A Medley of Mixtures

John Hinde

Statistics Group,
School of Mathematics, Statistics and Applied Mathematics
National University of Ireland, Galway
john.hinde@nuigalway.ie

Research Supported by SFI Award 07/MI/012

Australasian Region IBS, Hobart, Tasmania

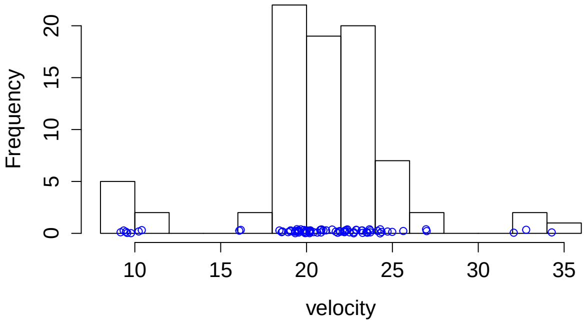
30 November 2015

Summary

1. Motivating Example
2. Mixture models
3. Fish growth studies
4. Mixtures of Mixed Models: *Time Course Microarray*
5. Clustering sugar cane SNP data
6. Acknowledgements
7. References

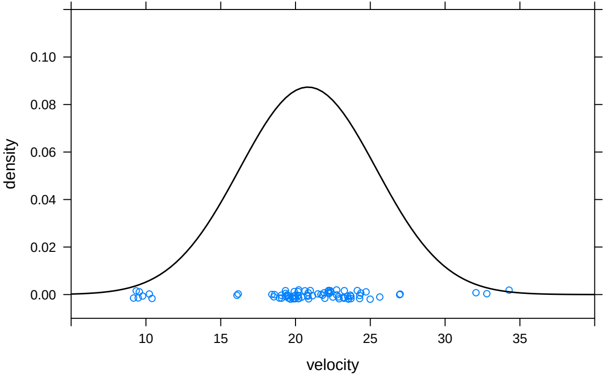
Galaxy Data – Recession Velocities

Recession velocities (in 10^3 km/s) of 82 galaxies.



Galaxy Data – Fitted Normal Density

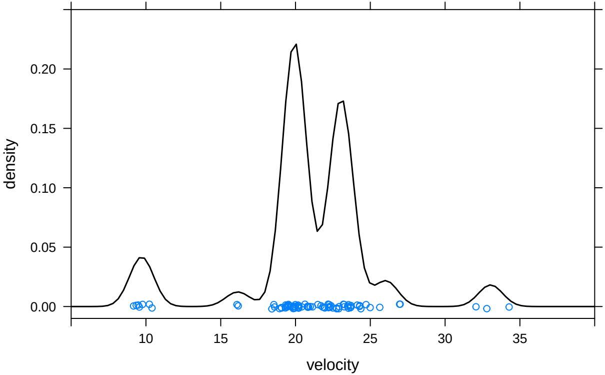
Single normal density based on sample mean and standard deviation



Is there any evidence of clustering?

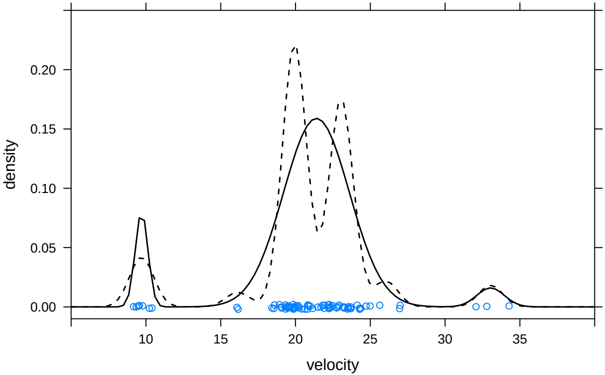
Galaxy Data – Fitted Mixture Model

Mixture of normal densities with equal variances



Galaxy Data – Fitted Mixtures Models

Mixtures of normal densities



——— three components with unequal variances

– – – – six components with equal variances

Mixture Models

- response y (uni/multivariate)
- explanatory variables $\mathbf{x}$

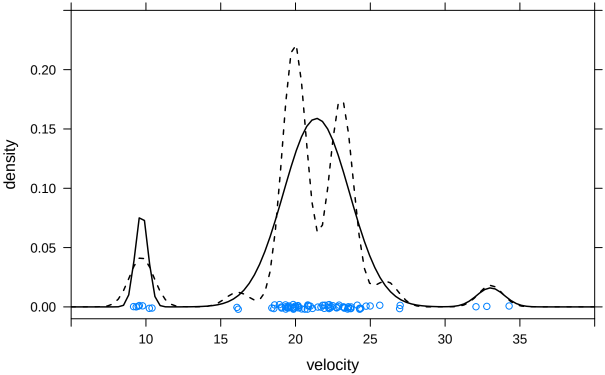
***K*-component mixture**

$$f(y \mid \boldsymbol{\Theta},\ \mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(y \mid \boldsymbol{\theta}_k,\, \mathbf{x})$$

- f_k — component densities (often same form)
- π_k — component probabilities ($\sum_k \pi_k = 1$)
- $\boldsymbol{\theta}_k$ — component parameter vectors (some may be equal across components)

Galaxy Data – Fitted Mixtures Models

Mixtures of normal densities



——— : three components with equal variances

– – – – : six components with unequal variances

Estimation — EM algorithm

Likelihood — *n* observations

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\Theta},\ \mathbf{x}_i) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(y_i \mid \boldsymbol{\theta}_k,\, \mathbf{x}_i)$$

Estimation for finite mixture conveniently viewed as EM algorithm.

E-Step: Calculate component weights w_{ik} – the posterior probability that observation y_i comes from component k (useful for **clustering)**:

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_\ell \pi_\ell f_{i\ell}}$$

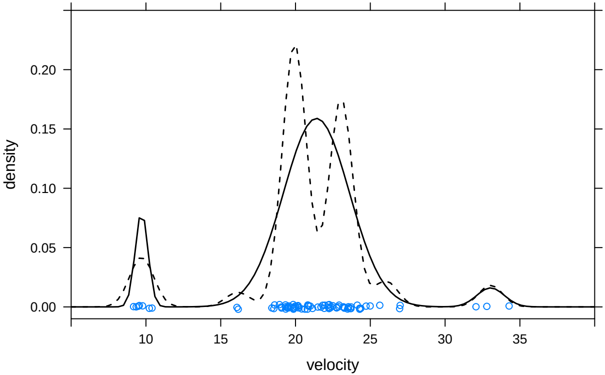
M-step:

- component parameters: estimate $\widehat{\boldsymbol{\theta}}_k$ from $(y_i, \mathbf{x}_i)$ with weights w_{ik}
- component proportions

$$\widehat{\pi}_k = \frac{\sum_{i=1}^{n} w_{ik}}{n}$$

Galaxy Data – Fitted Mixtures Models

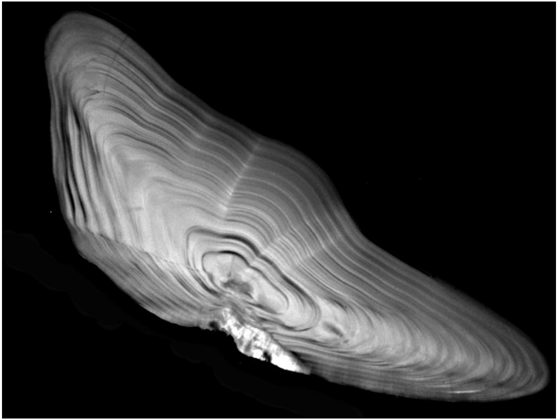
Mixtures of normal densities



——— : three components with equal variances
– – – – : six components with unequal variances

Fish ageing

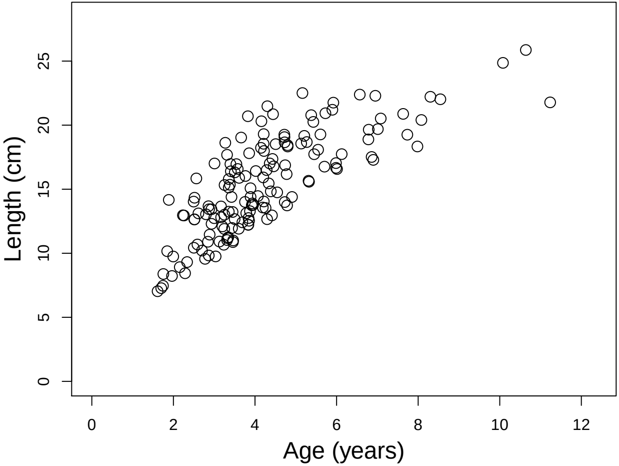
Count rings on sectioned otolith (ear bone)



Courtesy of Irish Marine Institute

Fish growth

Interested in the relationship between age and length or weight



Describing fish growth

Growth curves are typically 2-3 parameter non-linear models

Most common is the von Bertalanffy:

$$\ell = \ell_\infty \left(1 - e^{-ka}\right)$$

where

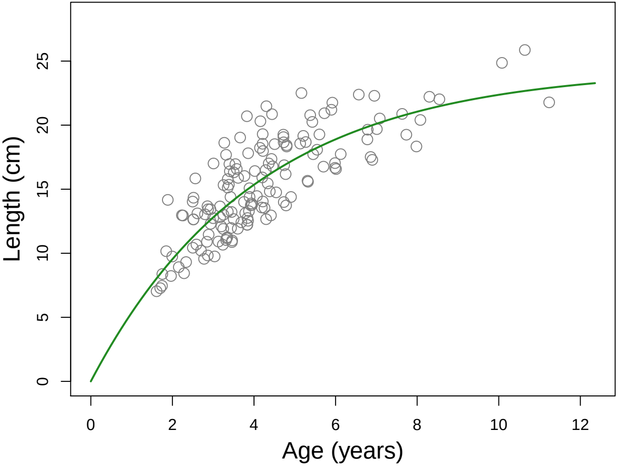
- ℓ: length
- a: age
- ℓ_∞: asymptotic length parameter
- k: growth rate parameter

Probabilistic (lognormal)

$$\ell_i = \ell_\infty \left(1 - e^{-ka_i}\right) e^{\varepsilon_i} \qquad \varepsilon_i \sim \mathrm{N}(0, \sigma_\ell^2)$$

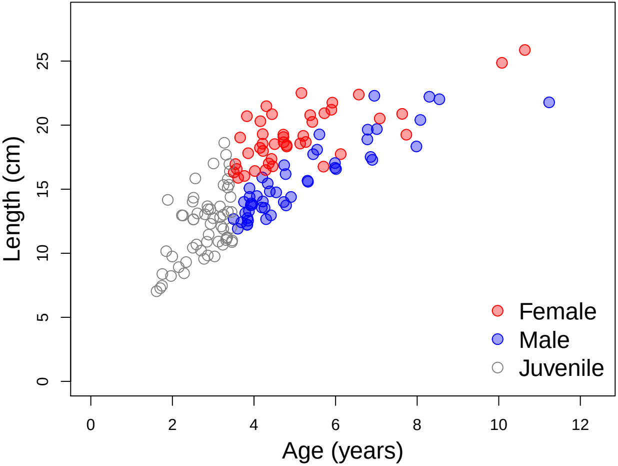
Fish growth

Fit a growth curve



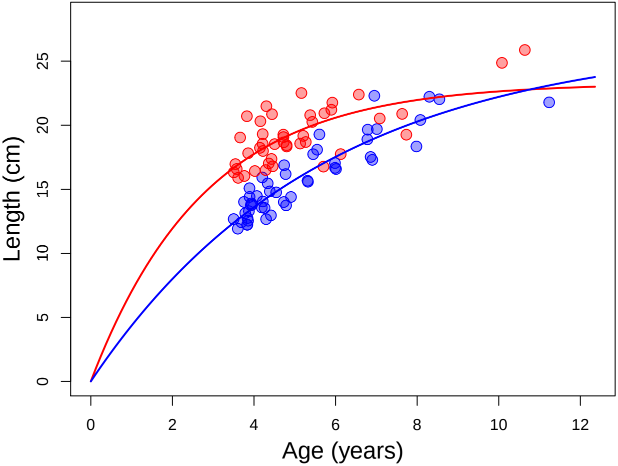
Fish growth

Sex-specific



Sex-specific growth

Standard practice is to discard juvenile data



Sex-specific growth

Standard practice may not make most of the data:

- Focuses on known sexes only
- Uses a reduced sub-region of the age-length space
- May be uninformative on growth rate
- Likely very uninformative when third parameter introduced
  (non-zero y-axis intercept)

Sex-specific growth

Suggested alternative:

- Keep all data when fitting sex-specific growth curves
- Treat the sex of the juveniles as a classification problem
- Simultaneously estimate the juvenile sexes and growth curves

How?

Fish growth: mixture model

Outline:

$$f(\ell|a,\boldsymbol{\theta}) = \pi_F f_F(\ell|a,\boldsymbol{\theta_F}) + \pi_M f_M(\ell|a,\boldsymbol{\theta_M})$$

where

$$\pi_F = \Pr(S = F),$$

where S is the sex

$$f_F(\ell|a,\boldsymbol{\theta_F}) = \frac{1}{\ell\sigma_F\sqrt{2\pi}} \exp\left(-\frac{(\ln(\ell) - \ln(v(a,\boldsymbol{\theta_F})))^2}{2\sigma_F^2}\right)$$

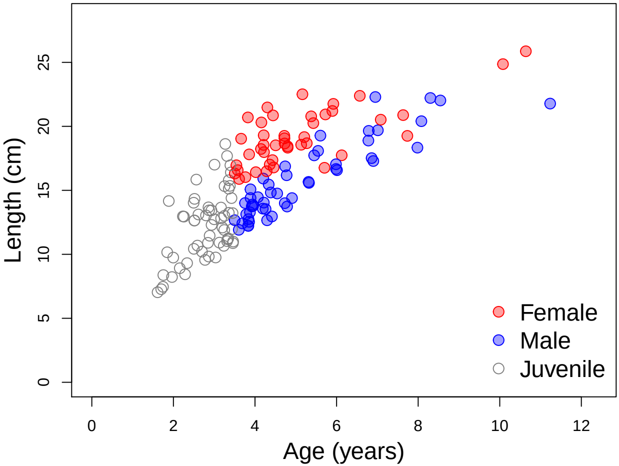
Lognormal where v is the von Bertalanffy function

$$Z_i = \begin{cases} 1, & \text{if observation } i \text{ is female,} \\ 0, & \text{if observation } i \text{ is male.} \end{cases}$$

Note: Z is **partially classified** — we know the sex of **some** of the individuals

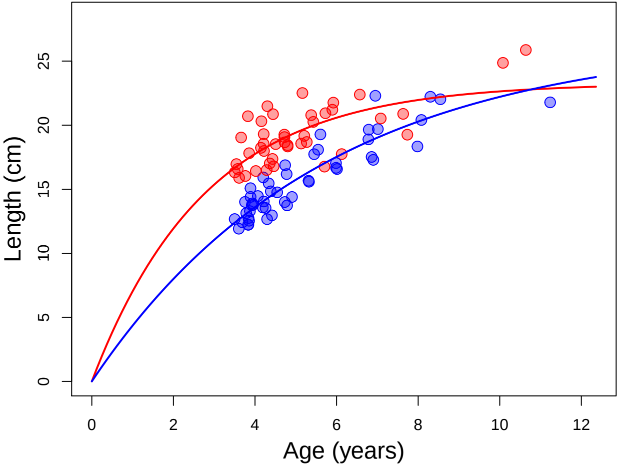
Example 1: separation

Data



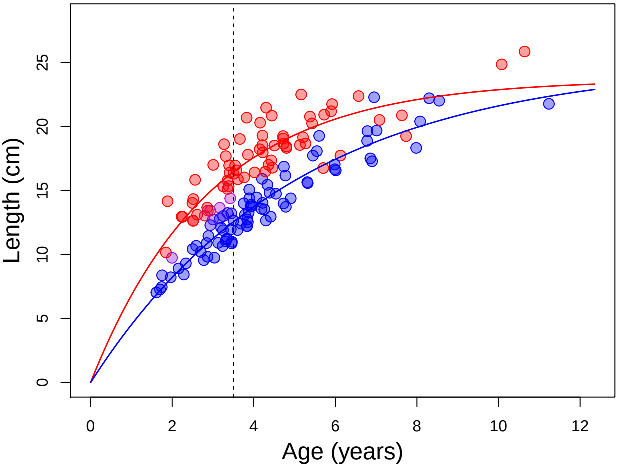
Example 1: separation

Standard practice



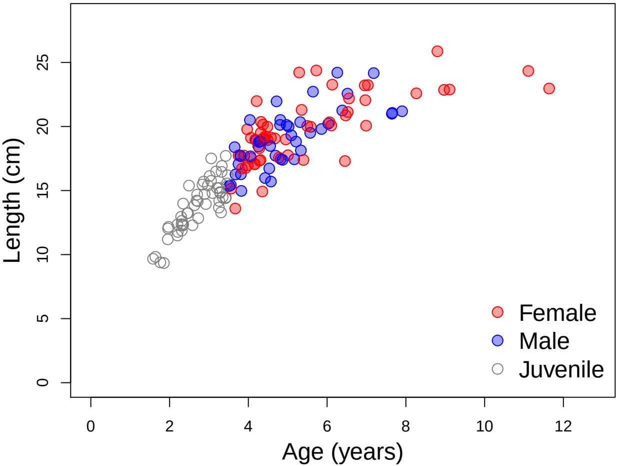
Example 1: separation

Finite mixture model fit



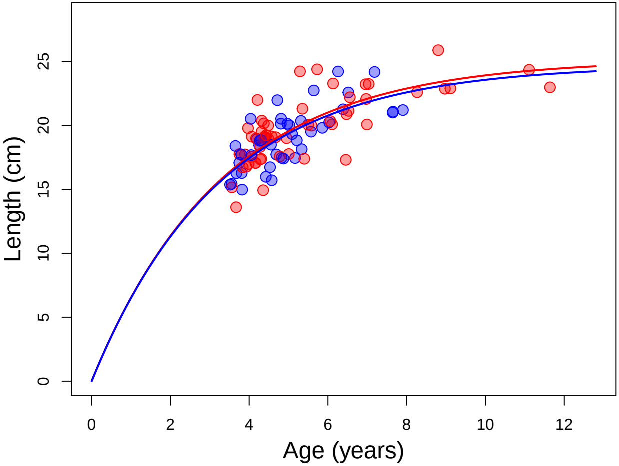
Example 2: overlapping

Data



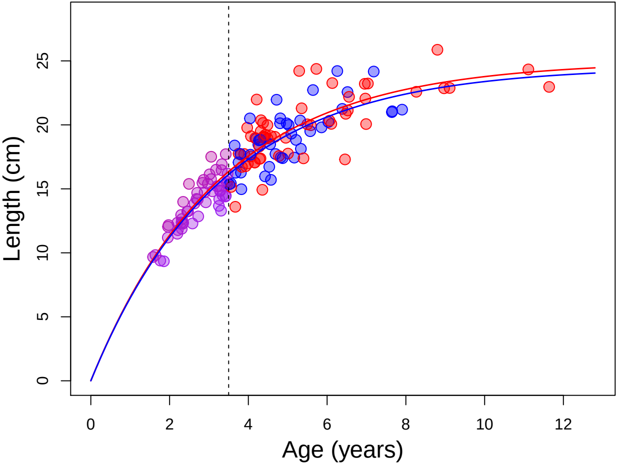
Example 2: overlapping

Status quo

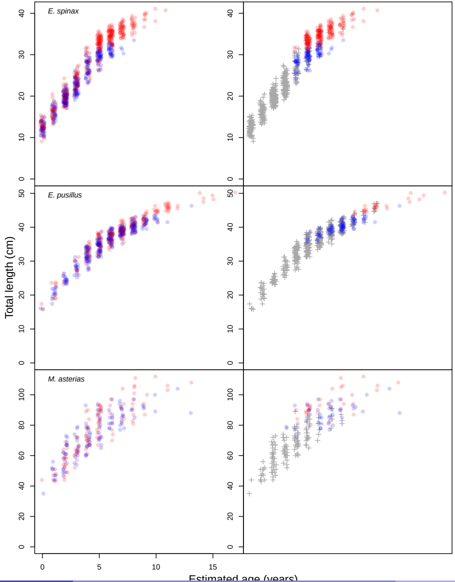


Example 2: overlapping

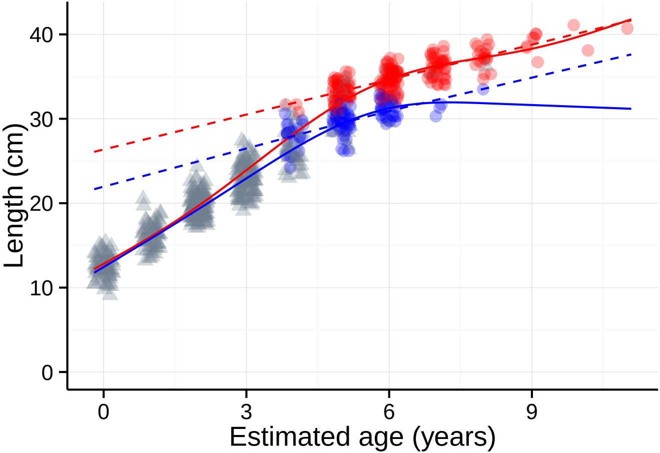
Finite mixture model fit



Real Data

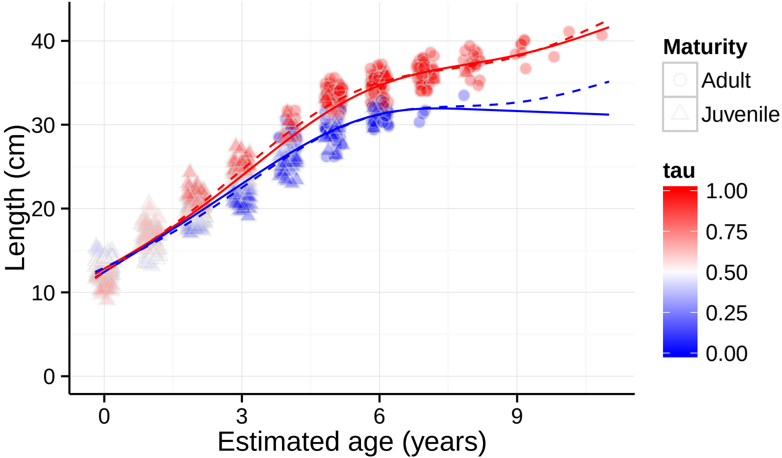

E. spinax
E. pusillus
M. asterias
Total length (cm)
Estimated age (years)
0
5
10
15
0
10
20
30
40
50
0
20
40
60
80
100

Traditional Modelling of Growth



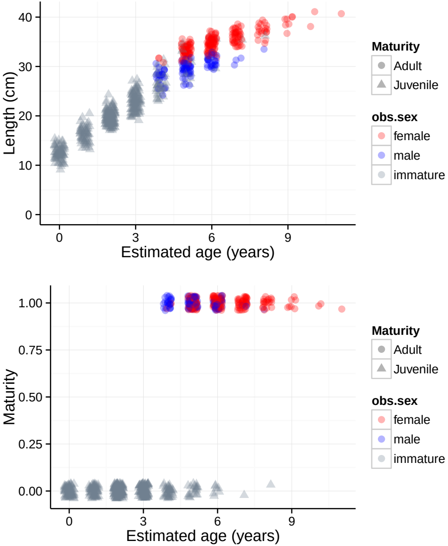
Dashed: only known sex data; **Solid**: full knowledge fit

Mixture Modelling of Growth



Dashed: EM mixture model fit; **Solid**: full knowledge fit

Modelling of Maturity and Growth



Modelling of Maturity and Growth

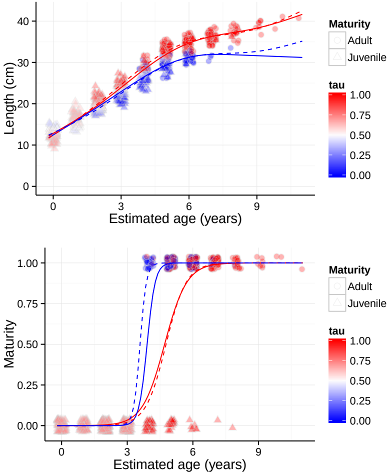
Gender specific models for

- **Growth** — length as a nonlinear model of age
- **Maturity** — logit model for maturity (known gender) depending on age

Fitting strategies

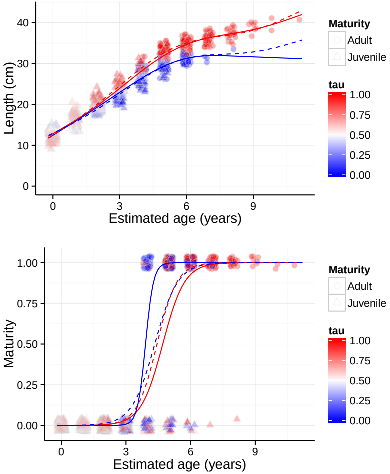
- Separate fits for **each** model
- **Joint** fit — missing gender estimation common to both models

Modelling of Maturity and Growth — Separate Fits



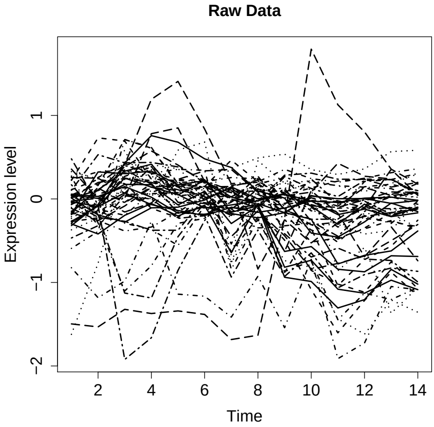
Dashed: EM fit; **Solid**: full knowledge fit

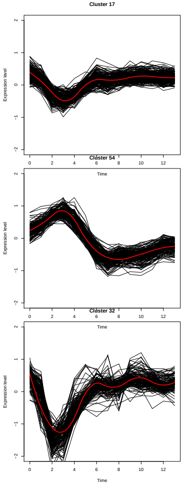
Modelling of Maturity and Growth — Joint Fit



Dashed: EM fit; **Solid**: full knowledge fit

Yeast data — time course microarray data





Previous analyses

- Have traditionally been clustered using multivariate clustering methods, e.g. k-means clustering, hierarchical clustering, finite mixture models, etc.
- Problems with gene expression data?
  - High dimensionality;
  - Missing values;
  - Large amounts of measurement error;
  - Correlation between measurements made over time on same gene.
- Multivariate techniques have difficulties handling these issues.

Smoothing

- Assume that there exists some underlying function $g(t)$ which generates the observed data.

- Observed data may contain a lot of measurement error/noise.

$$y_j = \underbrace{g(t_j)}_{\text{Signal}} + \underbrace{\varepsilon_j}_{\text{Noise}}$$

- Need to estimate smooth functions from noisy data.

- Use basis function expansions:

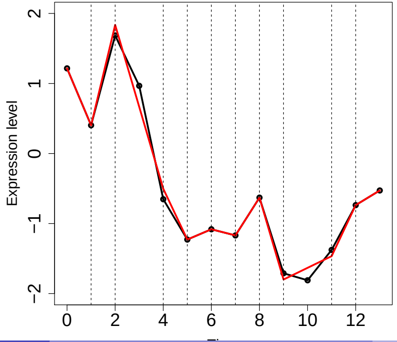
$$g(t) = \sum_{k=1}^{K} \beta_k \phi_k(t)$$

Basis functions

- Use pth degree truncated power basis (typically $p = 1$ or 2):

$$g(t_j) = \beta_0 + \beta_1 t_j + \ldots + \beta_p t_j^p + \sum_{\ell=1}^{L} \beta_{1\ell}(t_j - \kappa_\ell)_+^p,$$

$\kappa_\ell = \ell$th knot and $(t_j - \kappa_\ell)_+ = \max(0, t_j - \kappa_\ell)$.



P-spline smoothing as a mixed model

- Represent P-spline smoothing as a linear mixed effects model.
- Mixed effects model has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}.$$

- For simplicity assume $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$.
- For smoothing must also assume $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$.
- Estimates of $\boldsymbol{\beta}$, σ_ε^2, σ_u^2 and $\mathbf{u}$ determined using (RE)ML and BLUP.

Smoothing using mixed models: example

- Can smooth using a linear mixed effects model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

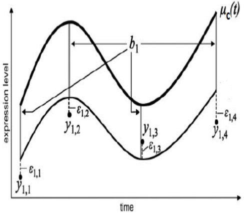
$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \mathbf{u} = \begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1L} \end{pmatrix}$$

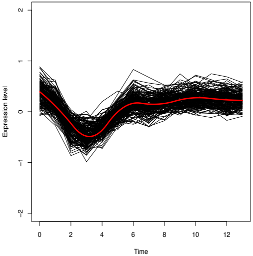
$$\mathbf{X} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} (t_1 - \kappa_1)_+ & \cdots & (t_1 - \kappa_L)_+ \\ (t_2 - \kappa_1)_+ & \cdots & (t_2 - \kappa_L)_+ \\ \vdots & \ddots & \vdots \\ (t_n - \kappa_1)_+ & \cdots & (t_n - \kappa_L)_+ \end{pmatrix}$$

- Assume $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$ and $\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$.

Why bother?

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$$





- Very flexible.
- Computationally efficient.
- Can be fitted using readily available software, e.g. SAS, R, S-Plus, etc.

Gene expression clusters

- Want to cluster genes into groups exhibiting the same/similar expression profiles.

- Write the expression level for gene i in cluster c at time j as

$$y_{ij} = \mu_g(t_{ij}) + b_i + \varepsilon_{ij}, \quad j = 1, \ldots, n_i,$$

where $b_i \sim N(0, \sigma^2_{bc})$ represent gene-specific shifts from the mean.

- Stack all data from genes in cluster c to get

$$\mathbf{Y}_c = \underbrace{\mathbf{X}_{c,s}\boldsymbol{\beta}_{c,s} + \mathbf{Z}_{c,s}\mathbf{u}_{c,s}}_{\mu_c(t)} + \mathbf{Z}_{c,b}\mathbf{b}_c + \boldsymbol{\varepsilon}_c,$$

$\mathbf{u}_{c,s} \sim N(\mathbf{0}, \sigma^2_{uc}\mathbf{I})$, $\mathbf{b}_c \sim N(\mathbf{0}, \sigma^2_{bc}\mathbf{I})$, $\boldsymbol{\varepsilon}_c \sim N(\mathbf{0}, \sigma^2_{\varepsilon c}\mathbf{I})$.

Gene expression clusters

- In practice, do not know cluster membership.
- Assume $\mathbf{y}_i$ comes from a mixture of C clusters:

$$\mathbf{y}_i \sim \pi_1 N(\mu_1(\mathbf{t}_i), \mathbf{V}_{i1}) + \pi_2 N(\mu_2(\mathbf{t}_i), \mathbf{V}_{i2}) + \ldots + \pi_C N(\mu_C(\mathbf{t}_i), \mathbf{V}_{iC})$$

where

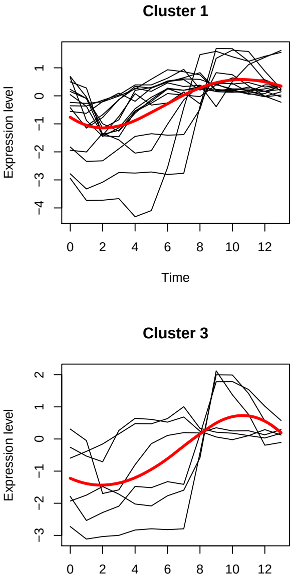
$$\mu_c(\mathbf{t}_i) = \mathbf{X}_{i,s}\boldsymbol{\beta}_{c,s} + \mathbf{Z}_{i,s}\mathbf{u}_{c,s}$$

and $\mathbf{V}_{ic} = \sigma^2_{bc} E_{n_i \times n_i} + \sigma^2_{\varepsilon c}\mathbf{I}_{n_i \times n_i}$

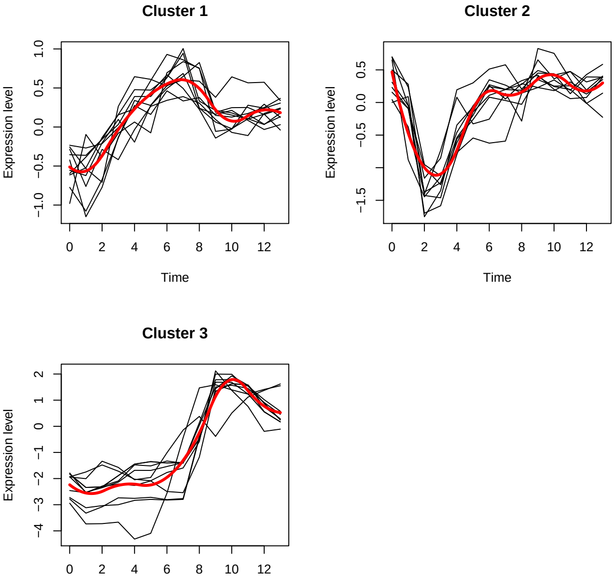
- $\pi_1, \pi_2, \ldots, \pi_C$ are mixing proportions such that $\sum_{c=1}^{C} \pi_c = 1$.
- Estimate $\pi_1, \ldots, \pi_C$, $(\boldsymbol{\mu}_1, \boldsymbol{V}_1), \ldots, (\boldsymbol{\mu}_C, \boldsymbol{V}_G)$.
- Obtain (posterior) probability that gene i is from cluster c.

Use EM algorithm.

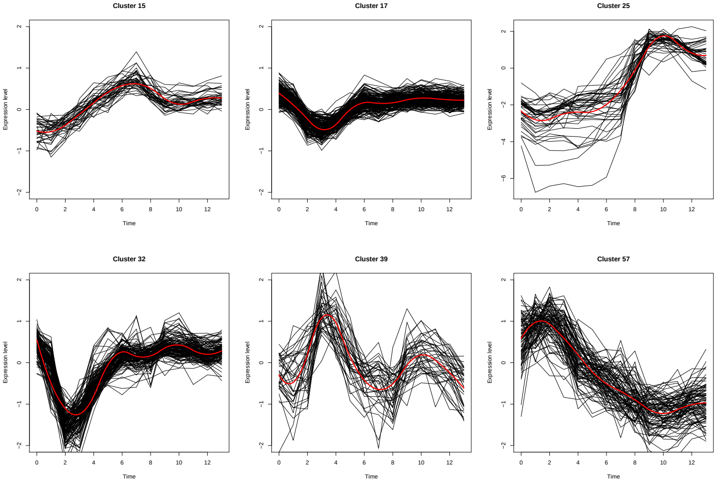
EM algorithm



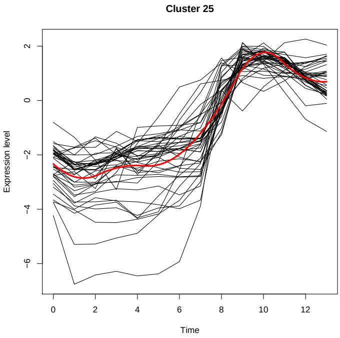
EM algorithm



Results: BIC suggests 58 clusters; 6 example groups



Results



- GO terms:- Sterol transport and stress-response.
- Sterols important in many cellular processes (usually synthesised in the ER membrane).
- Anaerobic conditions: must be imported into the cell ⇒ sterol import genes activated.
- Other genes from Seripauperin family only activated under anaerobic conditions.

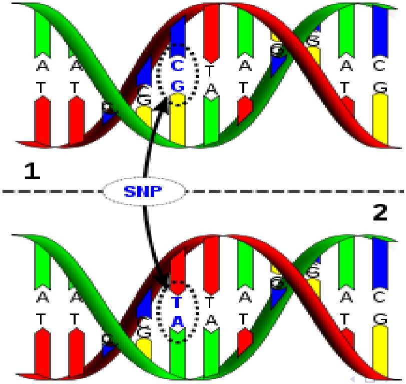
Sugar Cane

Motivation

- An allele is a particular form of a gene, e.g. the gene for eye colour has a number of alleles.
- Most organisms are diploid (2 sets of chromosomes).
- Sugarcane is polypoid (8 to 14 chromosomes) with individual alleles in varying numbers.
- Want to identify the many different alleles and associated genotypes/phenotypes.
- Can do this through the analysis of single nucleotide polymorphisms (SNPs).

SNPs

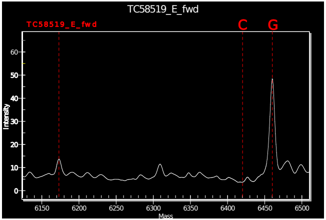
- SNPs occur during cell division, when cell divides in two by first copying its DNA.
- SNPs are mistakes that occur during the copying process i.e. changes that occur at a single base pair in DNA sequence.



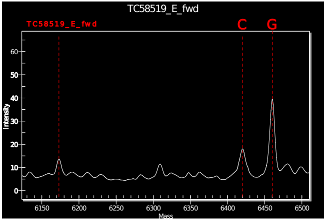
SNPs

- Frequency of a SNP base (A, T, C, G) at a locus determined by
  - the number of chromosomes carrying the gene;
  - the number of different alleles (or haplotypes);
  - the frequency of each allele possessing each SNP base.
- In sugarcane, the proportional frequencies of each SNP base varies depending on the number of alleles containing the SNP locus.
- Gives an indication of the number of allele haplotypes present for a gene.

Data — Spectra

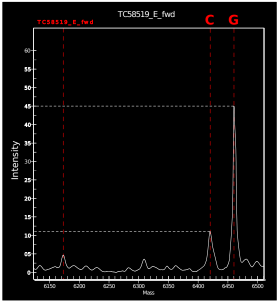


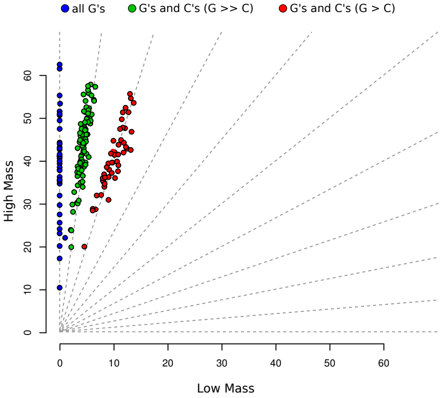
- Homozygous individual with allele G (nucleotide)



- Heterozygous individual with some copies of allele C and some copies of allele G (G > C)

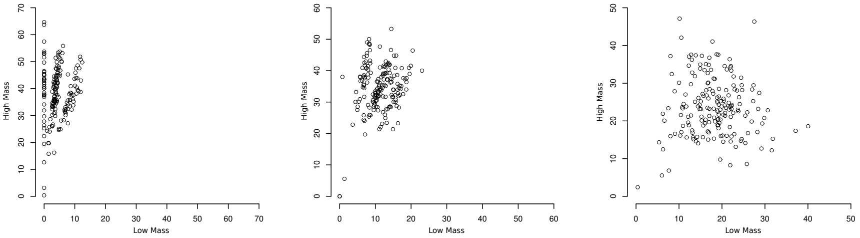
Data — Idealised



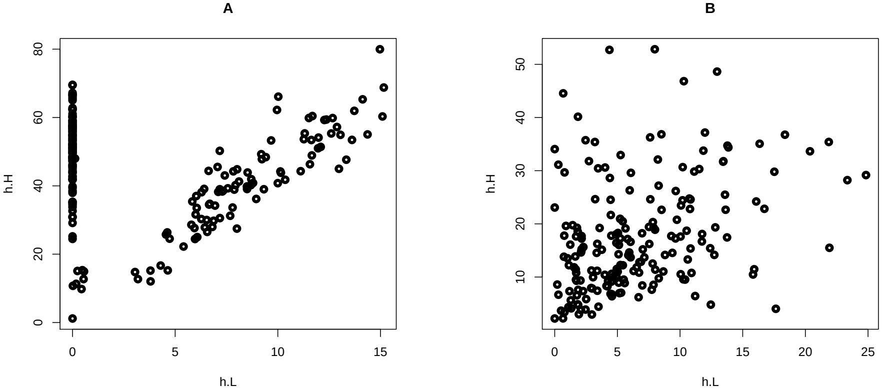


Real Data

- How many clusters?
- What are the angles (dosages) and proportions?
- How to allocate the individuals?



Data — Illustrative



Aims

Want to develop a technique that can:

- determine the number of clusters present;
- determine the angles between the lines that represent each cluster to identify different genotypes;
- provide a probabilistic clustering to identify points that have high probability of belonging to a particular cluster (i.e. points that have a particular genotype) and those that are regarded as an unclear genotype.

Finite Mixture Models

- Have a p-length data vector $\mathbf{y}_i = (\text{h.H, h.L})$ for each individual.
- Finite mixture models assume that the data come from a mixture of G clusters such that

$$f(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{y}_i; \boldsymbol{\theta}_g), \tag{1}$$

  where $f(\mathbf{y}_i; \boldsymbol{\theta})$ is the density of the data, $f_g(\mathbf{y}_i; \boldsymbol{\theta}_g)$ is the gth component density and π_g are mixing proportions such that

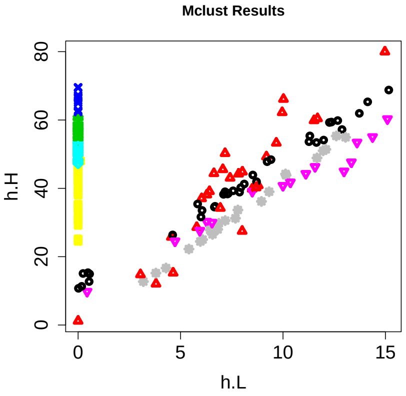
$$\sum_g \pi_g = 1.$$

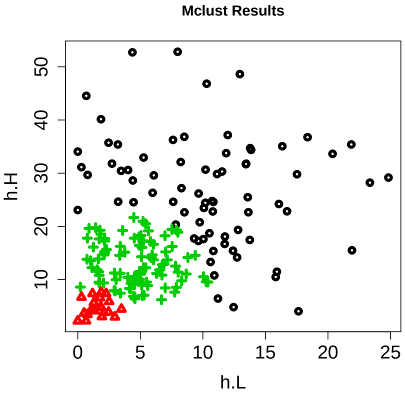
- Usually assume

$$f_g(\mathbf{y}_i; \boldsymbol{\theta}_g) = N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \tag{2}$$

- Need to estimate $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G, \pi_1, \ldots, \pi_{G-1})$ using EM algorithm.

Mclust


Mclust Results
h.H
h.L
0
20
40
60
80
0
5
10
15


Mclust Results
h.H
h.L
10
20
30
40
50
0
5
10
15
20
25

Linear Regression Lines

- Assume one of h.H/h.L is the response variable y_i and the other is the explanatory variable x_i.
- Fit a linear regression line through the origin

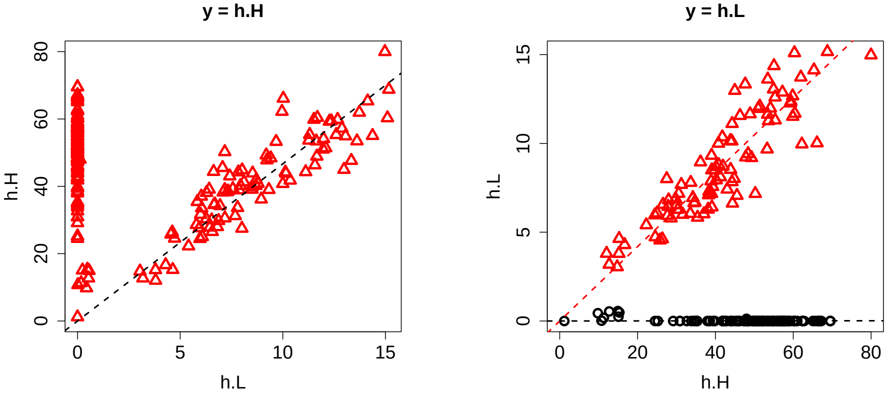
$$y_i = \beta_{1g} x_i + \varepsilon_i$$

  to the data in each component.
- Component densities now written as

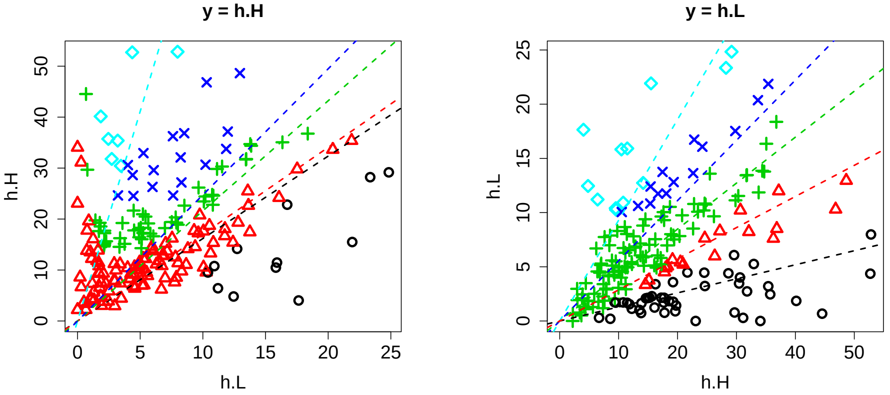
$$f_g(y_i | \beta_{1g} x_i, \sigma_g^2) = N(\beta_{1g} x_i, \sigma_g^2),$$

  where β_{1g} is the slope in the gth component and σ_g^2 is the variance.
- Need to determine estimates of β_{1g}, σ_g^2 (can be the same/different for each cluster) and $(\pi_1, \ldots, \pi_{G-1})$.

Results - Contig89b17



Results - Contig2312b2



Orthogonal Regression Lines

- Special case of Total Least Squares.
- Orthogonal regression line has the form

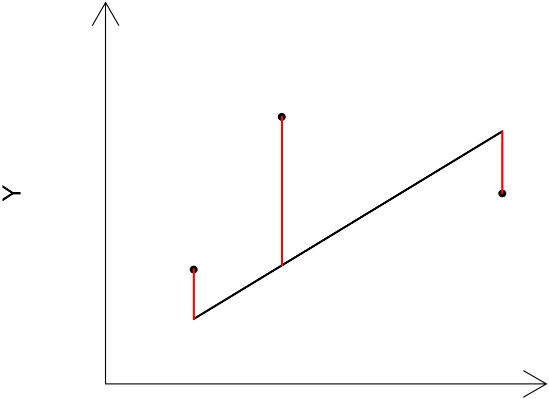
$$y_i = x_i \beta_{1g}$$

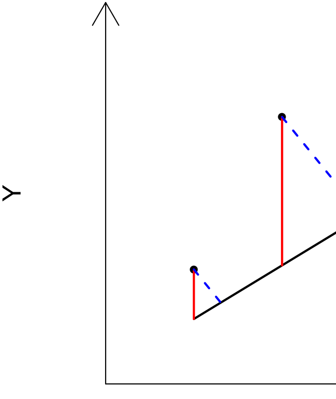
- Assumes both x and y are measured with error:

$$\begin{aligned} x_i &= x_i^* + \epsilon_i, \quad \mathrm{Var}(\epsilon) = \sigma_x^2 \\ y_i &= y_i^* + \tau_i, \quad \mathrm{Var}(\tau) = \sigma_y^2 \end{aligned}$$

- Orthogonal regression $\Rightarrow \sigma_x^2/\sigma_y^2 = \eta$ and independent.
- Suitable when both variables are linearly related and subject to error.
- Can fit a regression line to group parallel to the y-axis.

Orthogonal Regression Lines





Orthogonal Regression Lines

- Calculate $\hat{\beta}_{1g}$ using SVD.
- Need to find fitted values $(\hat{x}_i, \hat{y}_i)$.
- Find equation of line orthogonal to line with slope $\hat{\beta}_{1g}$ that goes through original data point (x_i, y_i).
- Point at which this line and orthogonal regression line intersect gives fitted values:

$$\hat{x}_i = \frac{y_i \hat{\beta}_{1g} + x_i}{1 + \hat{\beta}_{1g}^2}, \quad \hat{y}_i = \hat{x}_i \hat{\beta}_{1g}$$

- Assume $\sigma_x^2 = \sigma_y^2 \Rightarrow$ overall estimate of σ_g^2 given by

$$\hat{\sigma}_g^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{2(n-1)}$$

Orthogonal Regression Lines

- For clustering

$$f(x_i, y_i|\boldsymbol{\theta}_g) = \sum_{g=1}^{G} \pi_g f_g(x_i, y_i|\boldsymbol{\theta}_g)$$

- Component densities have form

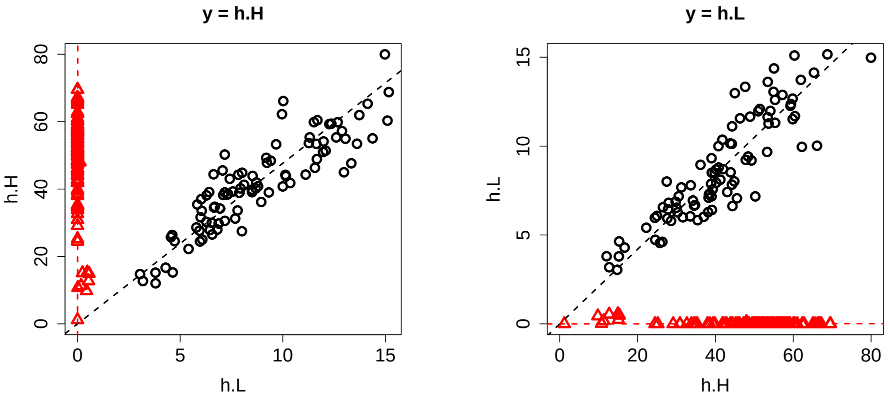
$$f_g(x_i, y_i|\boldsymbol{\theta}_g) = N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

where

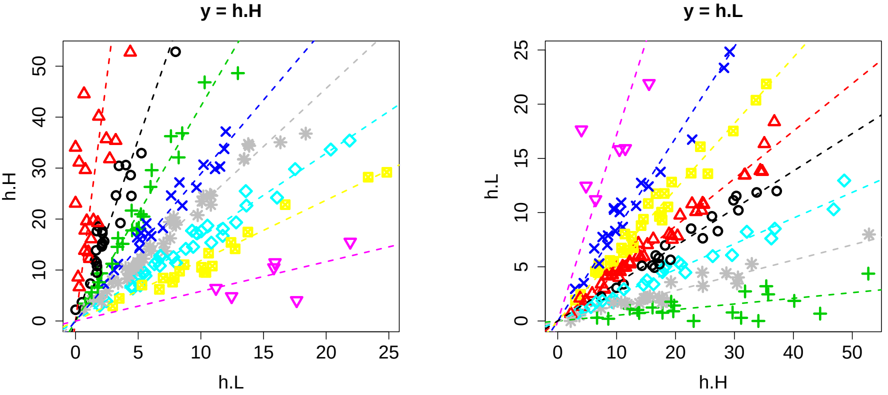
$$\boldsymbol{\mu}_g = \begin{pmatrix} \hat{x}_i \\ \hat{y}_i \end{pmatrix}, \quad \boldsymbol{\Sigma}_g = \begin{pmatrix} \hat{\sigma}_g^2 & 0 \\ 0 & \hat{\sigma}_g^2 \end{pmatrix}$$

- For each component find $\hat{x}_i$, $\hat{y}_i$ and $\hat{\sigma}_g^2$ as outlined in previous slide.

Results - Contig89b17



Results - Contig2312b2



Model choice & Use

- How many lines/groups?

  Use BIC or similar approaches, as in standard model-based clustering.

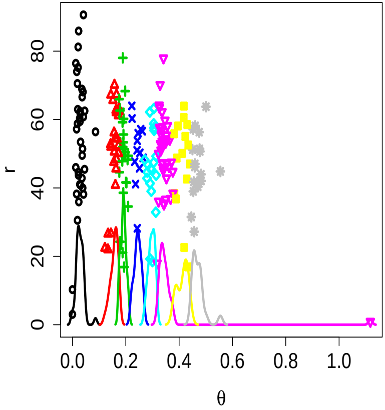
- Constrained models — lines at multiple of a *common angle*

- Which lines are present?

  Estimated mixture proportions — give information on ploidy level and genotype distribution

Polar Coordinates: Results - Contig2312b2

$$r = \sqrt{x^2 + y^2} \qquad \theta = \arctan(y/x)$$



Acknowledgements

- Norma Coffey
- Jochen Einbeck
- Marie-José Martinez
- Cóilín Minto
- Augusto Franco Garcia

References

- Aitkin, M.A., Francis, B.F., Hinde, J.P. and Darnell, R. (2009) *Statistical Modelling in R*. Oxford University Press, 576pp.
- Coffey, Norma and Hinde, John (2011) Analyzing Time-Course Microarray Data Using Functional Data Analysis — A Review, *Statistical Applications in Genetics and Molecular Biology*, **10**: 1, Article 23.
- Coffey, Norma, Hinde, John and Holian, Emma (2014) Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics and Data Analysis*, **71**, 14-29.
- Coffey, Norma, Hinde, John and Garcia, Augusto Franco. (2014) Finite mixture model clustering of SNP data. In *Statistical Modelling:Papers in Biostatistics and Bioinformatics*, eds MacKenzie, G. and Peng, D., Springer, 139-157.

The End !!!

Thank you for your attention