Time Series Data in Functional Genomics

Regression, Functional Data Models and Correlated Data

David A. Stephens

[1]Department of Mathematics and Statistics, McGill University

d.stephens@math.mcgill.ca

http://www.math.mcgill.ca/∼dstephens/MCB/

Outline

- Time course data
- Linear Regression Models
- Estimation and Inference
- Flexible Models: Splines
- Model-based Clustering
- Models for correlated data

Time Course Microarray Data

Microarrays are used to record the relative gene expression of many thousands of genes simultaneously.

This is useful when we wish to compare the functional activity across different genetic subgroups

- wild-type vs knockout
- heterozygote/double homozygote
- different developmental stages

Microarrays can detect differential expression across these subgroups.

Time Course Microarray Data

A more powerful experimental set-up is where we detect **changes in expression through time**.

A typical design would involve carrying out repeated microarray experiments on similar experimental units a different times over a number of hours or days that would allow us to measure how the differential expression changes in time.

This allows us to better understand **patterns of regulation**.

Motivation

We motivate the subsequent statistical models by looking at several real examples

- the principal model organism is the *Anopheles Gambiae* mosquito
- we study patterns in regulation of immune defence mounted in response to bacterial and chemical challenges
- this informs the study of malaria, and the immune defence of the mosquito to infestation by the protozoan parasite *Plasmodium Falciparum*

The Global Impact of Malaria.

- Malaria is caused by the parasite *Plasmodium falciparum* and is primarily spread by *Anopheles Gambiae* mosquito vectors.

- Alongside HIV and tuberculosis, it represents one of the world's most damaging infectious diseases.
- Malaria affects two to three hundred million people each year, one million of whom are children living in sub-Saharan Africa.
- Globally, two thousand million people (40% of the world's population) are at risk.
- Malaria research is ongoing: a key element is to understand genetic regulation in the mosquito and in the parasite.

Mosquito and Parasite Genomics

- 2002: genomes of *P. Falciparum* and *A. Gambiae* mapped
- it has been demonstrated that the mosquito employs its own immune system against the parasite.
- the components operating mosquito immune system and their potential relevance to antimalarial responses are being systematically dissected.
- special emphasis has been placed on the study of anti-malarial responses involved in limiting the extent of infection.

Microarray studies

Microarray experiments carried out at Imperial College have studied various aspects of genetic regulation.

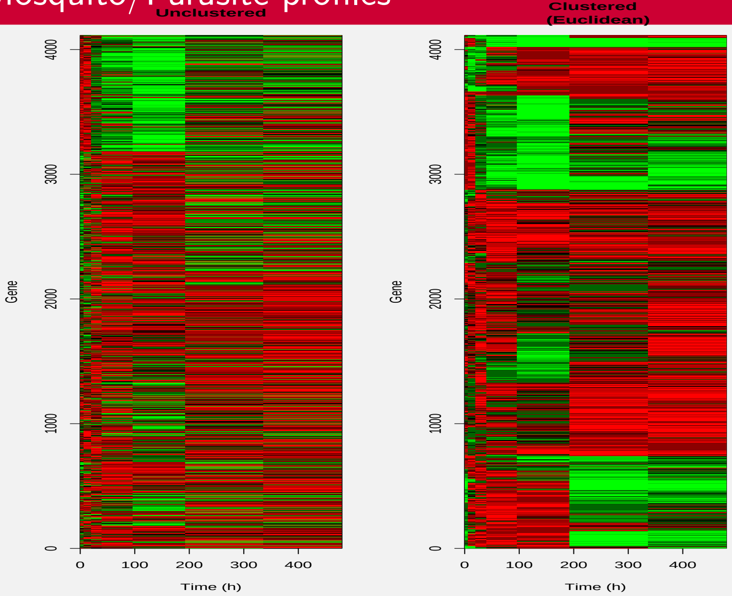
- The immune defense system of the mosquito to infestation by the parasite has come under study.
- An immune defense response is mounted whenever the host mosquito is infected by the parasite; genes in the mosquito genome known to be involved in immune defense have been identified.
- Key task is to find genes with similar regulation patterns, as they too may be involved with immune defense activity.

- clear correlation of immune responses with the passage of the parasite through the vector.
- the mosquito has become the organism of choice for directly studying antiparasitic innate immune responses.

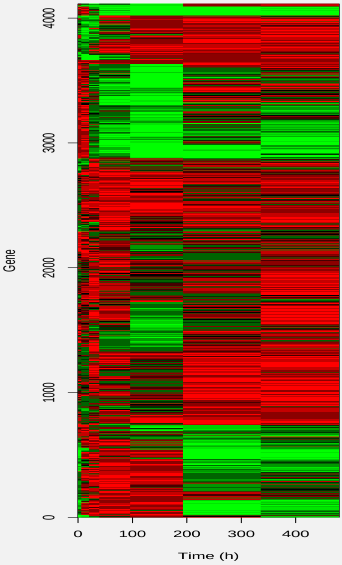
Recent genomic investigations in malaria include studies of

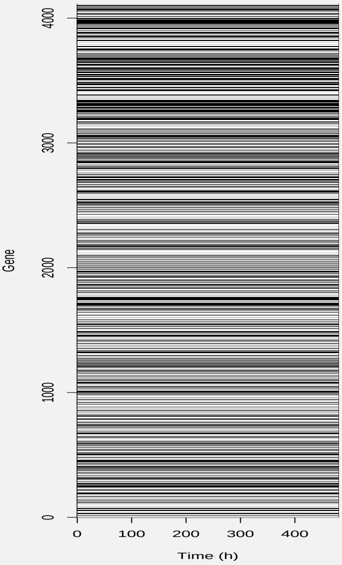
- mosquito with parasite infestation
- mosquito only under artificial experimental challenge

Mosquito/Parasite profiles



Profiles cluster within species ?





Objectives

- to understand regulatory mechanisms within each organism
- to produce plausible models for the patterns of regulation
- to extract subsets of genes that have similar patterns
- to classify genes to functional classes of interest (i.e. immune defense clusters)

Objectives

- to produce models and algorithms that can and will be used by researchers
- biologists often reluctant to use advanced statistical methods
- computational feasibility is an important factor

Time Course Data

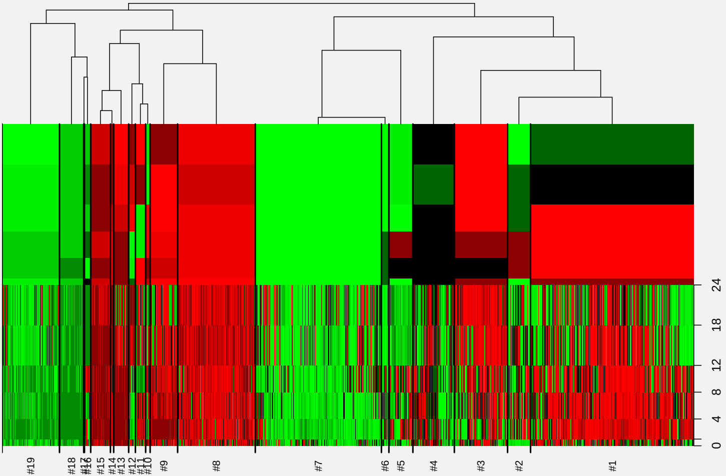
Statistical analysis of gene expression profiles obtained by microarray assays of mosquitos/cell-lines compromised by bacterial and chemical agents (challenges):

- data comprise relative expression of 2771 genes/sequence tags,
- probes selected from a specially constructed cDNA library.
- approximately 900 have associated/putative function.
- relative expression recorded at $T = 6$ time points, at 1, 4, 8, 12, 18 and 24 hours after the challenge.

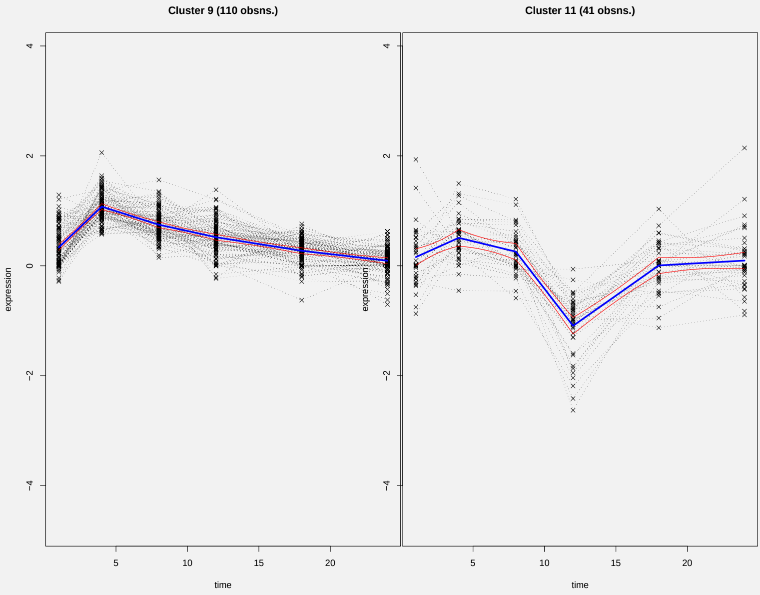
We focus on a single bacterial challenge, *Salmonella typhi*.

Salmonella typhi challenge data.

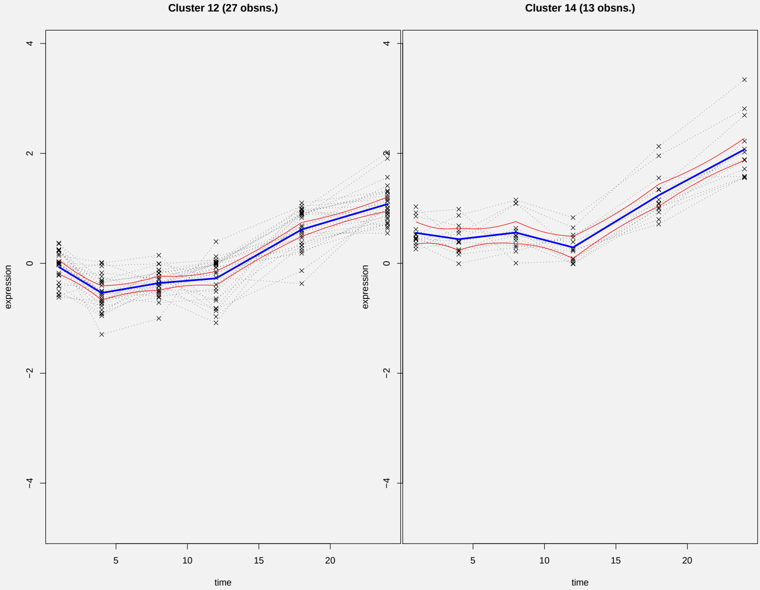
Clustering Results



Clusters 9 and 11



Clusters 12 and 14



Co-clustering for Multiple Challenges

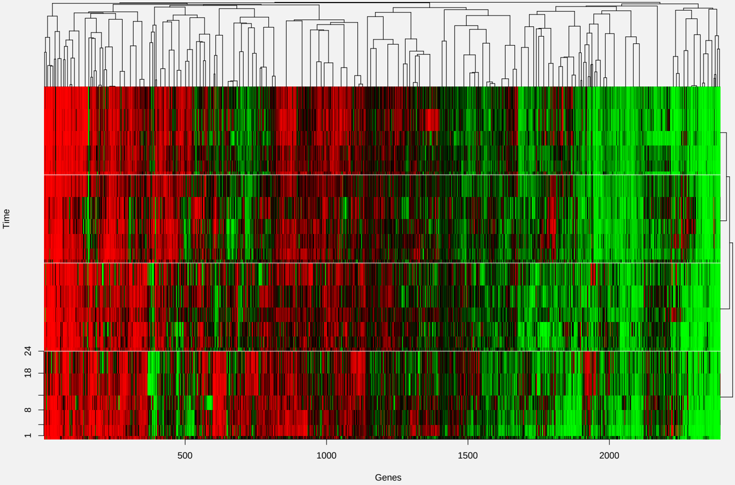
The *S. Typhi* challenge data were obtained as one of a series of experiments using ($\sim$ 15) different challenges on the same gene set. Here we look at only four:

- *S. Typhi*
- *Listeria*
- *M. Luteus*
- Zymosan (chemical)

We would expect similar patterns of regulation of immune defense genes under different challenges.

Unclustered

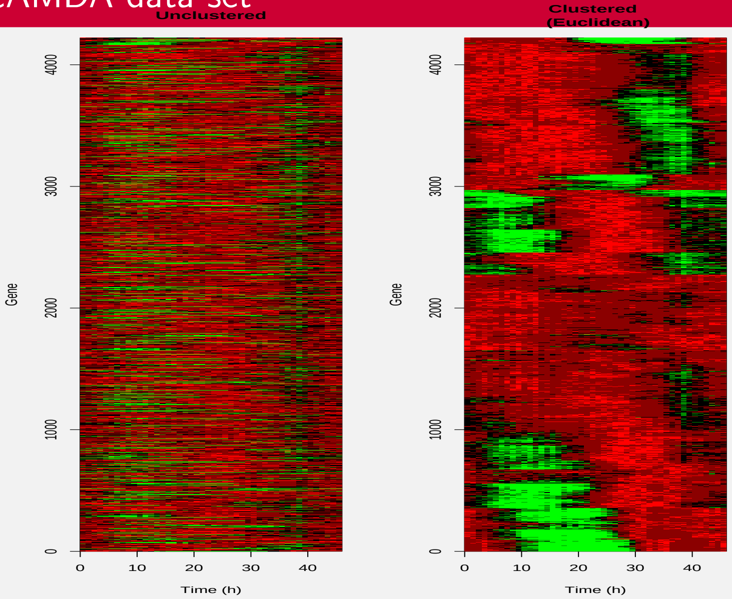
Co-clustered



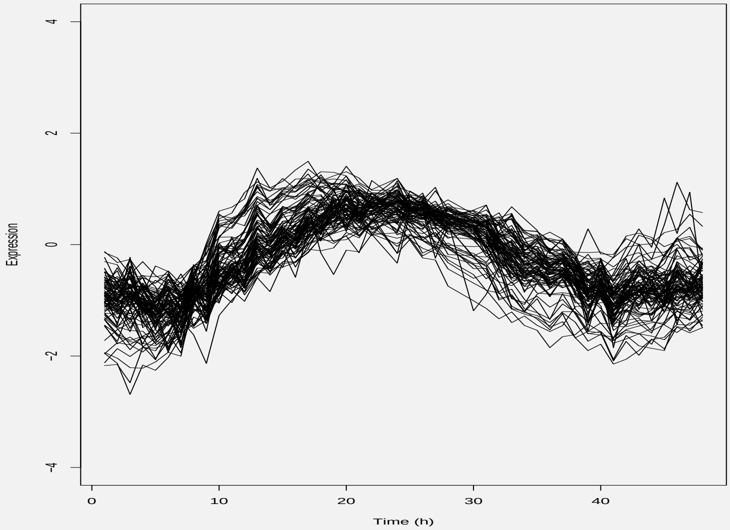
Gene Regulation in *P. Falciparum*

- Patterns of gene regulation in the parasite also under study
- CAMDA challenge data
- 4221 genes, 46 time points over 48 hours studied here
- expression relative to 48 hour transcriptome
- using hierarchical clustering, results in ~ 30 clusters
- not strictly a clustering problem ?

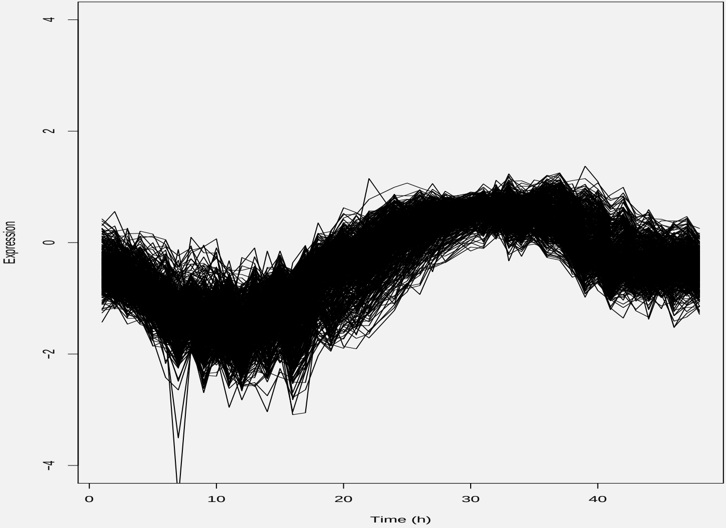
CAMDA data set



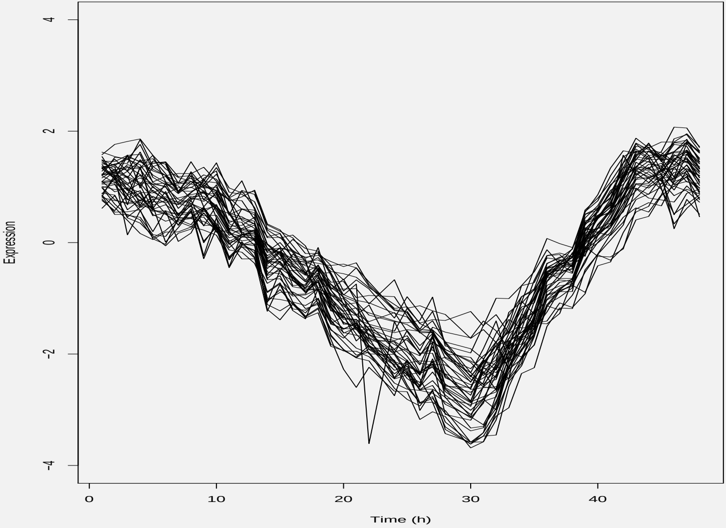
Cluster 4



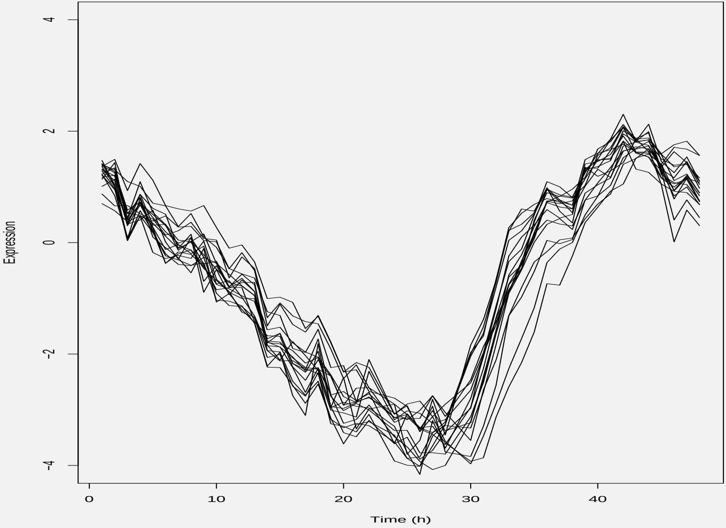
Cluster 6



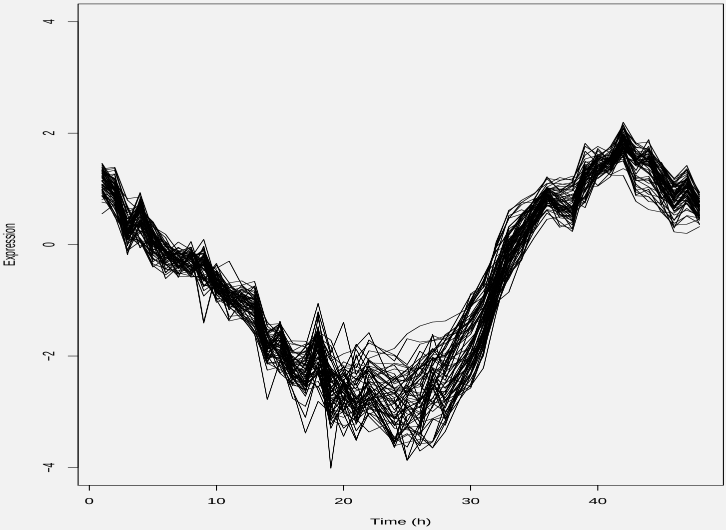
Cluster 1



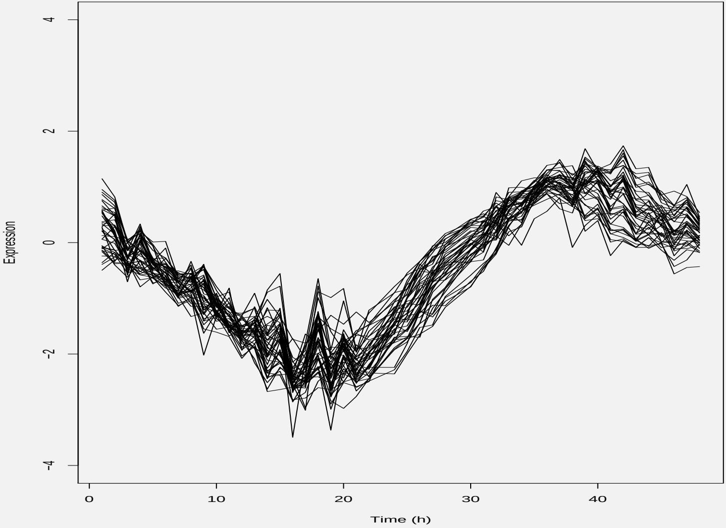
Cluster 2



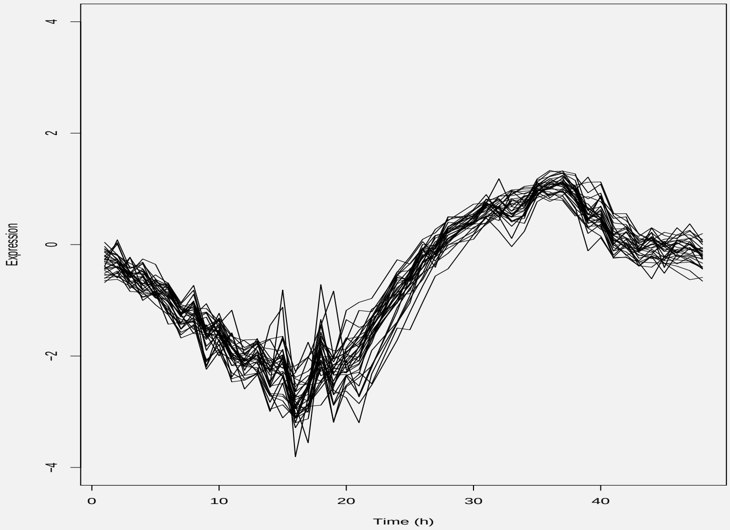
Cluster 7



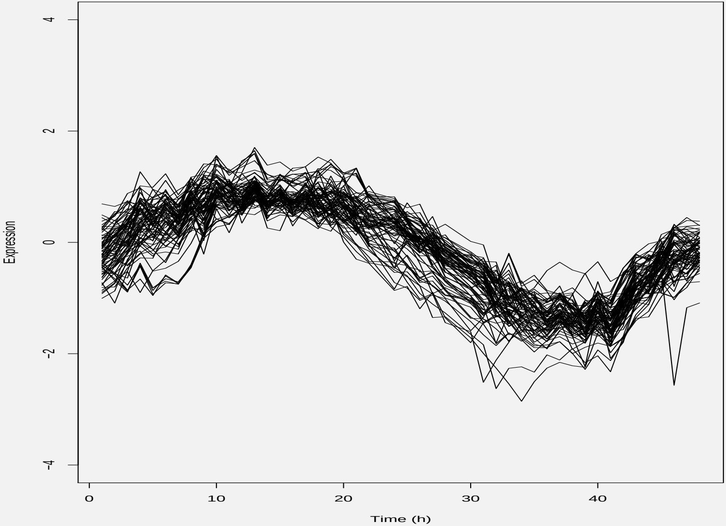
Cluster 25



Cluster 19



Cluster 16

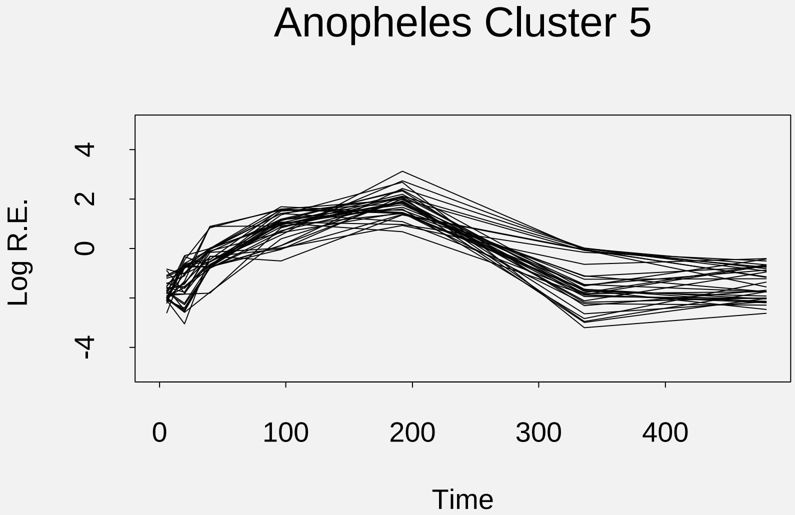


Mosquito and parasite interaction

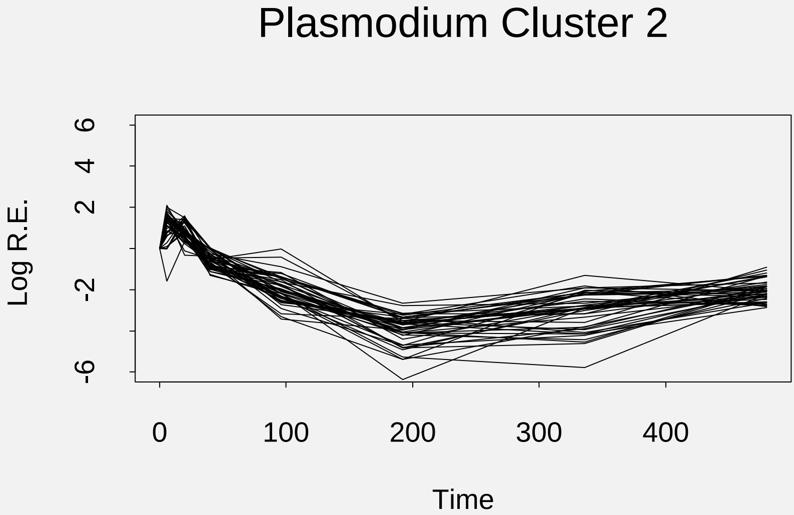
Mosquito and parasite regulation studied concurrently

- mosquitos studied longitudinally over 20 days after infected blood meal
- 4200 genes/ESTs
- 1400 from each the Anopheles and Plasmodium genomes
- 1400 unidentified ESTs
- seven time points

After clustering: Anopheles



After clustering: Plasmodium



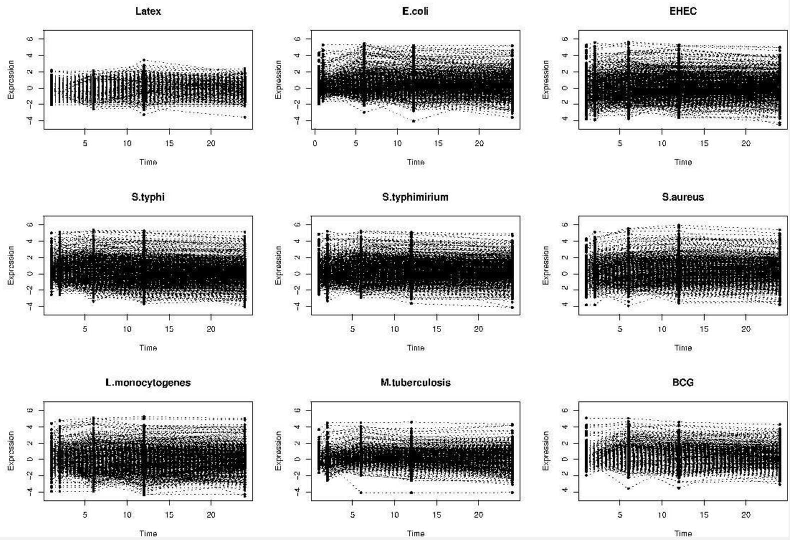
Example: immune defense in humans

Nau *et al* (PNAS, 2002) investigated human macrophage activation induced by different bacterial pathogens.

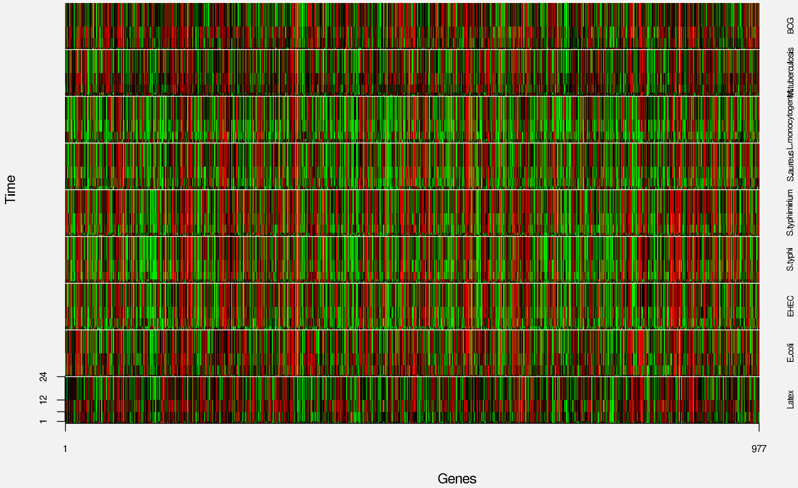
Learning about the gene response of innate immune cells to these pathogens may provide insight into host defenses and tactics used by pathogens to circumvent these defenses.

Sequences of microarray experiments were performed over a duration of 24 hours for 8 different bacterial pathogens plus a control.

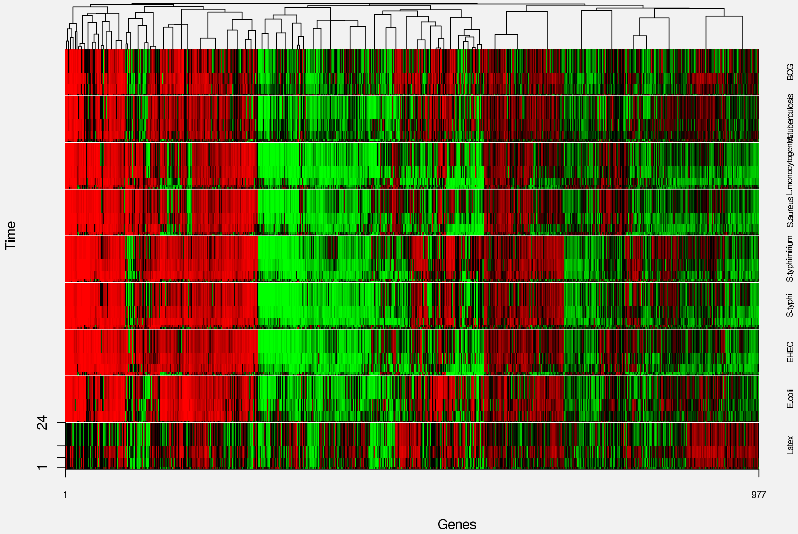
Nau *et al* data



Unclustered expression data



Clustered expression data



Regression Models

We wish to explain the variation of a time-varying *signal*. We will look at models of the form

$$y = f(x) + \epsilon$$

where

- x will represent time
- y will represent the measured relative expression of a gene or collection of genes
- $f(.)$ is a non-constant function
- ϵ is a random measurement error term.

Regression Models

The choice of f will crucially determine the types of patterns of regulation that can be captured.

- No differential expression: f is constant

$$f(x) = a$$

- Curvilinear Pattern: Quadratic form

$$f(x) = a + bx + cx^2$$

- Periodic Pattern: Trigonometric form

$$f(x) = a + b\cos(\lambda_1 \pi x) + c\sin(\lambda_1 \pi x)$$

We will see that each of these models can be regarded as special cases of a specific model; the

LINEAR REGRESSION MODEL

Regression Models

We will study the simplest form of a model in the class of Linear Regression Models where the relationship between time x and response y is a straight line.

Clearly for most real biological systems, the straight line model is an unrealistic simplification;

- the model implies that, beyond the time frame of the experiment the response increases (or decreases) over the whole range of x
- we expect periodic behaviour, or perhaps return to equilibrium after a stimulus, as time increases.

However, the properties of the model are best understood in the simplest case, and generalization to more realistic situations is then more straightforward.

Simple Linear Regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a straight-line).

Note: variables x and y will not be treated exchangeably - we will regard y as being a function of x.

Such models are **DETERMINISTIC**, that is, if we know x (and the values of the constants), we can compute y exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

Probabilistic Models

In a **probabilistic** model, we allow for the possibility that y is observed with random error, that is,

$$y = ax + b + ERROR$$

where *ERROR* is a random term that is present due to imperfect observation of the system due to (i) measurement error or (ii) missing information.

Note that we do not treat x and y exchangeably; x is a fixed observed variable that is measured *without error*, whereas y is an observed variable that is measured *with random error*.

We model the variation in y as a function of x. We observe pairs $(x_i, y_i), i = 1, \dots, n$.

A Basic Probabilistic Model

Terminology:

- y - *Dependent variable* or *independent variable*
- x - *Independent variable*, or *predictor*, or *covariate*

The model we study takes the form

$$y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a random error term, a random variable with mean zero and finite variance ($E[\epsilon] = 0$, $Var[\epsilon] = \sigma^2$); it represents the error present in the measurement of y.

- β_0 - *Intercept* parameter
- β_1 - *Slope* parameter

- $\beta_1 > 0$ - increasing y with increasing x
- $\beta_1 < 0$ - decreasing y with increasing x
- $\beta_1 = 0$ - no relationship between x and y

Note:

$$E[Y|x] = \beta_0 + \beta_1 x$$

where $E[Y|x]$ is the expected value of Y for fixed value of x.

Recall the notation

- Y - a random variable with a probability distribution
- y - a fixed value that the variable Y can take.

Fundamental Problem: If we believe the straight-line model with error is correct, how do we find the values of parameters β_0 and β_1. We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.

Least Squares Fitting

We select the best values of β_0 and β_1 by minimizing the *error in fit*. For two data points (x_1, y_1) and (x_2, y_2), the errors in fit are

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively. But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out. Therefore we look at **squared** errors (as a large negative error is as bad as a large positive error)

$$\begin{aligned} e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\ e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2 \end{aligned}$$

For n data, we obtain n misfit squared errors

$$e_1^2, \ldots, e_n^2$$

We select β_0 and β_1 as the values of the parameters that minimize SSE, where

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We wish to make the total misfit squared error as small as possible.

SSE - sum of squared errors.

We could write

$$SSE = SSE(\beta_0, \beta_1)$$

to show the dependence of SSE on the parameters.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved **analytically**.

It follows that the best parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- Sum of Squares SS_{xx}:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- Sum of Squares SS_{xy}:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the **least-squares estimates**

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the **least-squares line of best fit**. The **fitted-values** are

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the **residuals** or **residual errors** are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Model Assumptions for Least-Squares

To utilize least-squares for the probabilistic model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

we make the following assumptions

1. The expected error $E[\epsilon]$ is zero so that

$$E[Y] = \beta_0 + \beta_1 x$$

2. The variance of the error, $Var[\epsilon]$, is constant and does not depend on x.
3. The probability distribution of ϵ is a **symmetric** distribution about zero; a stronger assumption is that ϵ is **Normally** distributed.
4. The errors for two different measured responses are **independent**, i.e. the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Parameter Estimation: Estimating σ^2

Using the LS procedure, we can construct an estimate of the *error* or *residual error* variance

Recall that

$$Var[\epsilon] = \sigma^2$$

An estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = s^2$$

say.

The denominator $n-2$ is a *degrees of freedom* parameter of the form

$$\begin{array}{ccc}\text{TOTAL NUMBER} & - & \text{NUMBER OF PARAMETERS} \\ \text{OF DATA} & & \text{ESTIMATED}\end{array}$$

or $n-p$, where in the simple linear regression, $p=2$ ($\widehat{\beta}_0$ and $\widehat{\beta}_1$). Note also that

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

where

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Mathematical Formulation

To allow for generalization of the model, we consider a matrix formulation.

For data pairs $(x_1, y_1), \ldots, (x_n, y_n)$ we form

- $\mathbf{y} = [y_1, \ldots, y_n]^T$ (an $n \times 1$ column vector)
- $n \times 2$ matrix $\mathbf{X}$, where the first column of $\mathbf{X}$ is a column of 1s, and the second column is

  $$[x_1, \ldots, x_n]^T$$

  Thus the ith row is $\mathbf{x}_i = [1 \ x_i]^T$.
- $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ (an $n \times 1$ column vector)

The notation ^{T} means matrix transpose

Mathematical Formulation

We may then write the model in vector form

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

where $\mathbf{X}\beta$ is the matrix multiplication product of $\mathbf{X}$ and β, which yields an $n \times 1$ column vector.

The SSE quantity is then

$$\begin{aligned} SSE(\beta) &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \\ &= \sum_{i=1}^{n} (y_i - \mathbf{x}_i\beta)^2 \end{aligned}$$

Least-Squares Estimates

It can be shown that the least-squares estimates of the model parameters are given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where

- $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0, \widehat{\beta}_1]^T$ are the **least-squares estimates**
- the notation $^{-1}$ means matrix inversion

Least-Squares Estimates

The estimate of σ^2 is given by calculation of

$$\begin{aligned} SSE(\widehat{\beta}) &= (\mathbf{y} - \mathbf{X}\widehat{\beta})^T(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y} \end{aligned}$$

and then

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta})}{n-p}$$

where $p = 2$ is the dimension of β.

Least-Squares Estimates

The estimated standard errors of $\widehat{\beta}$ are given by the matrix variance calculation

$$\mathrm{Var}[\widehat{\beta}] = \widehat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$$

and the estimated standard errors are the diagonal elements of this 2×2 matrix.

These calculations look complicated, but are actually easy to compute. In the case of simple linear regression

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & S_x \\ S_x & S_{xx} \end{bmatrix}$$

where

$$S_x = \sum_{i=1}^{n} x_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2$$

Least-Squares Estimates

Then

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix}$$

and hence

$$\widehat{\beta} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx} & -S_x \\ -S_x & n \end{bmatrix} \begin{bmatrix} S_y \\ S_{xy} \end{bmatrix}$$

where

$$S_y = \sum_{i=1}^{n} y_i \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Least-Squares Estimates

Hence

$$\widehat{\boldsymbol{\beta}} = \frac{1}{nS_{xx} - S_x^2} \begin{bmatrix} S_{xx}S_y - S_xS_{xy} \\ nS_{xy} - S_xS_y \end{bmatrix}$$

and

$$\begin{aligned} \widehat{\beta}_0 &= \frac{S_{xx}S_y - S_xS_{xy}}{nS_{xx} - S_x^2} \\ \widehat{\beta}_1 &= \frac{nS_{xy} - S_xS_y}{nS_{xx} - S_x^2} = \frac{nSS_{xy}}{nSS_{xx}} \end{aligned}$$

as before.

Extending the model

The great advantage of writing the model in this apparently more complicated form is that the model can be **extended** from the simple straight line in a very straightforward way.

Provided the model can be written in the **Linear Model** form

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

where $\boldsymbol{\beta}$ contains the **parameters** in linear form, then the least-squares can be found easily.

Extending the model

Crucially, the form of the matrix $\mathbf{X}$ does not change the basic approach to estimation. Provided we can invert $\mathbf{X}^T\mathbf{X}$, we can form the LS estimates. Thus $\mathbf{X}$ can include

- more terms in involving x $(x^2, x^3, \ldots)$ (*polynomial regression*)
- non-linear functions of x $(\log x, \exp(x), \ldots)$
- other predictors if they available (*multiple regression*)

Extending the model

Example: A Harmonic Linear Regression Model.

Suppose that

$$f(x) = \beta_0 + \sum_{j=1}^{k} [\beta_{1j} \cos(\lambda_j x) + \beta_{2j} \sin(\lambda_j x)]$$

where $\lambda_1, \ldots, \lambda_k$ are frequencies

$$0 < \lambda < 2\pi.$$

The λ_j introduce **periodic** components that contribute to the overall signal variation.

Extending the model

Example: A Harmonic Linear Regression Model.

$k = 1$: Suppose that we wanted to fit a model with a 24 hour cycle to hourly data. Then we would choose $\lambda_1 = 2\pi/24$

$$f(x) = \beta_0 + \beta_{11}\cos(\lambda_1 x) + \beta_{21}\sin(\lambda_1 x)$$

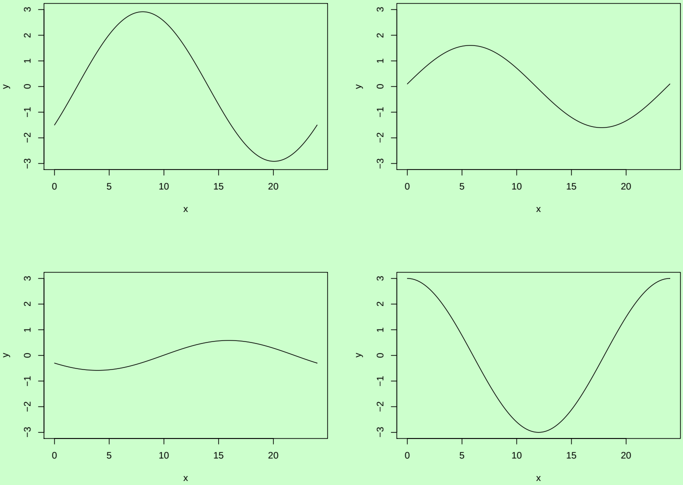
and for different choices of $\beta_0, \beta_{11}, \beta_{21}$ we can obtain different response profiles.

Without loss of generality, we assume $\beta_0 = 0$.

Extending the model

Example: A Harmonic Linear Regression Model.

$k = 1$:



Extending the model

Example: A Harmonic Linear Regression Model.

$k = 2$: Suppose that we wanted to fit a model with a 24 hour cycle and a 36 hour cycle to hourly data. Then we would choose $\lambda_1 = 2\pi/24$ and $\lambda_2 = 2\pi/36$

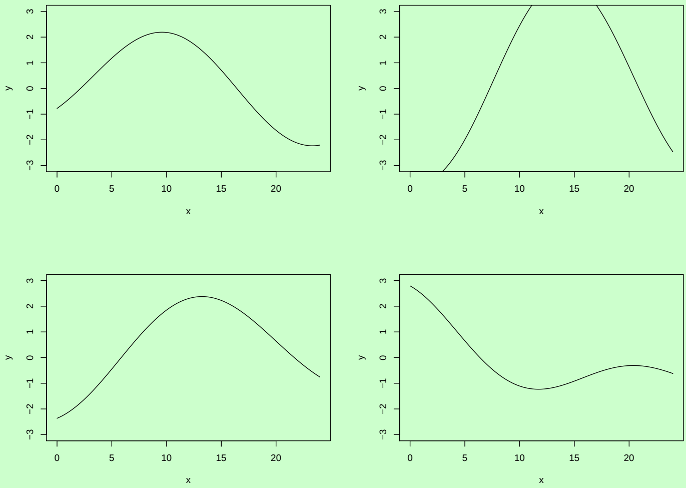
$$\begin{aligned} f(x) &= \beta_0 + \beta_{11}\cos(\lambda_1 x) + \beta_{21}\sin(\lambda_1 x) \\ &\quad + \beta_{12}\cos(\lambda_2 x) + \beta_{22}\sin(\lambda_2 x) \end{aligned}$$

The **period**, κ of the cycle is the reciprocal of the **characteristic frequency**, $\lambda/(2\pi)$

Extending the model

Example: A Harmonic Linear Regression Model.

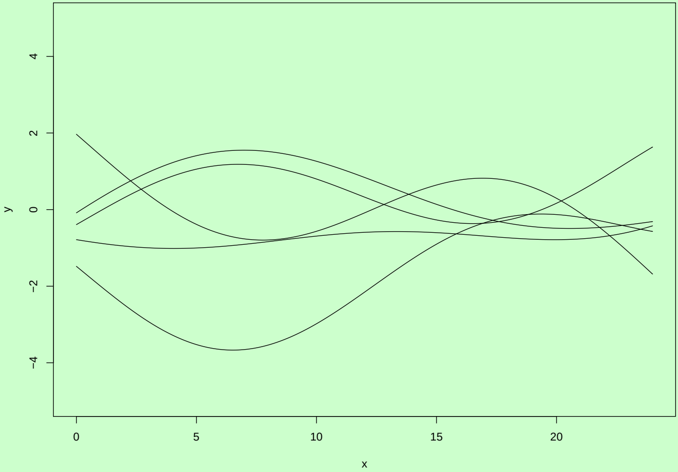
$k = 2$:



Extending the model

Example: A Harmonic Linear Regression Model.

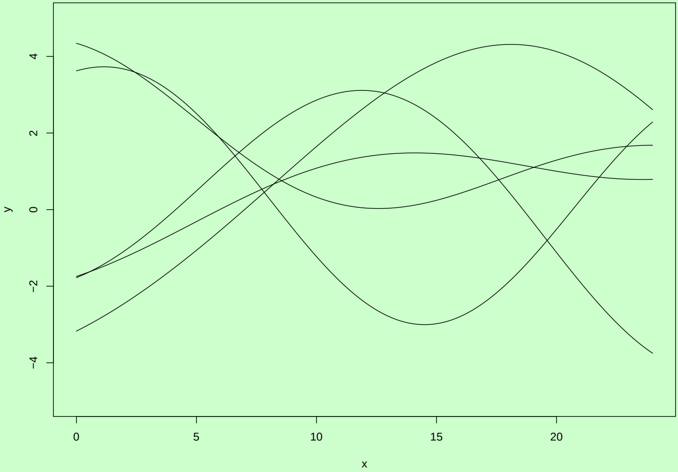
$k = 4$:



Extending the model

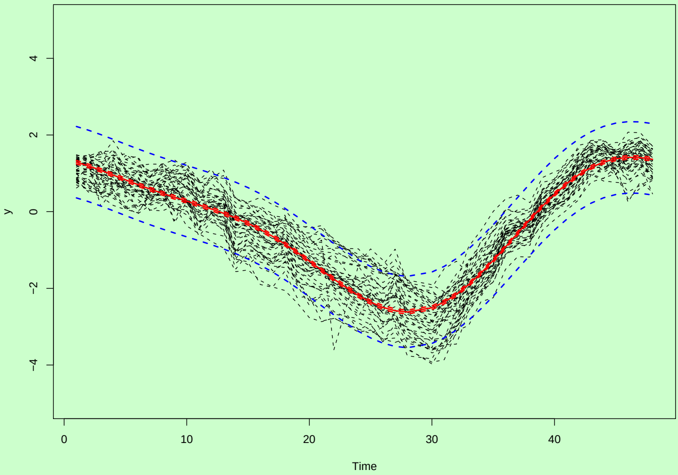
Example: A Harmonic Linear Regression Model.

$k = 8$:



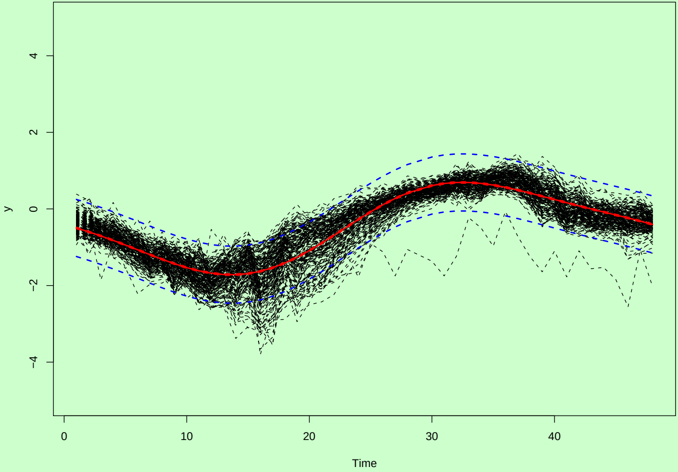
CAMDA Data Clusters

Example: CAMDA Data: Cluster 1 ($k = 2$).



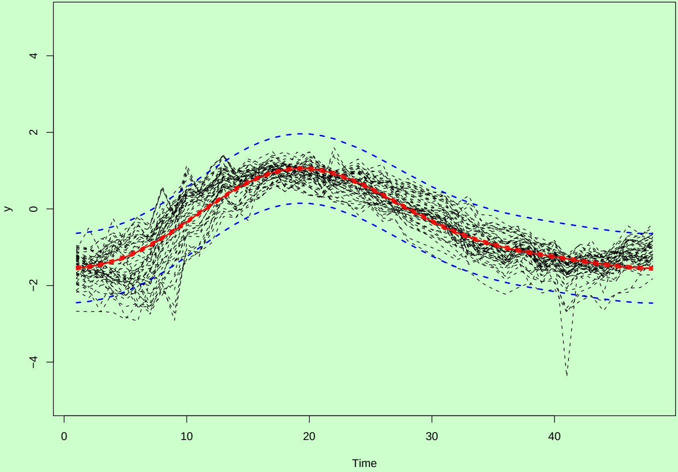
CAMDA Data Clusters

Example: CAMDA Data: Cluster 5 ($k = 2$).



CAMDA Data Clusters

Example: CAMDA Data: Cluster 10 ($k = 2$).



Extending the model

Note that

(i) The function $f(x)$ is **continuous** in x

(ii) For an observed data series, there is a limit to the number of terms we can include.

(iii) The model is **still a linear model** ! The design matrix $\mathbf{X}$ has n rows and $2k+1$ columns, and row i takes the form

$$\mathbf{x}_i = [1 \ \cos(\lambda_1 x_i) \ \sin(\lambda_1 x_i) \quad \cdots \quad \cos(\lambda_k x_i) \ \sin(\lambda_k x_i)]$$

This means that the usual least-squares approach can still be used to fit the models to data.

Extending the model

Example: A Piecewise Linear Model.

Define the truncation function $()_+$ as follows

$$(x)_+ = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Consider the piecewise linear model

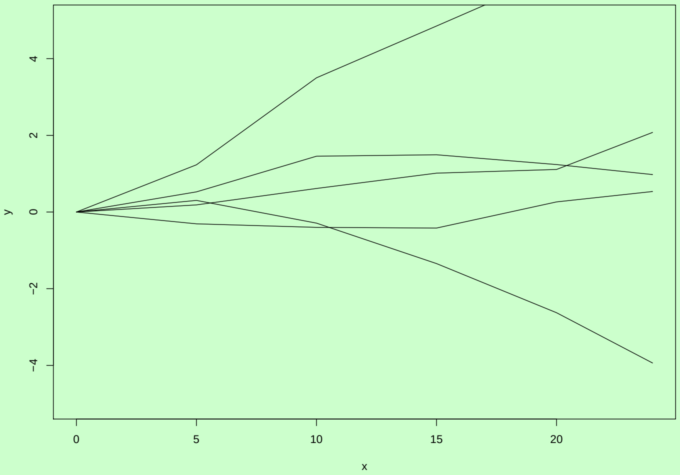
$$f(x) = \beta_0 + \sum_{j=1}^{k} \beta_j (x - \kappa_j)_+$$

where $0 = \kappa_1 < \kappa_2 < \cdots < \kappa_k$.

This f is also continuous, it is non-linear in x, but is still a linear model in terms of the parameters.

Extending the model

Example: A Piecewise Linear Model.



Extending the model

Example: A Piecewise Quadratic Model.

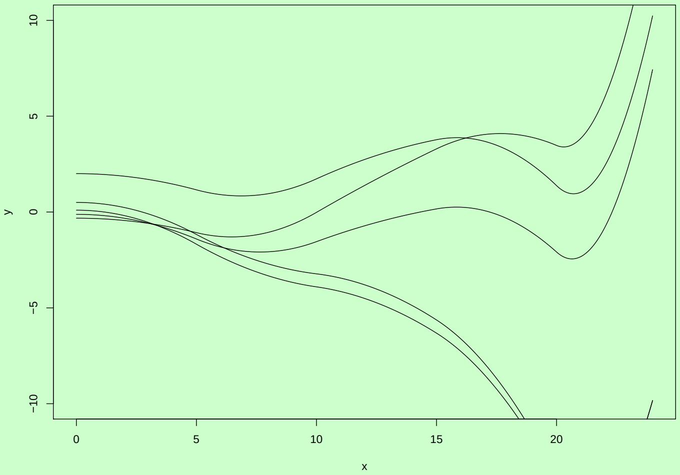
Consider the piecewise quadratic model

$$f(x) = \beta_0 + \sum_{j=1}^{k} \beta_j (x - \kappa_j)_+^2$$

Another continuous, non-linear in x, but linear in terms of the parameters model .

Extending the model

Example: A Piecewise Quadratic Model.



Summary

The formulation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \sum_{j=0}^{k} \beta_j g_j(\mathbf{x}) + \boldsymbol{\epsilon}$$

is still a linear model. Therefore least-squares fitting is straightforward.

The functions

$$g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_k(\mathbf{x})$$

are often called **basis functions**.

Summary

It is possible to construct sophisticated functions $f(x)$ by means of this linear model construction; the precise form of the design matrix $\mathbf{X}$ will depend on the choice of the basis functions used to decompose f.

- linear piecewise
- quadratic, cubic
- splines
- Fourier (sine and cosine)
- "wavelets"

We need to be able to form $\mathbf{X}$ and compute $(\mathbf{X}^T\mathbf{X})^{-1}$, but often that can be done routinely.

Extension to Correlated Errors

The formulation

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

where the entries in $\boldsymbol{\epsilon}$ are uncorrelated can be extended to the more general correlated case. If

$$Var[\boldsymbol{\epsilon}] = \Sigma$$

then

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}$$

are the **Generalized Least Squares** estimates.

Model-based Clustering

We now utilize the basis function models above to facilitate a model-based approach to clustering.

The advantages of using a flexible model-based approach are that

- we can more accurately represent the likely structure in the data
- we can perform model selection, that is, choose between plausible alternative models
- we can integrate the clustering analysis with other results or information.

Bayesian Models for the Profiles

We model the gene expression profiles using a linear regression model and non-linear basis functions in a Bayesian setting.

The model induces a stochastic process structure for the underlying variation in expression; our clustering approach utilizes the covariance structure of this process.

Aim is to use models that capture the characteristic behaviour of expression profiles corresponding to different forms of regulation.

Generically, we wish to capture the behaviour of the gene expression ratio y as a function of time t and measurement error.

Recall: Regression Representation

For gene i at time t

$$y_{i_t} = X(t)\beta + \varepsilon_t$$

where $X(t)$ is a p-vector of specified basis functions of t, β is a p-vector of basis coefficient parameters

Note: $\{\varepsilon_t\}$ is an independent and Gaussian error process;

- plausible for our experiments
- assumption can be relaxed
- $\{Y_{i_t}, t = 1, \ldots, T\}$ conditionally independent, unconditionally dependent.

Truncated Polynomial Spline Basis Model

We utilize the following basis function model:

$$X(t)\beta = \beta_0 + \sum_{j=1}^{p} \beta_j (t - \kappa_j)_+^q$$

for $q = 1, 2, ...$, where $(\kappa_1, \ldots, \kappa_p)$ are knot positions spanning the range of t, and

$$(t - \kappa_j)_+^q = \max\{0, (t - \kappa_j)\}^q$$

This function is continuous at the knot points; here we presume that the knot positions are **fixed at the data ordinates**.

Any suitable basis function set may be used.

Bayesian Inference

For inference, under the assumption that the random error terms $\{\varepsilon_{i_t}\}$ form an i.i.d Gaussian sequence with variance σ^2.

The conditional distribution of the concatenated response vectors of N genes Y is multivariate normal

$$\mathbf{Y}|\mathbf{X}, \beta, \sigma^2 \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_{NT}\right)$$

where now X is $NT \times p$ and I_{NT} is the NT-dimensional identity matrix.

Bayesian Inference

Rather than use the least-squares approach we adopt the **Bayesian framework** to obtain inference and model assessment.

The Bayesian framework requires the specification of **prior distributions** for the parameters of interest. These prior distributions can usually be specified from genuine prior knowledge of the experiment. For example,

- Microarray data have a measurement range (on the log-relative expression scale) of -5 to 5.
- Regulation of gene expression in most cases causes variation on the range -2 to 2
- The residual error variance is no greater than 1.

and so on.

Bayesian Inference: Prior distribution

The simplest Bayesian analysis uses a **conjugate prior specification** for (β, σ^2)

$$p\left(\beta|\sigma^2\right) \equiv \mathrm{N}\left(m, \sigma^2 V\right) \qquad p\left(\sigma^2\right) \equiv \mathrm{IGamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right)$$

m is $p \times 1$, V is $p \times p$ positive definite and symmetric, all other parameters are scalars.

Bayesian Inference: Posterior distribution

Then the posterior distribution can be computed as

$$p\left(\beta|\mathbf{y},\sigma^2\right) \equiv N\left(m^*,\sigma^2 V^*\right)$$

and

$$p\left(\sigma^2|\mathbf{y}\right) \equiv IGamma\left(\frac{T+\alpha}{2},\frac{c+\gamma}{2}\right)$$

which summarizes the information about the parameters in the data.

In the case where the prior parameters take limiting values, we recover the least-squares estimates.

Terms in the Posterior

We consider a **centered parameterization** for β so that $m = 0$, giving

$$m^* = \left(\mathbf{X}^T\mathbf{X} + V^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y} \qquad V^* = \left(\mathbf{X}^T\mathbf{X} + V^{-1}\right)^{-1}$$

and

$$\begin{aligned} c &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + V^{-1}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{y}^T\left(\mathbf{I}_n - X\left(\mathbf{X}^T\mathbf{X} + V^{-1}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \end{aligned}$$

Marginal Likelihood

Our clustering approach will be based on marginal likelihood considerations.

$$p(\mathbf{y}) = \int\int p\left(\mathbf{y}|\boldsymbol{\beta},\sigma^2\right) p\left(\boldsymbol{\beta}|\sigma^2\right) p\left(\sigma^2\right) d\beta d\sigma^2.$$

so that, for our basis function representation of a **single profile**

$$p(\mathbf{y}) = \left(\frac{1}{\pi}\right)^{T/2} \frac{\gamma^{\alpha/2}\Gamma\left(\dfrac{T+\alpha}{2}\right)}{\Gamma\left(\dfrac{\alpha}{2}\right)} \frac{|V^*|^{1/2}}{|V|^{1/2}} \frac{1}{\{c+\gamma\}^{(T+\alpha)/2}}$$

A Cautionary Note

It might be tempting to use a vague prior specification, where the prior input is as minimal as possible

However, a vague prior specification $V^{-1} \to 0$ in leads to $p(y) \to 0$ and impropriety or indeterminacy.

Thus a fully non-informative prior specification cannot be used, and can lead to the *Lindley-Bartlett paradox*; here this corresponds to choosing the model with 1 cluster, irrespective of the data.

In practice, V can be chosen from prior knowledge, or to approximately maximize the marginal likelihood.

Hierarchical Clustering

- a method of organizing a collection of objects into disjoint sets
- uses a similarity/discrepancy measure/overall potential function
- *agglomerative clustering* places each of the N items in its own cluster, and then recursively (optimally) merges currently existing clusters until a single cluster remains (i.e. performs $N-1$ merge operations in total)
- to find the ith optimal agglomeration requires $(N+1-i)(N-i)/2$ comparisons.
- at worst, an $O(N^3)$ procedure; here, $N \approx 2800$

Computationally Efficient Clustering

For both Euclidean distance clustering and our proposed method, the number of calculations is drastically reduced by noting the distance between any two clusters remains unchanged through successive iterations until one is agglomerated with another cluster.

The distance between two clusters will be based on the **change in marginal likelihood** caused by merging clusters.

The the number of marginal likelihood calculations actually required is

$$N^2 - N - 1$$

and so only $O\left(N^2\right)$.

Bayesian Clustering

We propose clustering on the basis of the **covariance structure** induced by the underlying stochastic process rather than on Euclidean distances.

- hierarchical clustering approach assigns profiles to the same cluster if they are similar in **covariance** terms.
- covariance determined by $\mathbf{X}$ and V; marginally, vector response Y_i has covariance

$$\mathbf{X} V \mathbf{X}^T + I$$

Advantages of the approach

- it respects the time ordering of the data
- it can lead to biologically appropriate covariance structures being discovered, as it can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.
- we wish to cluster using the marginal probability that the objects came from the same Gaussian process.
- choice of covariance matrix, induced by the basis function representation, leads to important simplifications in the calculation of the marginal likelihoods of each cluster, and for the hierarchical steps.

Computational Simplification

Let $y_{(K)} = (y_1, y_2, \ldots, y_{N_K})$ be N_K expression profiles in the Kth cluster. Then $Y_{(K)}|X_K, \beta, \sigma^2 \sim \mathrm{N}\left(X_K\beta, \sigma^2 I_{T_K}\right)$. As each profile has

- the same number of observed measurements T,
- identical observation points, then

$$\mathbf{X}_K^T = \left[\begin{array}{cccc} \mathbf{X}^T & \mathbf{X}^T & \cdots & \mathbf{X}^T \end{array}\right]$$

$$\mathbf{X}_K^T\mathbf{X}_K = N_K\mathbf{X}^T\mathbf{X} \qquad \mathbf{X}_K^T\mathbf{y}_{(K)} = \sum_{i=1}^{N_K}\mathbf{X}^T y_i.$$

with $\mathbf{Y}_{(K)}$ having length TN_K, and $\mathbf{X}_K$ being $(TN_K \times p)$.

Computational Simplification

Hence the pivotal quantities in the marginal likelihood can be presented in simple form, giving

$$p(\mathbf{y}_{(K)}) = \frac{g\left(N_K T, \alpha, \gamma\right) |V|^{-1/2}}{|N_K \mathbf{X}^T \mathbf{X} + V^{-1}|^{1/2} \left\{c_K + \gamma\right\}^{(N_K T + \alpha)/2}}.$$

where c_K is given by

$$\left(\sum_{i=1}^{N_K} \mathbf{y}_i^T \mathbf{y}_i\right) - \left(\sum_{i=1}^{N_K} \mathbf{X}^T \mathbf{y}_i\right) \left(N_K \mathbf{X}^T \mathbf{X} + V^{-1}\right)^{-1} \left(\sum_{i=1}^{N_K} \mathbf{X}^T \mathbf{y}_i\right)^T$$

Computational Simplification

Thus for each gene i we can compute

$$\mathbf{y}_i^T \mathbf{y}_i, \qquad \mathbf{X}^T \mathbf{y}_i$$

at the start, and then simply take sums over all genes in a cluster to get the required quantities.

Furthermore, $\forall n \in \{1, \ldots, N\}$ we can also compute

$$W_n^{-1} = \left(n\mathbf{X}^T\mathbf{X} + V^{-1}\right)^{-1}, \qquad |W_n|$$

beforehand, as these $\{n\}$ are the values which N_K will take.

A Full Bayesian Specification

To complete the Bayesian model, we need to specify the prior on

- the number of clusters, $\mathcal{C}$,
- the cluster sizes, $N_1, \ldots, N_C$,
- the cluster decomposition, $Z_1, \ldots, Z_n$

Natural default choice uses the specification

$$P(\mathcal{C} = C)P(N_1, \ldots, N_C | \mathcal{C} = C)P(Z_1, \ldots, Z_n | N_1, \ldots, N_C)$$

- $P(\mathcal{C} = C) = 1/N, C = 1, 2, \ldots, N$
- $P(N_1, \ldots, N_C | \mathcal{C} = C)$ Multinomial/Dirichlet
- $P(Z_1, \ldots, Z_n | N_1, \ldots, N_C)$ Uniform allocation

The Algorithm

- Starting with N clusters, each cluster containing the expression levels for one gene.
- At each step, merge the two clusters which cause the biggest increase (or smallest decrease) in the overall marginal likelihood
  - The posterior probability can be used in place of the marginal likelihood.
  - The Bayes Information Criterion (BIC) can also be used.
- Continue until one cluster remains.
- Take the optimal number of clusters as that which maximized the marginal likelihood/BIC/posterior probability.

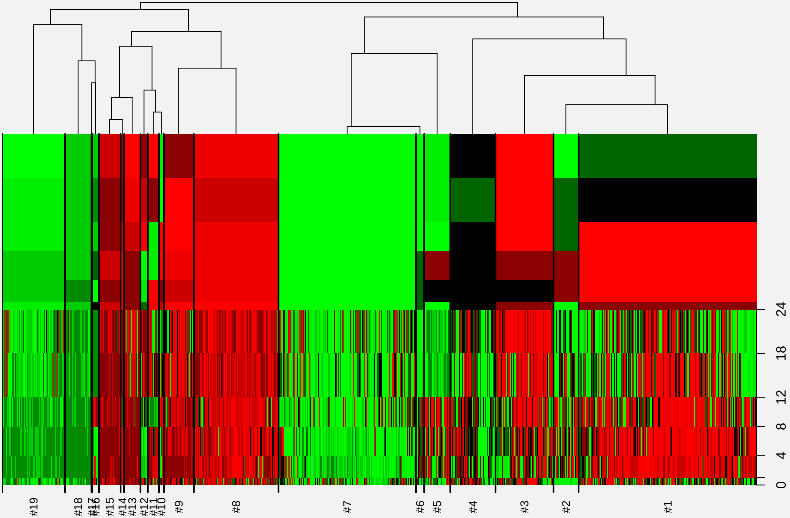
Results for the challenge experiment

- Multinomial-Dirichlet prior on the number and size of clusters.
- Linear splines $q = 1$, knots at data points.

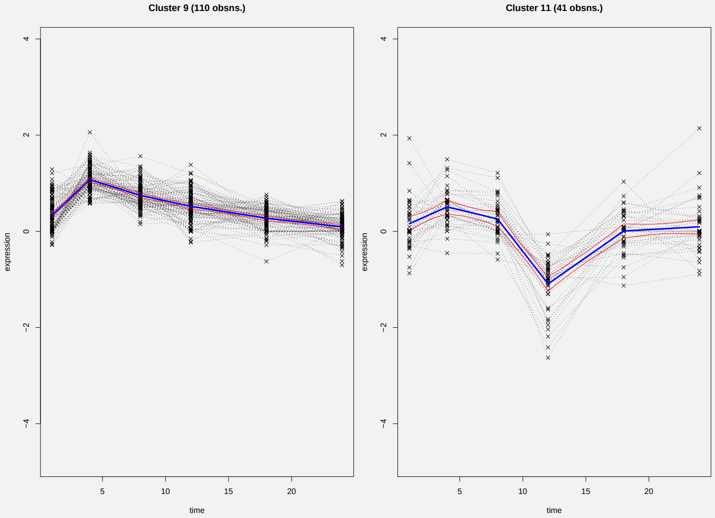
$$X(t)\beta = \beta_0 + \sum_{j=1}^{p} \beta_j (t - \kappa_j)_+$$

Optimal number of clusters was 19.

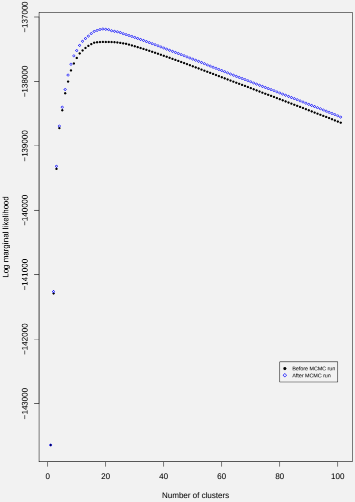
Clustered Series



An Immune Defence Cluster ?



Number of clusters



Results

- Get 'tighter' clusters than Euclidean hierarchical clustering.
- Computationally feasible - agglomerative clustering takes $\sim$ 2 minutes on a 2Gb processor PC.
- Probabilistic model enables inference on the number of clusters, and provides a rigorous framework for classification of unknown genes.
- More advanced methods can be use to better optimize the clusterings
  - *Markov chain Monte Carlo* (MCMC) takes longer, but can give clustering improvements.
  - *Simulated Annealing* MCMC can find even higher probability regions.

References

- Heard, NA, Holmes CC, Stephens DA, Hand, DJ and Dimopoulos, G, (2005), Bayesian co-clustering of gene expression profiles from multiple parallel immune defence Challenges, *PNAS*, **102** (47): Pages 16939-16944.
- Heard, NA, Holmes CC, Stephens DA, (2006), A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves, *Journal of the American Statistical Association*, Vol **101**, No 1, Pages: 18 - 29