(z)\bar{a}(z) \|^2$ ; $\forall z$

with

- $T$, the (not necessary orthogonal) templates
- $W$, the observation weights.

Normal equations solution:

$$\bar{a}(z) = (T^T(z)W^2T(z))^{-1}T^T(z)W^2\tilde{y}$$

⇒ Safest way to retrieve $z$
⇒ Solution used within SDSS-III using SVD

**Drawback**

Time complexity of $O(N^2)$ ⇒ Too slow to be used within Gaia pipeline
The good news

An $O(N \log N)$ algorithm exists that is stable and highly-threadable.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$N_T$</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^3$</td>
<td>10</td>
<td>0.243 s</td>
<td>2.049 ms</td>
</tr>
<tr>
<td>$10^4$</td>
<td>10</td>
<td>24.3 s</td>
<td>0.02 s</td>
</tr>
<tr>
<td>$10^6$</td>
<td>10</td>
<td>67.5 h</td>
<td>2.49 s</td>
</tr>
<tr>
<td>$10^9$</td>
<td>10</td>
<td>7705 y</td>
<td>48 m</td>
</tr>
<tr>
<td>$10^3$</td>
<td>$10^2$</td>
<td>20.7 s</td>
<td>1.09 s</td>
</tr>
<tr>
<td>$10^6$</td>
<td>$10^2$</td>
<td>236 d</td>
<td>18 m</td>
</tr>
<tr>
<td>$10^9$</td>
<td>$10^2$</td>
<td>64697 y</td>
<td>13 d</td>
</tr>
</tbody>
</table>

**Table**: Time complexity regarding the various algorithms 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...

TO BE CONTINUED....
2. Bailey S. (2012), Principal component analysis with Noisy and/or missing data, PASP, 124, 919, 1015-1023