



# On Class Visualisation for High Dimensional Data: Exploring Scientific Data Sets

---

Ata Kabán<sup>1</sup>, Jianyong Sun<sup>1,2</sup>,

Somak Raychaudhury<sup>2</sup> & Louisa Nolan<sup>2</sup>

<sup>1</sup>School of Computer Science

<sup>2</sup>School of Physics & Astronomy

The University of Birmingham



# Motivation

---

- Clustering & visualisation
  - Widespread data analysis principles
- Strategies
  - Machine Learning approach
    - Map the data first
    - Detect clusters afterwards
  - Data Mining approach
    - Cluster the data first
    - Visualise the structure afterwards
- A simultaneous approach



# Outline

---

- A latent variable model for joint clustering & class projection
  - Multi-objective interpretation
- MAP parameter estimation
- Evaluation & accommodating new data points
- A modification to take into account known measurement errors
- Experiments & applications
  - Visualisation & exploratory analysis of high-dimensional scientific data sets



# The model

---

$\mathbf{d}_n \in \mathcal{R}^T, n = 1, \dots, N$       multivariate data points

$\mathbf{x}_n \in \mathcal{R}^2$       2D latent points

$$p(\mathbf{d}_n | \mathbf{x}_n) = \prod_t p(d_{tn} | \mathbf{x}_n)$$

where  $p(d_{tn} | \mathbf{x}_n) = \sum_k N(d_{tn} | \mu_{tk}, v_{tk}) P_{c_k}(k | \mathbf{x}_n)$

$$\text{where } P_{c_k}(k | \mathbf{x}_n) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}_n - \mathbf{c}_k)^2\}}{\sum_{k'} \exp\{-\frac{1}{2}(\mathbf{x}_n - \mathbf{c}_{k'})^2\}}$$

Smoothness priors:  $p(\mathbf{x}_n) = N(\mathbf{x}_n | \mathbf{0}, \alpha \mathbf{I}); p(\mathbf{c}_k) = N(\mathbf{c}_k | \mathbf{0}, \beta \mathbf{I})$

Complexity reducing priors:  $p(v_{tk}) = \gamma e^{-\gamma v_{tk}}$



# Parameter estimation

---

$$\begin{aligned} L &= \sum_{n,t} \log \sum_k N(d_{tn} | \mu_{tk}, v_{tk}) P_{\mathbf{c}_k}(k | \mathbf{x}_n) + \sum_n \log P(\mathbf{x}_n) + \sum_k \log P(\mathbf{c}_k) + \sum_{t,k} P(v_{tk}) \\ &\geq \sum_{n,t,k} r_{ktn} \{ \log N(d_{tn} | \mu_{tk}, v_{tk}) + \log P_{\mathbf{c}_k}(k | \mathbf{x}_n) - \log r_{ktn} \} + \\ &\quad + \sum_n \log P(\mathbf{x}_n) + \sum_k \log P(\mathbf{c}_k) + \sum_{t,k} P(v_{tk}) \end{aligned}$$

where  $r_{ktn} > 0$ ,  $\sum_k r_{ktn} = 1$

# Parameter estimation

$$\begin{aligned}
 L &= \sum_{n,t} \log \sum_k N(d_{tn} | \mu_{tk}, v_{tk}) P_{\mathbf{c}_k}(k | \mathbf{x}_n) + \sum_n \log P(\mathbf{x}_n) + \sum_k \log P(\mathbf{c}_k) + \sum_{t,k} P(v_{tk}) \\
 &\geq \sum_{n,t,k} r_{ktn} \{ \log N(d_{tn} | \mu_{tk}, v_{tk}) + \log P_{\mathbf{c}_k}(k | \mathbf{x}_n) - \log r_{ktn} \} + \\
 &\quad + \sum_n \log P(\mathbf{x}_n) + \sum_k \log P(\mathbf{c}_k) + \sum_{t,k} P(v_{tk}) \quad \text{where } r_{ktn} > 0, \sum_k r_{ktn} = 1
 \end{aligned}$$

$$\text{Term}_1 = \sum_{n,t,k} r_{ktn} \log N(d_{tn} | \mu_{tk}, v_{tk}) - \gamma \sum_{t,k} v_{tk} + \text{const}$$

$$\begin{aligned}
 \text{Term}_2 &= \sum_{n,t,k} r_{ktn} \{ \log P_{\mathbf{c}_k}(k | \mathbf{x}_n) - \log r_{ktn} \} + \sum_n \log P(\mathbf{x}_n) + \sum_k \log P(\mathbf{c}_k) \\
 &= \sum_{n,k} -KL(\mathbf{r}_{\cdot,t,n} \| P_{\mathbf{c}_k}(\cdot | \mathbf{x}_n)) - \alpha \sum_n \|\mathbf{x}_n\|^2 - \beta \sum_k \|\mathbf{c}_k\|^2 + \text{const}.
 \end{aligned}$$

# Parameter estimation

$$\text{Term}_1 = \sum_{n,t,k} r_{ktn} \log N(d_{tn} | \mu_{tk}, v_{tk}) - \gamma \sum_{t,k} v_{tk} + \text{const}$$

This is like a “Modular mixture” [Attias] essentially an “aspect Gaussian” [T Hoffman] model log-likelihood

➤ useful for clustering of high-dimensional data

$$\begin{aligned} \text{Term}_2 &= \sum_{n,t,k} r_{ktn} \{ \log P_{\mathbf{c}_k}(k | \mathbf{x}_n) - \log r_{ktn} \} + \sum_n \log P(\mathbf{x}_n) + \sum_k \log P(\mathbf{c}_k) \\ &= \sum_{n,k} -KL(\mathbf{r}_{.,t,n} \| P_{\mathbf{c}_k}(\cdot | \mathbf{x}_n)) - \alpha \sum_n \|\mathbf{x}_n\|^2 - \beta \sum_k \|\mathbf{c}_k\|^2 + \text{const}. \end{aligned}$$

This is like a “Parametric Embedding” [T Iwata et al.] objective

➤ useful for embedding & visualising class posteriors in 2D



# Algorithm

---

- E-step

$$r_{ktn} = \frac{N(d_{tn} | \mu_{tk}, v_{tk}) P_{\mathbf{c}_k}(k | \mathbf{x}_n)}{\sum_{k'} N(d_{tn} | \mu_{tk'}, v_{tk'}) P_{\mathbf{c}_{k'}}(k' | \mathbf{x}_n)}$$

- M-step

$$\mu_{tk} = \frac{\sum_n r_{ktn} d_{tn}}{\sum_n r_{ktn}}; \quad v_{tk} = \frac{\sum_n r_{ktn}}{\sum_n r_{ktn} (d_{tn} - \mu_{tk})^2 + 2\gamma}$$

$$\frac{\partial}{\partial \mathbf{x}_n} = \sum_k (\mathbf{c}_k - \mathbf{x}_n) \sum_t (r_{ktn} - P_{\mathbf{c}_k}(k | \mathbf{x}_n)) - \alpha \mathbf{x}_n = \mathbf{0}$$

$$\frac{\partial}{\partial \mathbf{c}_k} = \sum_n (\mathbf{x}_n - \mathbf{c}_k) \sum_t (r_{ktn} - P_{\mathbf{c}_k}(k | \mathbf{x}_n)) - \beta \mathbf{c}_k = \mathbf{0}$$



# Accommodating new points

---

- Empirical Bayes test likelihood
  - Integrating over the empirical distribution of the latent variable  $\frac{1}{N} \sum_n \delta(\mathbf{x} - \mathbf{x}_n)$

$$p(\mathbf{d}_{test}) = \frac{1}{N} \sum_n p(\mathbf{d}_{test} | \mathbf{x}_n)$$

- Visualisation

$$\mathbf{x}_{test} = \arg \max_{\mathbf{x}} p(\mathbf{d}_{test} | \mathbf{x})$$

- This is a convex optimisation

# Taking into account known measurement errors

- We want to use this method for exploring of *scientific* data
  - Nice that we can deal with high-dimensional data
  - In addition, often we have knowledge of the errors in the measurements
    - Many of the traditional methods would need to disregard them
    - Danger of finding false interesting patterns

So,

$$p(y_{tn}|k) = \int dd_{tn} \mathcal{N}(y_{tn}|d_{tn}, 1/\sigma_{tn}^2) \mathcal{N}(d_{tn}|\mu_{tk}, v_{tk}) = \mathcal{N}(y_{tn}|\mu_{tk}, (\sigma_{tn}^2 + 1/v_{tk})^{-1})$$

error model  $d_{tn}$  become hidden variables



# Modified algorithm

---

- E-step

$$r_{ktn} = \frac{\mathcal{N}(y_{tn} | \mu_{tk}, (\sigma_{tn}^2 + 1/v_{tk})^{-1}) P_{\mathbf{c}_k}(k | \mathbf{x}_n)}{\sum_{k'} \mathcal{N}(y_{tn} | \mu_{tk}, \sigma_{tn}^2 + 1/v_{tk}) P_{\mathbf{c}_k}(k | \mathbf{x}_n)}$$

- M-step

$$\mu_{tk} = \frac{\sum_n y_{tn} r_{ktn} / (\sigma_{tn}^2 + 1/v_{tk})}{\sum_n r_{ktn} / (\sigma_{tn}^2 + 1/v_{tk})}$$

$$\frac{\partial}{\partial \log v_{tk}} = \frac{1}{v_{tk}} \sum_n r_{ktn} \left\{ \frac{1}{2(\sigma_{tn}^2 + 1/v_{tk})} - \frac{(d_{tn} - \mu_{tk})^2}{2(\sigma_{tn}^2 + 1/v_{tk})^2} \right\} - \gamma v_{tk} = 0$$

& the  $\mathbf{x}_n$  and  $\mathbf{c}_k$  updates are as before



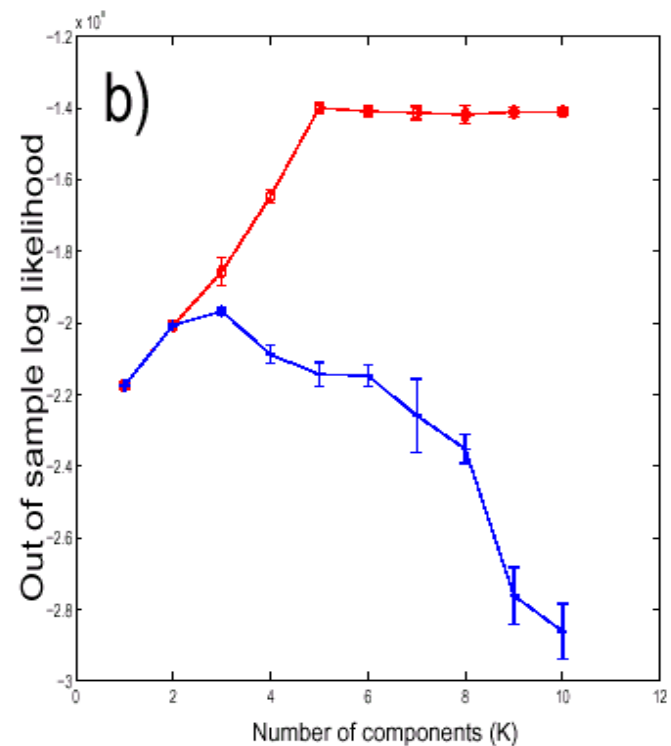
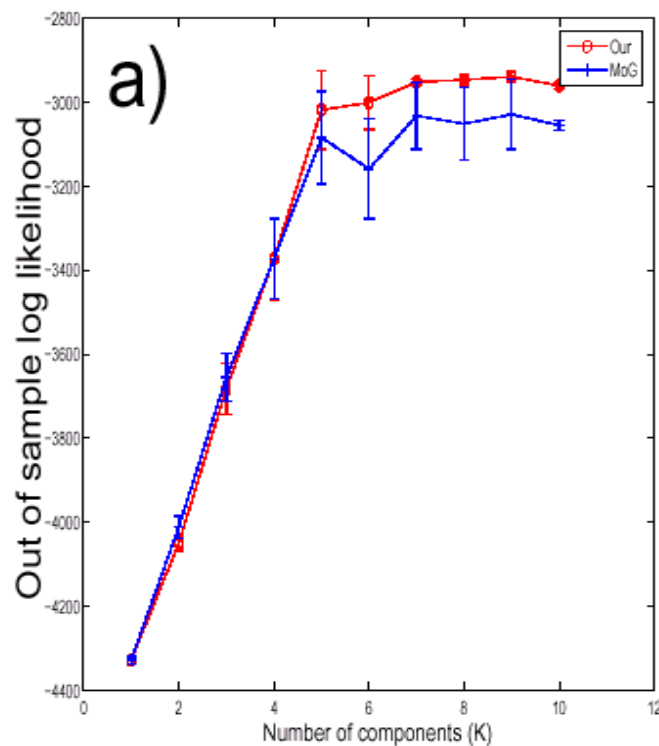
# Experiments

---

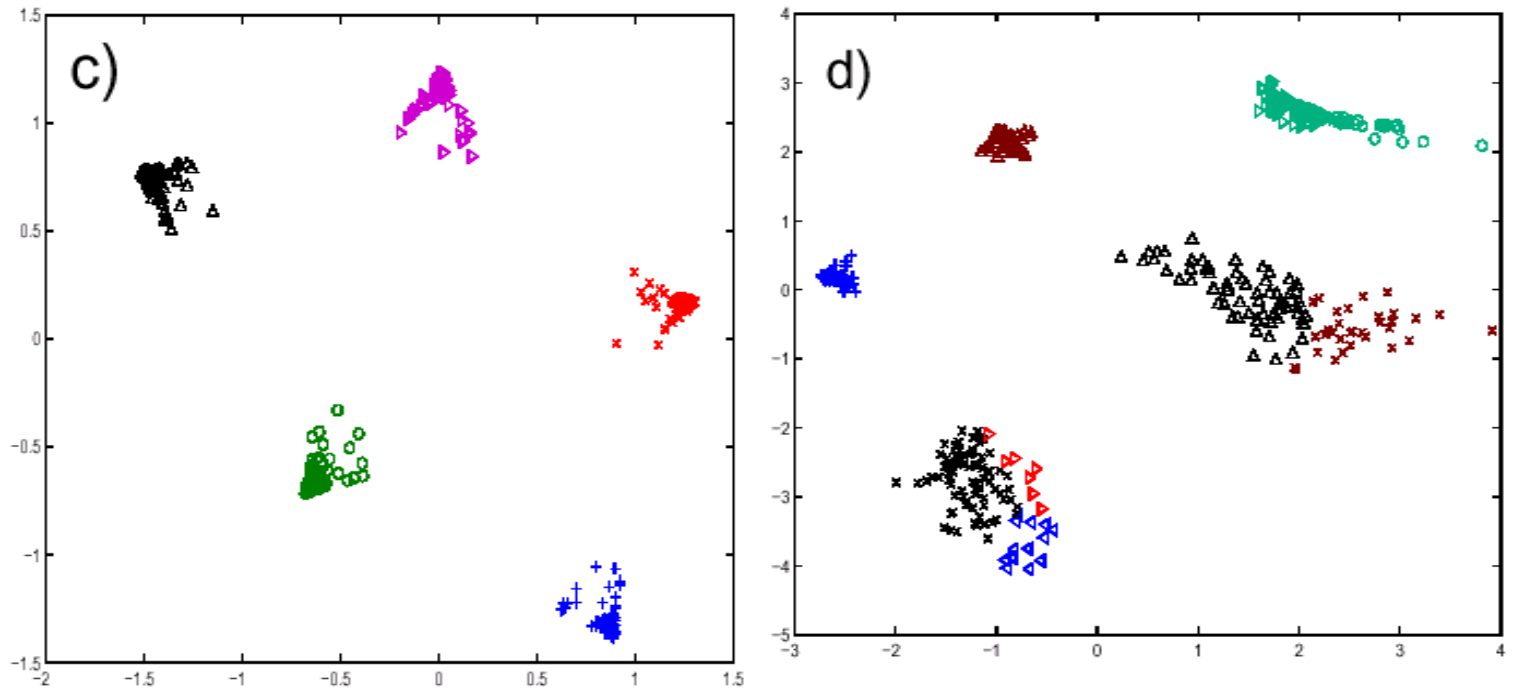
- Hyper-parameters (initially let  $K$ =big, e.g. 10)
  - $\alpha = \beta = 1$  was used;  $\gamma$  located by Cross Validation
  - Due to the prior, part of the unnecessary clusters become empty
- Number of clusters  $K$  determined by Cross-Validation, keeping the hyp's fixed
- Parameter initialisation from K-means
- Each experiment repeated 20 times to get a good-enough local optimum prior to the testing phase

# Toy data

- (a): 6D mixture of 5 independent Gaussians; 300 points
- (b): 300D mixture of 5 independent Gaussians; 300 points



## Visualisation of the five 300D clusters



c) The final visual image at  $K=5$

d) An intermediate visual image, with  $K=8 > 5$  non-empty model components

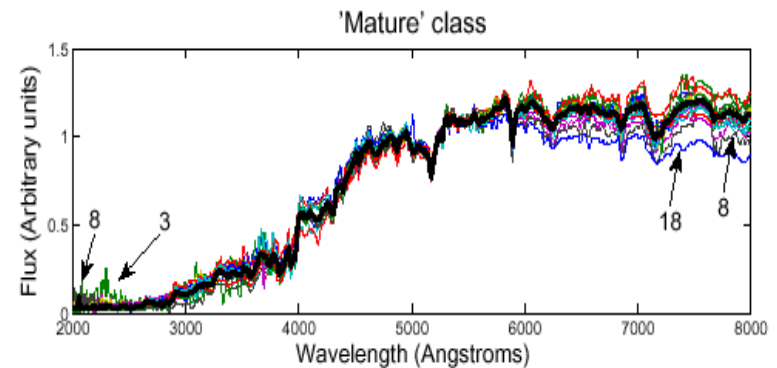
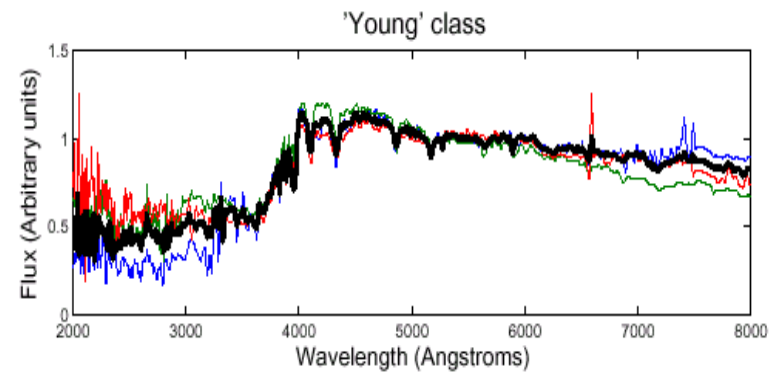
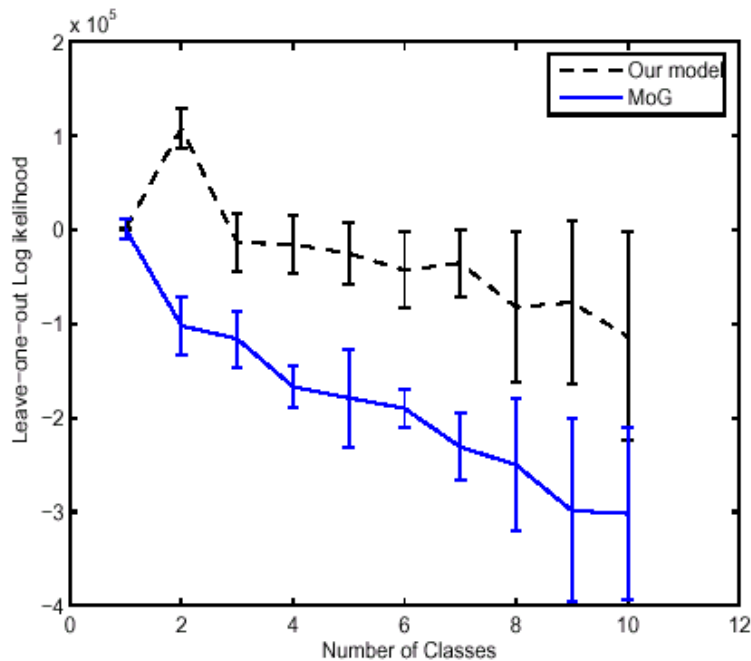


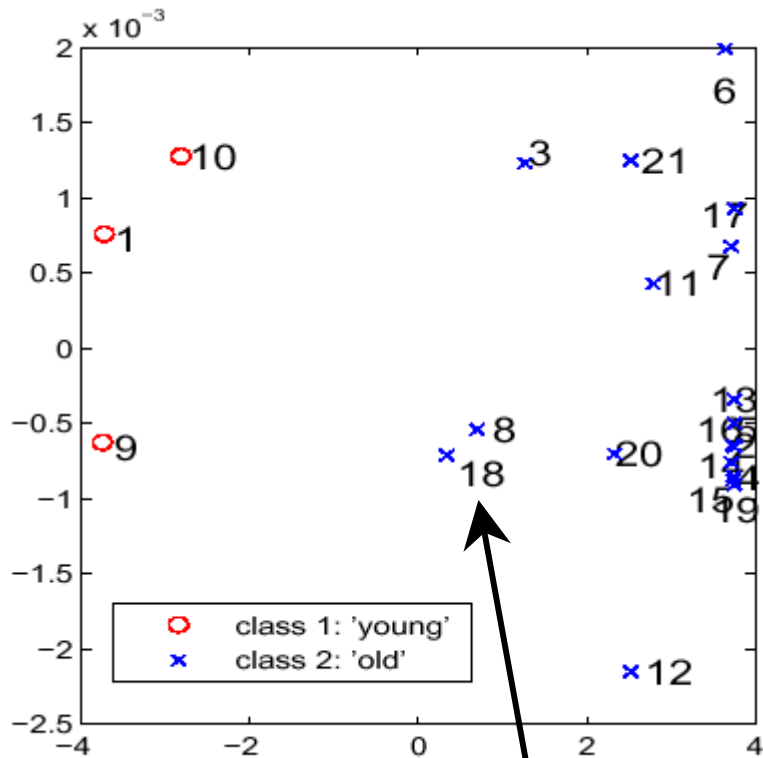
# Exploring scientific data sets

---

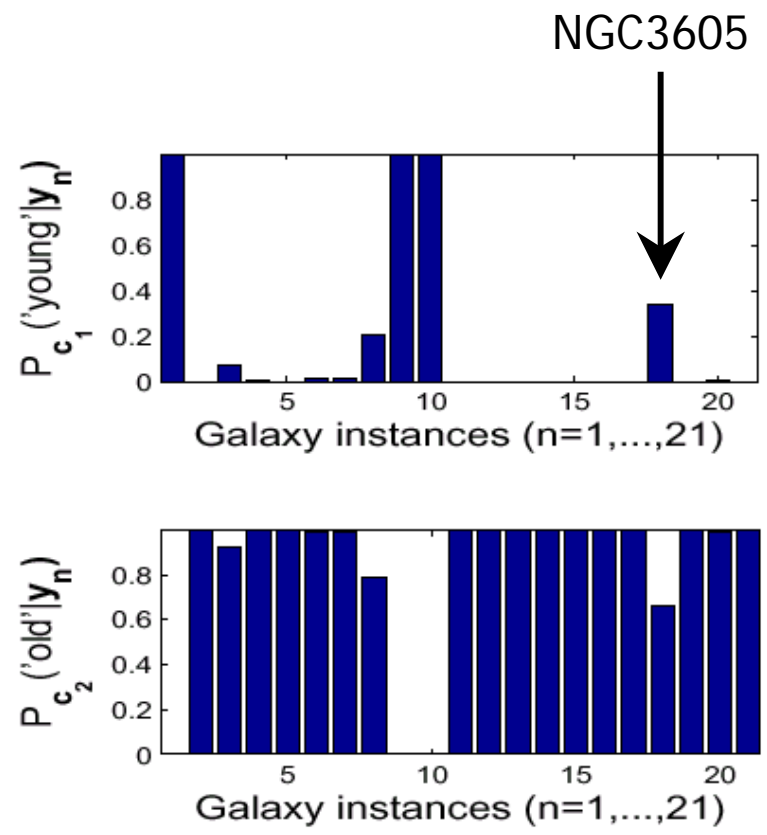
- Observed spectra of 21 early-type galaxies
  - Rare detailed coverage in wavelength: 348 flux measurements in the range of 2000-8000 Angstroms, in equal bins
  - Observational errors associated to each value, due to known instrumental characteristics and calibration
  - Pilot data set for a study in galaxy evolution
- Gene expression arrays
  - Benchmark data set
  - 62 samples of cancer vs normal colon tissues

# Visualisation of observed spectra of early-type galaxies





NGC3605

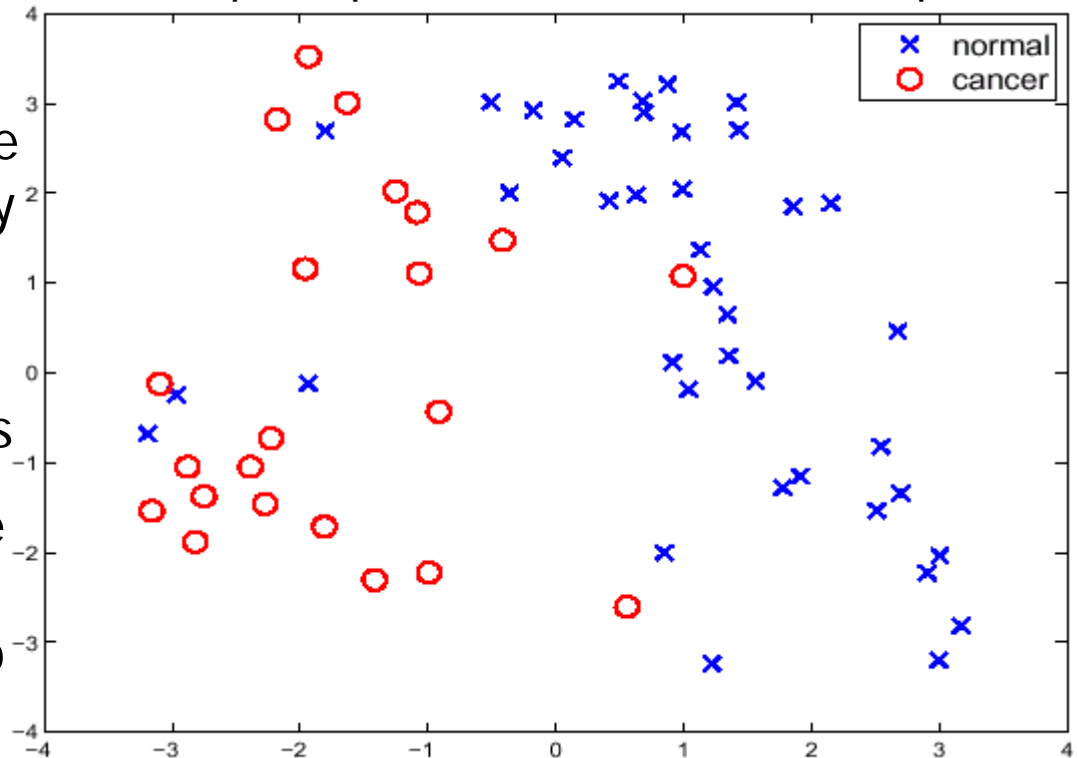


According to recent astrophysical analysis, although 85% of the stellar mass of NGC3605 is associated with an old (9-12Gyr) stellar population, it does contain a younger stellar population too, at approx. 1Gyr

# Visual analysis of gene expressions

- Expression levels of thousands of genes
- Difficulty: the sample size is of the order of tens only
- Eliminating genes with little variation still leaves us with hundreds of genes
- Extensive research in the literature, both on classification and trying to find interpretations

ColonCancer (62 high-dim samples) – the true labels superimposed with the obtained x points





# Conclusions

---

- We presented a method of class visualisation for high-dimensional data
- Integrates both clustering and class projection objectives into a consistent probabilistic model
- Measurement uncertainty straightforwardly included in the model
- We demonstrated applications in two scientific data analysis domains