Complexity in Systems Level Biology and Genetics: Statistical Perspectives

David A. Stephens[1]

Department of Mathematics and Statistics
McGill University, Montreal

Definition Of The Subject And Its Importance

This chapter identifies the challenges posed to biologists, geneticists and other scientists by advances in technology that have made the observation and study of biological systems increasingly possible. High-throughput platforms have made routine the collection vast amounts of structural and functional data, and have provided insights into the working cell, and helped to explain the role of genetics in common diseases. Associated with the improvements in technology is the need for statistical procedures that extract the biological information from the available data in a coherent fashion, and perhaps more importantly, can quantify the certainty with which conclusions can be made. This chapter outlines a biological hierarchy of structures, functions and interactions that can now be observed, and detail the statistical procedures that are necessary for analyzing the resulting data. The chapter has four main sections. The first section details the historical connection between statistics and the analysis of biological and genetic data, and summarizes fundamental concepts in biology and genetics. The second section outlines specific mathematical and statistical methods that are useful in the modelling of data arising in bioinformatics. In sections three and four, two particular issues are discussed in detail: functional genomics via microrray analysis, and metabolomics. Section five identifies some future directions for biological research in which statisticians will play a vital role.

Glossary

[1]Department of Mathematics and Statistics, McGill University, Montreal, H3A 2K6, Canada, Tel: +1-514-398-2005, Fax: +1-514-398-3899, Email: d.stephens@math.mcgill.ca

1. Introduction

The observation of biological systems, their processes and inter-reactions, is one of the most important activities in modern science. It has the capacity to provide direct insight into fundamental aspects of biology, genetics, evolution, and indirectly will inform many aspects of public health. Recent advances in technology - high-throughput measurement platforms, imaging - have brought a new era of increasingly precise methods of investigation. In parallel to this, there is an increasingly important focus on statistical methods that allow the information gathered to be processed and synthesized. This chapter outlines key statistical techniques that allow the information gathered to be used in an optimal fashion.

Although its origin is dated rather earlier, the term *Systems Biology* (see for example, [1, 2, 3]) has, since 2000, been used to describe the study of the operation of biological systems, using tools from mathematics, statistics and computer science, supplanting *computational biology* and *bioinformatics* as an all-encompassing term for quantitative investigation in molecular biology. Most biological systems are hugely complex, involving chemical and mechanical processes operating at different scales. It is important therefore that information gathered is processed coherently, according to self-consistent rules and practices, in the presence of the uncertainty induced by imperfect observation of the underlying system. The most natural framework for coherent processing of information is that of probabilistic modelling

1.1. Statistical versus Mathematical Modelling

There is a great tradition of mathematical and probabilistic modelling of biology and genetics; see **(author?)** [4] for a thorough review. The mathematization of biology, evolution and heredity began at the end of the nineteenth century, and continued for the first half of the twentieth century, by far pre-dating the era of molecular biology and genetics that culminated at the turn of the last millennium with the human genome project. Consequently, the mathematical models of, say, evolutionary processes that were developed by Yule [5] and Fisher and Wright [6, 7, 8], and classical models of heredity, could only be experimentally verified and developed many years after their conception. It could also be convincingly argued that through the work of F. Galton, K. S. Pearson and R. A. Fisher, modern statistics has its foundation in biology and genetics.

In parallel to statistical and *stochastic* formulation of models for biological systems, there has been a more recent focus on the construction of *deterministic* models to describe observed biological phenomena. Such models fall under the broad description *Mathematical Biology*, and have their roots in applied mathematics and dynamical systems; see, for example, **(author?)** [9, 10] for a comprehensive treatment. The distinction between stochastic and deterministic models is important to make, as the objectives and tools used often differ considerably. This chapter will restrict attention to stochastic models, and the processing of observed data, and thus is perhaps more closely tied to the immediate interests of the scientist, although some of the models utilized will be inspired by mathematical models of the phenomena being observed.

1.2. Fundamental Concepts in Biology and Genetics

To facilitate the discussion of statistical methods applied to systems biology, it is necessary to introduce fundamental concepts from molecular biology and genetics; see the classic text **(author?)** [11] for full details. Attention is restricted to eukaryotes, organisms whose cells are constructed to contain a nucleus within coding information is encapsulated.

- The cell is a complex architecture containing several nuclear domains [12] whose organization is not completely understood, but the fundamental activity that occurs within the nucleus is the production and distribution of proteins.
- Dioxyribonucleic acid (DNA) is a long string of nucleotides that encodes biological information, and that is copied or *transcribed* into ribonucleic acid (RNA), which in turn enables the formation of proteins. Specific segments of the DNA, genes, encode the proteins, although non-coding regions of DNA - for example, promoter regions, transcription factor binding sites - also have important roles. Genetic variation at the nucleotide level, even involving a single nucleotide, can disrupt cellular activity. In humans and most other complex organisms, DNA is arranged into chromosomes, which are duplicated in the process of mitosis. The entire content of the DNA of an organism is termed the *genome*.
- Proteins are macromolecules formed by the *translation* of RNA, comprising amino acids arranged (in primary structure) in a linear fashion, comprising domains with different roles, and physically configured in three dimensional space. Proteins are responsible for all biological activities that take place in the cell, although proteins may have different roles in different tissues at different times, due to the *regulation* of transcription.
- Proteins interact with each other in different ways in different contexts in interaction networks that may be dynamically organized. Genes are also regarded as having indirect interactions through gene regulatory networks.
- Genetic variation amongst individuals in a population is due to mutation and selection, which can be regarded as stochastic mechanisms. Genetic information in the form of DNA passes from parent to offspring, which promulgates genetic variation. Individuals in a population are typically related in evolutionary history. Similarly, proteins can also thought to be related through evolutionary history.
- Genetic disorders are the result of genetic variation, but the nature of the genetic variation can be large- or small-scale; at the smallest scale, variation in single nucleotides (*single nucleotide polymorphims* (SNPs)) can contribute to the variation in observed traits.

Broadly, attention is focussed on the study of *structure* and *function* of DNA, genes and proteins, and the nature of their *interactions*. It is useful, if simplistic, to view biological activities in terms of an organizational hierarchy of inter-related chemical reactions at the level of DNA, protein, nucleus, network and cellular levels. A holistic view of mathematical modelling and statistical inference requires the experimenter to model simultaneously actions and interactions of all the component features, whilst recognizing that the component features cannot observed directly, and can only be studied through separate experiments on often widely different platforms. It is the role of the bioinformatician or systems biologist to synthesize the data available from separate experiments in an optimal fashion.

2. Mathematical Representations of The Organizational Hierarchy

A mathematical representation of a biological system is required that recognizes, first, the complexity of the system, secondly, its potentially temporally changing nature, and thirdly the inherent uncertainties that are present. It is the last feature that necessitates the use of probabilistic or stochastic modelling.

An aphorism commonly ascribed to D.V. Lindley states that "*Probability is the language of uncertainty*"; probability provides a coherent framework for processing information in the presence of imperfect knowledge, and through the paradigm of Bayesian theory [13] provides the mathematical template for statistical inference and prediction. In the modelling of complex systems, three sorts of uncertainty are typically present

- *Uncertainty of Structure:* Imperfect knowledge of the connections between the interacting components is typically present. For example, in a gene regulatory network, it may be possible via the measurement of gene co-expression to establish which genes interact within the network, but it may not be apparent precisely how the organization of regulation operates, that is which genes regulate the expression of other genes.
- *Uncertainty concerning Model Components :* In any mathematical or probabilistic model of a biological system, there are model components (differential equations, probability distributions, parameter settings) that must be chosen to facilitate implementation of the model. These components reflect, but are not determined by, stuctural considerations.
- *Uncertainty of Observation :* Any experimental procedure carries with it uncertainty induced by the measurement of underlying system, that is typically subject to random measurement error, or noise. For example, many biological systems rely on imaging technology, and the extraction of the level of signal of a fluorescent probe, for a representation of the amount of biological material present. In microarray studies (see section §3.1.), comparative hybridization of messenger RNA (mRNA) to a medium is a technique for measuring gene expression that is noisy due to several factors (imaging noise, variation in hybridization) not attributable to a biological cause.

The framework to be built must handle these types of uncertainty, and permit inference about structure and model components.

2.1. Models derived from Differential Equations

A deterministic model reflecting the dynamic relationships often present in biological systems may be based on the system of ordinary differential equations (ODEs)

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) \tag{1}$$

where $\boldsymbol{x}(t) = (x_1(t), \ldots, x_d(t))^T$ represent the levels of d quantities being observed $\dot{\boldsymbol{x}}(t)$ represents time derivative, and $\boldsymbol{g}$ is some potentially non-linear system of equations, that may be suggested by biological prior knowledge or prior experimentation. The model in equation (1) is a classical "Mathematical Biology" model, that has been successful in representing forms of organization in

many biological systems (see, for example **(author?)** [14] for general applications). Suppressed in the notation is a dependence on system parameters, $\boldsymbol{\theta}$, a k-dimensional vector that may be presumed fixed, and "tuned" to replicate observed behaviour, or estimated from observed data. When data representing a partial observation of the system are available, inferences about $\boldsymbol{\theta}$ can be made, and models defined by ODE systems are of growing interest to statisticians; see, for example, **(author?)** [15, 16, 17]).

Equation (1) can be readily extended to a *stochastic differential equation* (SDE) system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) + d\boldsymbol{z}(t) \tag{2}$$

where $\boldsymbol{z}(t)$ is some stochastic process that renders the solution to equation (2) a stochastic process (see, for example, **(author?)** [18] for a comprehensive recent summary of modelling approaches and inference procedures, and a specific application in **(author?)** [19]). The final term $d\boldsymbol{z}(t)$ represents the infinitesimal stochastic increment in $\boldsymbol{z}(t)$. Such models, although particularly useful for modelling activity at the molecular level, often rely on simplifying assumptions (linearity of $\boldsymbol{g}$, Gaussianity of $\boldsymbol{z}$) and the fact that the relationship structure captured by $\boldsymbol{g}$ is known. Inference for the parameters of the system can be made, but in general require advanced computational methods (Monte Carlo (MC) and Markov chain Monte Carlo (MCMC)).

2.2. Probabilistic Graphical Models

A simple and often directly implementable approach is based on a *probabilistic graphical model*, comprising a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, described by a series of nodes $\mathcal{N}$, edges $\mathcal{E}$, and a collection of random variables $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ placed at the nodes, all of which may be dynamically changing. See, for example [20] for a recent summary, [21, Chapter 2] for mathematical details and [22] for a biological application.

The objective of constructing such a model is to identify the joint probability structure of $\boldsymbol{X}$ given the graph $\mathcal{G}$, which possibly is parameterized by parameters ϕ, $f_{\boldsymbol{X}}(\boldsymbol{x}|\phi, \mathcal{G})$. In many applications, $\boldsymbol{X}$ is not directly observed, but is instead inferred from observed data, $\boldsymbol{Y}$, arising as noisy observations derived from $\boldsymbol{X}$. Again, a k-dimensional parameter vector $\boldsymbol{\theta}$ helps to characterize the stochastic dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ by parameterizing the conditional probability density $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$. The joint probability model encapsulating the probabilistic structure of the model is

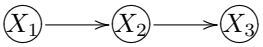
$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \phi, \mathcal{G}) = f_{\boldsymbol{X}}(\boldsymbol{x}|\phi, \mathcal{G}) f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{3}$$

The objectives of inference are to learn about $\mathcal{G}$ (the uncertain structural component) and parameters $(\boldsymbol{\theta}, \phi)$ (the uncertain model parameters and observation components).

The graph structure $\mathcal{G}$ is described by $\mathcal{N}$ and $\mathcal{E}$. In holistic models, $\mathcal{G}$ represents the interconnections between interacting modules (genomic modules, transcription modules, regulatory modules, proteomic modules, metabolic modules etc.) and also the interconnections within modules in the form of subgraphs. The nodes $\mathcal{N}$ (and hence $\boldsymbol{X}$) represent influential variables in the model structure, and the edges $\mathcal{E}$ represent dependencies. The edge connecting two nodes, if present, may be *directed* or *undirected* according to the nature of the influence; a directed edge indicates the direction of *causation*, an undirected edge indicates a *dependence*.

Causality is a concept distinct from dependence (association, covariation or correlation), and represents the influence of one node on one or more other nodes (see, for example, [23] for a recent

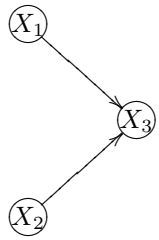
discussion of the distinction with examples, and [24, 25] for early influential papers discussing how functional dependence may be learned from real data). A simple causal relationship between three variables X_1, X_2, X_3 can be represented

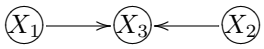


which encodes a *conditional independence relationship* between X_1 and X_3 given X_2, and a factorization of the joint distribution

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2).$$

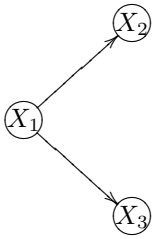
Similarly, the equivalent graphs

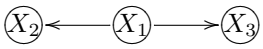




encode $\quad p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$

whereas the graphs for conditional independence of X_2 and X_3 given X_1 are





encoding $\quad p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1).$

(4)

Such simple model assumptions are the building blocks for the construction of highly complex graphical representations of biological systems. There is an important difference between analyses based purely on simultaneous observation of all components of the system, which can typically on yield inference on *dependencies* (say, covariances measured in the joint probability model $p(\boldsymbol{x})$ - see for example [26, 27, 28] - and analyses based on *interventions* - genomic knock-out experiments, chemical or biological challenges, transcriptional/translational peturbation such as RNA interference (RNAi) - that may yield information on casual links; see, for example [29, 30].

2.3. Bayesian Statistical Inference

Given a statistical model for observed data such as equation (3), inference for the parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and the graph structure $\mathcal{G}$ is required. The optimal coherent framework is that of *Bayesian statistical inference* (see for example [31]), that requires computation of the *posterior distribution* for the unknown (or *unobservable*) quantities given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y}) \propto f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) = L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y})p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \tag{5}$$

a probability distribution from which can be computed parameter estimates with associated uncertainties, and predictions from the model. The terms $L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})$ are termed *likelihood* and *prior probability distribution* respectively. The likelihood reflects the observed data, and the prior distribution encapsulates biological prior knowledge about the system under study. If

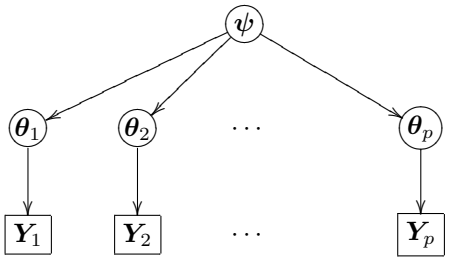
the graph structure is known in advance, the prior distribution for that component can be set to be degenerate. If, as in many cases of probabilistic graphical models, the $\boldsymbol{x}$ are unobserved, then the posterior distribution incorporates them also,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}, \boldsymbol{x}|, \boldsymbol{y}) \propto f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \tag{6}$$

yielding a *latent* or *state-space* model, otherwise interpreted as a *missing data* model.

The likelihood and prior can often be formulated in a hierarchical fashion to reflect believed causal or conditional independence structures. If a graph $\mathcal{G}$ is separable into two subgraphs $\mathcal{G}_1, \mathcal{G}_2$ conditional on a connecting node η, similar to the graph in (4), then the probability model also factorizes into a similar fashion; for example, X_1 might represent the amount of expressed mRNA of a gene that regulates two separate functional modules, and X_2 and X_3 might be the levels of expression of collections of related proteins. The hierarchical specification also extends to parameters in probability models; a standard formulation of a Bayesian hierarchical model involves specification of conditional independence structures at multiple levels of within a graph. The following three-level hierarchical model relates data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ at level 1, to a population of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^T$ at level 2, to hyperparameters $\boldsymbol{\psi}$ at level 3

Level 3 (7)



Level 2

Level 1

yielding the factorization of the Bayesian full joint distribution as

$$f_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\psi},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{\psi}) \left\{ \prod_{i=1}^{p} p(\boldsymbol{\theta}_i|\boldsymbol{\psi}) \right\} \left\{ \prod_{i=1}^{p} p(\boldsymbol{Y}_i|\boldsymbol{\phi}_i) \right\}.$$

2.4. Bayesian Computation

The posterior distribution is, potentially, a high-dimensional multivariate function on a complicated parameter space. The proportionality constant in equation (5) takes the form

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}) = \int f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \, d\boldsymbol{\theta} \, d\boldsymbol{\phi} \, d\mathcal{G} \tag{8}$$

and in equation (6) takes the form

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \, d\boldsymbol{\theta} \, d\boldsymbol{\phi} \, d\mathcal{G} \, d\boldsymbol{x} \tag{9}$$

and is termed the *marginal likelihood* or *prior predictive distribution* for the *observable* quantities $\boldsymbol{x}$ and $\boldsymbol{y}$. In formal Bayesian theory, it is the representation of the distribution of the observable quantities through the paradigm of *exchangeability* that justifies the decomposition in equation (8) into likelihood and prior, and justifies, via asymptotic arguments, the use of the posterior distribution for inference (see [13, Chapters 1-4] for full details). It is evident from these equations that exact computation of the posterior distribution necessitates high-dimensional integration, and in many cases this cannot be carried out analytically.

2.4.1 Numerical Integration Approaches

Classical numerical integration methods, or analytic approximation methods are suitable only in low dimensions. Stochastic numerical integration, for example Monte Carlo integration, approximates expectations by using empirical averages of functionals of samples obtained from the target distribution; for probability distribution $\pi(\boldsymbol{x})$, the approximation of $\mathrm{E}_\pi[g(\boldsymbol{X})]$,

$$\mathrm{E}_\pi[g(\boldsymbol{X})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\, d\boldsymbol{x} < \infty$$

is achieved by randomly sampling $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ (N large) from $\pi()$, and using the estimate

$$\widehat{\mathrm{E}}_\pi[g(\boldsymbol{X})] = \frac{1}{N}\sum_{i=1}^{n} g(\boldsymbol{x}_i).$$

An adaptation of the Monte Carlo method can be used if the functions g and π are not "similar" (in the sense that g is large in magnitude where π is not, and *vice versa*); *importance sampling* uses the representation

$$\mathrm{E}_\pi[g(\boldsymbol{X})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\, d\boldsymbol{x} = \int \frac{g(\boldsymbol{x})\pi(\boldsymbol{x})}{p(\boldsymbol{x})} p(\boldsymbol{x})\, d\boldsymbol{x}$$

for some pdf $p()$ having common support with π, and constructs an estimate from a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from $p()$ of the form

$$\widehat{\mathrm{E}}_\pi[g(\boldsymbol{X})] = \frac{1}{N}\sum_{i=1}^{n} \frac{g(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}.$$

Under standard regularity conditions, the corresponding estimators converge to the required expectation. Further extensions are also useful:

- *Sequential Monte Carlo :* Sequential Monte Carlo (SMC) is an adaptive procedure that constructs a sequence of improving importance sampling distributions. SMC is a technique that is especially useful for inference problems where data are collected sequentially in time, but is also used in standard Monte Carlo problems. See [32].
- *Quasi Monte Carlo :* Quasi Monte Carlo (QMC) utilizes uniform **but not random** samples to approximate the required expectations. It can be shown that QMC can produce estimators with lower variance than standard Monte Carlo.

2.4.2 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a stochastic Monte Carlo method for sampling from a high-dimensional probability distribution $\pi(\boldsymbol{x})$, and using the samples to approximate expectations with respect to that distribution. An ergodic, discrete-time Markov chain is defined on the support of π in such a way that the stationary distribution of the chain exists, and is equal to π. Dependent samples from π are obtained by collecting realized values of the chain after it has reached its stationary phase, and then used as the basis of a Monte Carlo strategy.

The most common MCMC algorithm is known as the *Metropolis-Hastings* algorithm which proceeds as follows. If the state of the d-dimensional chain $\{\boldsymbol{X}_t\}$ at iteration t is given by $\boldsymbol{X}_t = \boldsymbol{u}$, then a candidate state $\boldsymbol{v}$ is generated from conditional density $q(\boldsymbol{u}, \boldsymbol{v}) = q(\boldsymbol{v}|\boldsymbol{u})$, and accepted as the new state of the chain (that is, $\boldsymbol{X}_{t+1} \stackrel{\text{def}}{=} \boldsymbol{v}$) with probability $\alpha(\boldsymbol{u}, \boldsymbol{v})$ given by

$$\alpha(\boldsymbol{u}, \boldsymbol{v}) = \min\left\{1, \frac{\pi(\boldsymbol{v})q(\boldsymbol{v}, \boldsymbol{u})}{\pi(\boldsymbol{u})q(\boldsymbol{u}, \boldsymbol{v})}\right\}.$$

A common MCMC approach involves using a *Gibbs sampler strategy* that performs iterative sampling with updating from the collection of *full conditional* distributions

$$\pi(\boldsymbol{x}_j|\boldsymbol{x}_{(j)}) = \pi(\boldsymbol{x}_j|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{j-1},\boldsymbol{x}_{j+1},\boldsymbol{x}_d) = \frac{\pi(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d)}{\pi(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{j-1},\boldsymbol{x}_{j+1},\boldsymbol{x}_d)} \qquad j = 1,\ldots,d$$

rather than updating the components of $\boldsymbol{x}$ simultaneously. There is a vast literature on MCMC theory and applications; see [33, 34] for comprehensive treatments.

MCMC re-focusses inferential interest from computing posterior analytic functional forms to producing posterior samples. It is an extremely flexible framework for computational inference that carries with it certain well-documented problems, most important amongst them being the assessment of *convergence*. It is not always straightforward to assess when the Markov chain has reached its stationary phase, so certain monitoring steps are usually carried out.

2.5. Bayesian Modelling: Examples

Three models that are especially useful in the modelling of systems biological data are *regression models*, *mixture models state-space models*. Brief details of each type of model follow.

2.5.1 Regression Models

Linear regression models relate an observed response variable Y to a collection of predictor variables $X_1, X_2, \ldots, X_d$ via the model for the ith response

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ij} + \epsilon_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \epsilon_i$$

say, or in vector form, for $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ is a vector of real-valued parameters, and $\boldsymbol{\epsilon}$ is a vector random variable with zero-mean and variance-covariance matrix Σ. The objective in the analysis is to make inference about $\boldsymbol{\beta}$, to understand the influence of the predictors on the response, and to perform prediction for Y. The linear regression model (or General Linear Model) is extremely flexible: the design matrix $\boldsymbol{X}$ can be formed from arbitrary, possibly non-linear *basis* functions of the predictor variables. By introducing a covariance structure into Σ, it is possible to allow for dependence amongst the components of $\boldsymbol{Y}$, and allows for the possibility of modelling repeated measures, longitudinal or time-series data that might arise from multiple observation of the same experimental units.

An extension that is often also useful is to *random effect* or *mixed* models that take into account any repeated measures aspect to the recorded data. If data on an individual (person, sample, gene etc) is $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{id})^T$, then

$$\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U}_i + \boldsymbol{\epsilon}_i \tag{10}$$

where $\boldsymbol{Z}$ is a $d \times p$ constant design matrix, and $\boldsymbol{U}_i$ is a $p \times 1$ vector of *random effects* specific to individual i. Typically the random effect vectors are assumed to be drawn from a common population. Similar formulations can be used to construct semi-parametric models that are useful for flexible modelling in regression.

2.5.2 Mixture Models

A mixture model presumes that the probability distribution of variable $\boldsymbol{Y}$ can be written

$$f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{y}|\theta_k) \tag{11}$$

where $f_1, f_2, \ldots, f_K$ are distinct component densities indexed by parameters $\theta_1, \ldots, \theta_K$, and for all k, $0 < \omega_k < 1$, with

$$\sum_{k=1}^{K} \omega_k = 1.$$

The model can be interpreted as one that specifies that with probability ω_k, $\boldsymbol{Y}$ is drawn from density f_k, for $k = 1, \ldots, K$. Hence the model is suitable for modelling in cluster analysis problems.

This model can be extended to an *infinite mixture model*, which has close links with *Bayesian non-parametric* modelling. A simple infinite mixture/Bayesian non-parametric model is the *mixture of Dirichlet processes* (MDP) model [35, 36]: for parameter $\alpha > 0$ and distribution function F_0, an MDP model can be specified using the following hierarchical specification: for a sample of size n, we have

$$\begin{aligned} Y_i|\theta_i &\sim f_{Y|\theta}(y|\theta_i) \qquad i = 1, \ldots, n \\ \theta_1, \ldots, \theta_n &\sim DP(\alpha, F_0) \end{aligned}$$

where $DP(\alpha, F_0)$ denotes a *Dirichlet process*. The $DP(\alpha, F_0)$ model may be sampled to produce $\theta_1, \theta_2, \ldots, \theta_n$ using the *Polya-Urn* scheme

$$\begin{aligned} \theta_1 &\sim F_0 \\ \theta_k|\theta_1, \ldots, \theta_{k-1} &\sim \frac{\alpha}{\alpha + k - 1} F_0 + \frac{1}{\alpha + k - 1} \sum_{j=1}^{k-1} \delta_{\theta_j} \end{aligned}$$

where δ_x is a point mass at x. For θ_k, conditional on $\theta_1, \ldots, \theta_{k-1}$, the Polya-Urn scheme either samples θ_k from F_0 (with probability $\alpha/(\alpha + k - 1)$), or samples $\theta_k = \theta_j$ for some $j = 1, \ldots, k - 1$ (with probability $1/(\alpha + k - 1)$). This model therefore induces *clustering* amongst the θ values, and hence has a structure similar to the finite mixture model - the distinct values of $\theta_1, \ldots, \theta_n$ are identified as the cluster "centers" that index the component densities in the mixture model in equation (11). The degree of clustering is determined by α; high values of α encourage large numbers of clusters.

The MDP model is a flexible model for statistical inference, and is used in a wide range of applications such as density estimation, cluster analysis, functional data analysis and survival analysis. The component densities can be univariate or multivariate, and the model itself can be used to represent the variability in observed data or as a prior density. Inference for such models is typically carried out using MCMC or SMC methods ([32, 33]). For applications in bioinformatics and functional genomics, see [37, 38].

2.5.3 State-space Models

A state-space model is specified through a pair of equations that relate a collection of *states*, $\boldsymbol{X}_t$, to *observations* $\boldsymbol{Y}_t$ that represent a system and how that system develops over time. For example, the

relationship could be modelled as

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{U}_t) \\ \boldsymbol{X}_{t+1} &= \boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{V}_t) \end{aligned}$$

where $\boldsymbol{f}$ and $\boldsymbol{g}$ are vector-valued functions, and $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ are random error terms. A *linear state-space model* takes the form

$$\begin{aligned} \boldsymbol{Y}_t &= A_t \boldsymbol{X}_t + \boldsymbol{c}_t + \boldsymbol{U}_t \\ \boldsymbol{X}_{t+1} &= B_t \boldsymbol{X}_t + \boldsymbol{d}_t + \boldsymbol{V}_t \end{aligned}$$

for deterministic matrices A_t and B_t and vectors $\boldsymbol{c}_t$ and $\boldsymbol{d}_t$. The $\boldsymbol{X}_t$ represent the values of unobserved states, and the second equation represents the evolution of these states through time. See [39].

State-space models can be used as models for scalar, vector and matrix-valued quantities. One application is evolution of a covariance structure, for example, representing dependencies in a biological network. If the network is dynamically changing through time, a model similar to those above is required but where $\boldsymbol{X}_t$ is a square, positive-definite matrix. For such a network, therefore, a probabilistic model for positive-definite matrices can be constructed from the Wishart/Inverse Wishart distributions [40]. For example, we may have for $t = 1, 2, \ldots$,

$$\begin{aligned} \boldsymbol{Y}_t &\sim \text{Normal}(0, \boldsymbol{X}_t) \\ \boldsymbol{X}_{t+1} &\sim \text{InverseWishart}(\nu_t, \boldsymbol{X}_t) \end{aligned}$$

where degrees of freedom parameter ν_t is chosen to induce desirable properties (stationarity, constant expectation etc.) in the sequence of $\boldsymbol{X}_t$ matrices.

3. Transcriptomics & Functional Genomics

A key objective in the study of biological organization is to understand the mechanisms of the transcription of genomic DNA into mRNA that initiates the production of proteins and hence lies at the centre of the functioning of the nuclear engine. In a cell in a particular tissue at a particular time, the nucleus contains the entire mRNA profile (*transcriptome*) which, if it could be measured, would provide direct insight into the functioning of the cell. If this profile could be measured in a dynamic fashion, then the patterns of gene regulation for one, several or many genes could be studied. Broadly, if a gene is "active" at any time point, it is producing mRNA transcripts, sometimes at a high rate, sometimes at a lower rate, and understanding the relationships between patterns of up- and down-regulation lies at the heart of uncovering pathways, or networks of interacting genes. *Transcriptomics* is the study of the entirety of recorded transcripts for a given genome in a given condition. *Functional genomics*, broadly, is the study of gene function via measured expression levels and how it relates to genome structure and protein expression.

3.1. Microarrays

A common biological problem is to detect *differential expression* levels of a gene in two or more tissue or cell types, as any differences may contribute to the understanding of the cellular organization (pathways, regulatory networks), or may provide a mechanism for discrimination between

future unlabelled samples. An important tool for the analysis of these aspects of gene function is the *microarray*, a medium onto which DNA fragments (or *probes*) are placed or etched. Test sample mRNA fragments are tagged with a fluorescent marker, and then allowed to bond or *hybidize* with the matching DNA probes specific to that nucleotide sequence, according to the usual biochemical bonding process. The microarray thus produces a measurement of the mRNA content of the test sample for each of the large number of DNA sequences bound to the microarray as probes. Microarrays typically now contain tens of thousands of probes for simultaneous investigation of gene expression in whole chromosomes, or even whole genomes for simple organisms. The hybridization experiments are carried out under strict protocols, and every effort is made to regularize the production procedures, from the preparation stage through to imaging. Typically, replicate experiments are carried out.

Microarray experiments have made the study of gene expression routine; instantaneous measurements of mRNA levels for large numbers of different genes can be obtained for different tissue or cell types in a matter of hours. The most important aspects of a statistical analysis of gene expression data are, therefore, twofold; the analysis should be readily implementable for large data sets (large numbers of genes, and/or large numbers of samples), and should give representative, robust and reliable results over a wide range of experiments.

Since their initial use as experimental platforms, microarrays have become increasingly sophisticated allowing measurement of different important functional aspects. Arrays containing whole genomes of organisms can be used for investigation of function, copy-number variation, SNP variation, deletion/insertion sites and other forms of DNA sequence variation (see [41] for a recent summary). High-throughput technologies similar in the form of printed arrays are now at the centre of transcriptomic investigation in several different organisms, and also widely used for genome-wide investigation of common diseases in humans [42, 43]. The statistical analysis of such data represents a major computational challenge. In the list below, a description of details of *first* and *second* generation microarrays is given.

- **First Generation Microarray Studies**

  From the mid 1990s, comparative hybridization experiments using microarrays or gene-chips began to be widely used for the investigation of gene expression. The two principal types of array used were cDNA arrays and oligonucleotide arrays:

  - **cDNA microarrays**: In cDNA microarray competitive hybridization experiments, the mRNA levels of a genes in a target sample are compared to the mRNA level of a control sample by attaching fluorescent tags (usually red and green respectively for the two samples) and measuring the relative fluorescence in the two channels. Thus, in a test sample (containing equal amounts of target and control material), differential expression **relative** to the control is either in terms of *up-regulation* or *down-regulation* of the genes in the target sample. Any genes that are up-regulated in the target compared to the control and hence that have larger amounts of the relevant mRNA, will fluoresce as predominantly red, and any that are down-regulated will fluoresce green. Absence of differences in regulation will give equal amounts of red and green, giving a yellow fluor. Relative expression is measured on the log scale

    $$y = \log \frac{x_{TARGET}}{x_{CONTROL}} = \log \frac{x_R}{x_G} \tag{12}$$

    where x_R and x_G are the fluorescence levels in the RED and GREEN channels respectively.

– **Oligonucleotide arrays:** The basic concept oligonucleotide arrays is that the array is produced to interrogate specific target mRNAs or genes by means of a number of oligo probes usually of length no longer than 25 bases; typically 10-15 probes are used to hybridize to a specific mRNA, with each oligo probe designed to target a specific segment of the mRNA sequence. Hybridization occurs between oligos and test DNA in the usual way. The novel aspect of the oligonucleotide array is the means by which the **absolute** level of the target mRNA is determined; each *perfect match* (PM) probe is paired with a *mismatch* (MM) probe that is identical to the prefect match probe **except** for the nucleotide in the centre of the probe, for which a mismatch nucleotide is substituted, as indicated in the diagram below.

$$\begin{aligned} PM &: ATGTATACTATT\boxed{A}TGCCTAGAGTAC \\ MM &: ATGTATACTATT\boxed{C}TGCCTAGAGTAC \end{aligned}$$

The logic is that the target mRNA, which has been fluorescently tagged, will bind perfectly to the PM oligo, and not bind at all to the MM oligo, and hence the absolute amount of the target mRNA present can be obtained as the difference $x_{PM} - x_{MM}$ where x_{PM} and x_{MM} are the measurements of for the PM and MM oligos respectively.

- **Second Generation Microarrays**

  In the current decade, the number of array platforms has increased greatly. The principle of of hybridization of transcripts to probes on a printed array is often still the fundamental biological component, but the design of the new arrays is often radically different. Some of the new types of array are described below (see [44] for a summary).

  – **ChiP-Chip:** ChIP-chip (*chromatin immunoprecipitation* chip) arrays are *tiling* array with genomic probes systematically covering whole genomes or chromosomes that is used to relate protein expression to DNA sequence by mapping the binding sites of transcription factor and other DNA-binding proteins. See [45] for an application and details of statistical issues.

  – **ArrayCGH :** Array comparative genome hybridization (ArrayCGH) is another form of tiling array that is used to detect *copy number variation* (the variation in the numbers of repeated DNA segments) in subgroups of individuals with the aim of detecting important variations related to common diseases. See [46, 47]

  – **SAGE :** Serial Analysis of Gene Expression (SAGE) is a platform for monitoring the patterns of expression of many thousands of transcripts in one sample, which relies on the sequencing of short cDNA tags that correspond to a sequence near one end of every transcript in a tissue sample. See [48, 49, 50].

  – **Single Molecule Arrays :** Single Molecule Arrays rely on the binding of single mRNA transcripts to the spots on the array surface, and thus allows for extremely precise measurement of transcript levels: see [51]. Similar technology is used for precise protein measurement and antibody detection. See [52]

3.2. Statistical Analysis Of Microarray Data

In a microarray experiment, the experimenter has access to expression/expression profile data, possibly for a number of replicate experiments, for each of a (usually large) number of genes. Conventional statistical analysis techniques and principles (hypothesis testing, significance testing,

estimation, simulation methods/Monte Carlo procedures) are used in the analysis of microarray data. The principal biological objectives of a typical microarray analysis are:

- **Detection of differential expression**: up- or down-regulation of genes in particular experimental contexts, or in particular tissue samples, or cell lines at a given time instant.
- **Understanding of temporal aspects of gene regulation**: the representation and modelling of patterns of changes in gene regulation over time.
- **Discovery of gene clusters:** the partitioning of large sets of genes into smaller sets that have common patterns of regulation.
- **Inference for gene networks/biological pathways:** the analysis of co-regulation of genes, and inference about the biological processes involving many genes concurrently.

There are typically several key issues and models that arise in the analysis of microarray data: such methods are described in detail in [53, 54, 55, 56]. For a Bayesian modelling perspective, see [57].

- **array normalization:** arrays are often imaged under slightly different experimental conditions, and therefore the data are often very different even from replicate to replicate. This is a systematic experimental effect, and therefore needs to be adjusted for in the analysis of differential expression. A misdiagnosis of differential expression may be made purely due to this systematic experimental effect.
- **measurement error**: the reported (relative) gene expression levels models are only in fact proxies for the true level gene expression in the sample. This requires a further level of variability to be incorporated into the model.
- **random effects modelling**: it may be necessary to use *mixed* regression models, where gene specific *random-effects* terms are incorporated into the model.
- **multivariate analysis**: the covariability of response measurements, in time course experiments, or between PM and MM measurements for an oligonucleotide array experiment, is best handled using multivariate modelling.
- **testing:** one- and two-sample hypothesis testing techniques, based on parametric and non-parametric testing procedures can be used in the assessment of the presence of differential expression. For detecting more complex (patterns of) differential expression, in more general structured models, the tools of analysis of variance (ANOVA) can be used to identify the chief sources of variability.
- **multiple testing/False discovery**: in microarray analysis, a classical statistical analysis using significance testing needs to take into account the fact that a very large number of tests are carried out. Hence significance levels of tests must be chosen to maintain a required *family-wise error rate*, and to control the *false discovery rate***.**
- **classification**: the genetic information contained in a gene expression profile derived from microarray experiments for, say, an individual tissue or tumour type may be sufficient to enable the construction of a *classification rule* that will enable subsequent classification of new tissue or tumour samples.

- **cluster analysis:** discovery of subsets of sets of genes that have common patterns of regulation can be achieved using the statistical techniques of *cluster analysis* (see section §3.3.).
- **computer-intensive inference**: for many testing and estimation procedures needed for microarray data analysis, simulation-based methods (bootstrap estimation, Monte Carlo and permutation tests, Monte Carlo and MCMC) are often necessary, especially when complex Bayesian models are used.
- **data compression/feature extraction:** the methods of principal components analysis and extended linear modelling via *basis functions* can be used to extract the most pertinent features of the large microarray data sets.
- **experimental design**: statistical experimental design can assist in determining the number of replicates, the number of samples, the choice of time points at which the array data are collected and many other aspects of microarray experiments. In addition, power and sample size assessments can inform the experimenter as to the statistical worth of the microarray experiments that have been carried out.

Typically, data derived from both types of microarray highly noise and artefact corrupted. The statistical analysis of such data is therefore quite a challenging process. In many cases, the replicate experiments are very variable. The other main difficulty that arises in the statistical analysis of microarray data is the dimensionality; a vast number of gene expression measurements are available, usually only on a relatively small number of individual observations or samples, and thus it is hard to establish any general distributional models for the expression of a single gene.

3.3. Clustering

Cluster analysis is an unsupervised statistical procedure that aims to establish the presence of identifiable subgroups (or *clusters*) in the data, so that objects belonging to the same cluster resemble each other more closely than objects in different clusters; see [58, 59] for comprehensive summaries.

In two or three dimensions, clusters can be visualized by plotting the raw data. With more than three dimensions, or in the case of dissimilarity data (see below), analytical assistance is needed. Broadly, clustering algorithms fall into two categories:

- **Partitioning Algorithms :** A partitioning algorithm divides the data set into K clusters, where and the algorithm is run for a range of K -values. Partitioning methods are based on specifying an initial number of groups, and iteratively reallocating observations between groups until some equilibrium is attained. The most famous algorithm is the K-*Means* algorithm in which the observations are iteratively classified as belonging to one of K groups, with group membership is determined by calculating the centroid for each group (the multidimensional version of the mean) and assigning each observation to the group with the closest centroid. The K-means algorithm alternates between calculating the centroids based on the current group memberships, and reassigning observations to groups based on the new centroids. A more robust method uses mediods rather than centroids (that is, medians rather than means in each dimension, and more generally, any distance-based allocation algorithm could be used.

- **Hierarchical Algorithms :** A hierarchical algorithm yields an entire hierarchy of clusterings for the given data set. *Agglomerative methods* start with each object in the data set in its own cluster, and then successively merges clusters until only one large cluster remains. *Divisive methods* start by considering the whole data set as one cluster, and then splits up clusters until each object is separated. Hierarchical algorithms are discussed in detail in section §3.3.1.

Data sets for clustering of N observations can either take the form of an $N \times p$ data matrix, where rows contain the different observations, and columns contain the different variables, or an $N \times N$ dissimilarity matrix, whose $(i,j)^{th}$ element is d_{ij}, the distance or dissimilarity between observations i and j that obeys the usual properties of a metric. Typical data distance measures between two data points i and j with measurement vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are the L_1 and L_2 Euclidean distances, and the grid-based *Manhattan distance* for discrete variables, or the *Hamming distance* for binary variables. For *ordinal* (ordered categorical) or *nominal* (label) data, other dissimilarities can be defined.

3.3.1 Hierarchical Clustering

Agglomerative hierarchical clustering initially places each of the N items in its own cluster. At the first level, two objects are to be clustered together, and the pair is selected such that the potential function increases by the largest amount, leaving $N-1$ clusters, one with two members, the remaining $N-2$ each with one. At the next level, the optimal configuration of $N-2$ clusters is found, by joining two of the existing clusters. This process continuous until a single cluster remains containing all N items. At each level of the hierarchy, the merger chosen is the one that leads to the smallest increase in some objective function.

Classical versions of the hierarchical agglomeration algorithm are typically used with *average*, *single* or *complete* linkage methods, depending on the nature of the merging mechanism. Such criteria are inherently heuristic, and more formal *model-based* criteria can also be used. Model-based clustering is based on the assumption that the data are generated by a mixture of underlying probability distributions. Specifically, it is assumed that the population of interest consists of K different subpopulations, and that the density of an observation from the the subpopulation is for some unknown vector of parameters. Model-based clustering is described in more detail in section §3.3.2.

The principal display plot for a clustering analysis is the *dendrogram* which plots all of the individual data objects linked by means of a binary "tree". The dendrogram represents the structure inferred from a hierarchical clustering procedure which can be used to partition the data into subgroups as required if it is cut at a certain "height" up the tree structure. As with many of the aspects of the clustering procedures described above, it is more of a heuristic graphical representation rather than formal inferential summary. However, the dendrogram is readily interpretable, and favoured by biologists.

3.3.2 Model-Based Hierarchical Clustering

Another approach to hierarchical clustering is *model-based clustering* (see for example [60, 61]), which is based on the assumption that the data are generated by a mixture of K underlying probability distributions as in equation (11). Given data matrix $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_N)^T$, let $\gamma = (\gamma_1, ..., \gamma_N)$ denote

the cluster labels, where $\gamma_i = k$ if the i^{th} data point comes from the k^{th} subpopulation. In the classification procedure, the maximum likelihood procedure is used to choose the parameters in the model.

Commonly, the assumption is made that the data in the different subpopulations follow multivariate normal distributions, with mean $\boldsymbol{\mu}_k$ and covariance matrix Σ_k for cluster k, so that

$$f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{y}|\boldsymbol{\mu}_k, \Sigma_k) = \sum_{k=1}^{K} \omega_k \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_k)\right\}$$

where $\Pr[\gamma_i = k] = \omega_k$. If $\Sigma_k = \sigma^2 \boldsymbol{I}_p$ is a $p \times p$ matrix, then maximizing the likelihood is the same as minimizing the sum of within-group sums of squares and corresponds to the case of hyperspherical clusters with the same variance. Other forms of Σ_k yield clustering methods that are appropriate in different situations. The key to specifying this is the singular value or eigen decomposition of Σ_k, given by eigenvalues $\lambda_1, ..., \lambda_p$ and eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_p$, as in Principal Components Analysis **(author?)** [62]. The eigenvectors of Σ_k, specify the orientation of the k^{th} cluster, the largest eigenvalue λ_1 specifies its variance or size, and the ratios of the other eigenvalues to the largest one specify its shape. Further, if $\Sigma_k = \sigma_k^2 \boldsymbol{I}_p$, the criterion corresponds to hyperspherical clusters of different sizes, and by fixing the eigenvalue ratios $\alpha_j = \lambda_j/\lambda_1$ for $j = 2, 3, ..., p$ across clusters, other cluster shapes are encouraged.

3.4. Model-Based Analysis Of Gene Expression Profiles

The clustering problem for vector-valued observations can be formulated using models used to represent the gene expression patterns via the *extended linear model*, that is, a linear model in non-linear basis functions; see, for example, [63, 64] for details.

Generically, the aim of a statistical model is to capture the behaviour of the gene expression ratio y_t as a function of time t. The basis of the modelling strategy would be to use models that capture the characteristic behaviour of expression profiles likely to be observed due to different forms of regulation. A regression framework and model can be adopted. Suppose that Y_t is modelled using a linear model

$$Y_t = \boldsymbol{X}_t \boldsymbol{\beta} + \varepsilon_t$$

where $\boldsymbol{X}_t$ is (in general) a $1 \times p$ vector of specified functions of t, and $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector. In vector representation, the gene expression profile over times $t_1, \ldots, t_T$ can be written $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \tag{13}$$

The precise form of design matrix $\boldsymbol{X}$ will be specified to model the time-variation in signal. Typically the random error terms $\{\varepsilon_t\}$ are taken as independent and identically distributed Normal random variables with variance σ^2, implying that the conditional distribution of the responses $\boldsymbol{Y}$ is multivariate normal

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_T\right) \tag{14}$$

where now $\boldsymbol{X}$ is $T \times p$ where $\boldsymbol{I}_T$ is the $T \times T$ identity matrix.

In order to characterize the underlying gene expression profile, the parameter vector $\boldsymbol{\beta}$ must be estimated. For this model, the maximum likelihood/ordinary least squares estimates of $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}}_{ML} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad \widehat{\sigma}^2 = \frac{1}{T-p}\left(\boldsymbol{y}-\widehat{\boldsymbol{y}}\right)^T\left(\boldsymbol{y}-\widehat{\boldsymbol{y}}\right)$$

for fitted values $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{ML} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

3.4.1 Bayesian Analysis In Model-Based Clustering

In a Bayesian analysis of the model in (13) a joint prior distribution $\pi\left(\boldsymbol{\beta}, \sigma^2\right)$ is specified for $\left(\boldsymbol{\beta}, \sigma^2\right)$, and a posterior distribution conditional on the observed data is computed for the parameters. The calculation proceeds using equation (5) (essentially with $\mathcal{G}$ fixed).

$$\pi\left(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}\right) = \frac{L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}, \sigma^2\right)}{\int L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}, \sigma^2\right)\, d\boldsymbol{\beta}\, d\sigma^2}$$

where $L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)$ is the likelihood function. In the linear model context, a conjugate prior specification is used where

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \equiv \text{Normal}\left(\boldsymbol{v}, \sigma^2\boldsymbol{V}\right) \qquad \pi\left(\sigma^2\right) \equiv \text{IGamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right) \tag{15}$$

($\boldsymbol{v}$ is $p \times 1$, $\boldsymbol{V}$ is $p \times p$ positive-definite and symmetric, all other parameters are scalars) and IGamma denotes the inverse Gamma distribution. Using this prior, standard Bayesian calculations show that conditional on the data

$$\pi\left(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2\right) \equiv \text{Normal}\left(\boldsymbol{v}^*, \sigma^2\boldsymbol{V}^*\right) \qquad \pi\left(\sigma^2|\boldsymbol{y}\right) \equiv \text{IGamma}\left(\frac{T+\alpha}{2}, \frac{c+\gamma}{2}\right) \tag{16}$$

where

$$\begin{aligned} \boldsymbol{V}^* &= \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1} \qquad \boldsymbol{v}^* = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right) \\ c &= \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{v}^T\boldsymbol{V}^{-1}\boldsymbol{v} - \left(\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right)^T\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right) \end{aligned} \tag{17}$$

In regression modelling, it is usual to consider a centered parameterization for $\boldsymbol{\beta}$ so that $\boldsymbol{v} = 0$, giving

$$\begin{aligned} \boldsymbol{v}^* &= \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ c &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}^T\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{y}^T\left(\boldsymbol{I}_T - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^T\right)\boldsymbol{y} \end{aligned}$$

A critical quantity in a Bayesian clustering procedure is the marginal likelihood, as in equation (8), for the data in light of the model.

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int f_{\boldsymbol{Y}|\boldsymbol{\beta},\sigma^2}\left(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}|\sigma^2\right)\pi\left(\sigma^2\right) d\boldsymbol{\beta} d\sigma^2. \tag{18}$$

Combining terms above gives that

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \left(\frac{1}{\pi}\right)^{T/2}\frac{\gamma^{\alpha/2}\Gamma\left(\frac{T+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\frac{|\boldsymbol{V}^*|^{1/2}}{|\boldsymbol{V}|^{1/2}}\frac{1}{\{c+\gamma\}^{(T+\alpha)/2}} \tag{19}$$

This expression is the marginal likelihood for a single gene expression profile. For a collection of profiles belonging to a single cluster, $\boldsymbol{y}_1, ..., \boldsymbol{y}_N$, equation (19) can again be evaluated and used as the basis of a dissimilarity measure as an input into a hierarchical clustering procedure. The marginal likelihood in equation (19) can easily be re-expressed for clustered data. The basis of the hierarchical clustering method outlined in [64] proceeds by agglomeration of clusters from N to 1, with the two clusters that lead to the **greatest increase** marginal likelihood score at each stage of the hierarchy. This method works for profiles of arbitrary length, potentially with different observation time points, however it is computationally most efficient when the time points are the same for each profile.

The design matrix $\boldsymbol{X}$ is typically expressed via non-linear basis functions, for example truncated polynomial splines, Fourier bases or wavelets. For T large, it is usually necessary to use a projection through a lower number of bases; for example, for a single profile, $\boldsymbol{X}$ becomes $T \times p$ and $\boldsymbol{\beta}$ becomes $p \times 1$, for $T > p$. Using different designs, many flexible models for the expression profiles can be fitted. In some cases, the linear mixed effect formulation in equation (10) can be used to construct the spline-based models; in such models, some of the β parameters are themselves assumed to be random effects. See [65].

For example, in *harmonic regression*, regression in the Fourier bases is carried out. Consider the extended linear model

$$Y_t = \sum_{j=0}^{p} \beta_j g_j(t) + \varepsilon_t$$

where $g_0(t) = 1$ and

$$g_j(t) = \begin{cases} \cos(\phi_j t) & j \text{ odd} \\ \sin(\phi_j t) & j \text{ even} \end{cases}$$

where p is an even number, $p = 2k$ say, and $\phi_j, j = 1, 2, \ldots, k$ are constants with $\phi_1 < \phi_2 < \cdots < \phi_k$. For fixed t, $\cos(\phi_j t)$ and $\sin(\phi_j x)$ are also fixed and this model is a linear model in parameters

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

This model can be readily fitted to time-course expression profiles. The plot below is a fit of the model with $k = 2$ to a cluster of profiles extracted using the method described in [64] from the malaria protozoa *Plasmodium falciparum* data set described in [66].

One major advantage of the Bayesian inferential approach is that any biological prior knowledge that is available can be incorporated in a coherent fashion. For example, the data in Figure 1 illustrate periodic behaviour to the cyclical nature of cellular organization, and thus the choice of the Fourier bases is a natural one.

3.4.2 Choosing The Number Of Clusters: Bayesian Information Criterion

A hierarchical clustering procedure gives the sequence by which the clusters are merged (in agglomerative clustering) or split (in divisive clustering) according the model or distance measure used, but does not give an indication for the number of clusters that are present in the data (under the model specification). This is obviously an important consideration. One advantage of the model-based approach to clustering is that it allows the use of statistical model assessment procedures to assist in the choice of the number of clusters. A common method is to use approximate *Bayes factors* to compare models of different orders (i.e. models with different numbers of clusters), and gives a systematic means of selecting the parameterization of the model, the clustering method, and also the number of clusters. See [67].

The Bayes factor is the posterior odds for one model against the other assuming neither is favored *a priori*. A reliable approximation to twice the log Bayes factor called the *Bayesian Information Criterion* (BIC), which, for model M fitted to n data points is given by

$$\text{BIC}_M = -2 \log L_M(\widehat{\boldsymbol{\theta}}) + d_M \log n$$

where L_M is the Bayesian marginal likelihood from equation (18), $L_M(\widehat{\boldsymbol{\theta}})$ is the maximized log likelihood of the data for the model M, and d_M is the number of parameters estimated in the model.

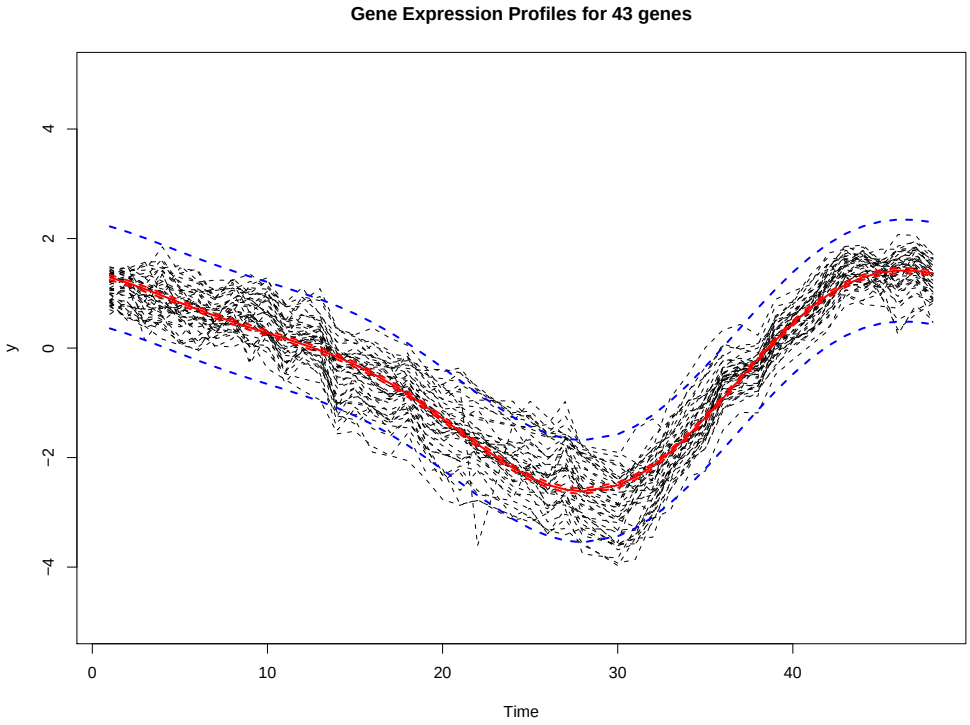


Figure 1: Cluster of gene expression profiles obtained using Bayesian hierarchical model-based clustering: data from the intraerythrocytic developmental cycle of protozoa *Plasmodium falciparum*. Clustering achieved using harmonic regression model with $k = 2$. Solid red line is posterior mean for this cluster, dotted red lines are pointwise 95 % credible intervals for the cluster mean profile, and dotted blue lines are pointwise 95 % credible intervals for the observations.

The number of clusters is not considered a parameter for the purposes of computing the BIC. The smaller (more negative) the value of the BIC, the stronger the evidence for the model.

3.4.3 Classification Via Model-Based Clustering

Any clustering procedure can be used as the first step in the construction of *classification* rules. Suppose that it, on the basis of an appropriate decision procedure, it is known that there are C clusters, and that a set of existing expression profiles $y_1, ..., y_N$ have been allocated in turn to the clusters. Let $z_1, ..., z_N$ be the cluster allocation labels for the profiles. Now, suppose further that the C clusters can be decomposed further into two subsets of sizes C_0 and C_1, where the subsets represent perhaps clusters having some common, known biological function or genomic origin. For example, in a cDNA microarray, it might be known that the clones are distinguishable in terms of the organism from which they were derived. A new objective could be to allocate a novel gene and expression profile to one of the subsets, and one of the clusters within that subset. ,

Let y_{ijk} denote, for $i = 0, 1$, $j = 1, 2, ..., C_i$, $k = 1, 2, ..., N_{ij}$ denote the kth profile in cluster j in subset i. Let y^* denote a new profile to be classified, and ξ^* be the binary classification-to-subset, and z^* the classification-to-cluster variable for y^*. Then, by Bayes Rule, for $i = 1, 2$,

$$P\left[\xi^* = i | y^*, y, z\right] \propto p\left(y^* | \xi^* = i, y, z\right) P\left[\xi^* = i | y, z\right] \tag{20}$$

The two terms in (20) can be determined on the basis of the clustering output.

4. Metabolomics

The term *metabolome* refers to the total metabolite content of an organic sample (tissue, blood, urine etc) obtained from a living organism which represents the products of a higher level of biological interaction than that which occurs within the cell. *Metabolomics* and *metabonomics* are the fields in biomedical investigation that combines the application of nuclear magnetic resonance (NMR) *spectroscopy* with multivariate statistical analysis in studies of the composition of the samples. Metabonomics is often used in reference to the static chemical content of the sample, whereas metabolomics is used to refer to the dynamic evolution of the metabolome. Both involve the measurement of the metabolic response to interventions - see for example [68] - and applications of metabolomics include several in public health and medicine [69, 70].

4.1. Statistical Methods for Spectral Data

The two principal spectroscopic measurement platforms, NMR and Mass Spectrometry (MS) yield alternative representations of the metabolic spectrum. They produce spectra (or profiles) that consist of several thousands of individual measurements at different resonances or masses. There are several phases of processing of such data; pre-processing using smoothing, alignment and de-noising, peak separation, registration and signal extraction. For an extensive discussion, see [62].

An NMR spectrum consists of measurements of the intensity or frequency of different biochemical compounds (metabolites) represented by a set of *resonances* dependent upon the chemical structure, and can be regarded as a linear combination of peaks (nominally of various widths) that correspond to singletons or multiple peaks according to the neighbouring chemical environment. A typical spectrum extracted from rat urine is depicted in Figure 2; see [71]. Two dominant sharp peaks are visible.

Features of the spectra that require specific statistical modelling include multiple peaks for a single compound, variation in peak shape, and chemical shifts induced by variation in experimental pH Signals from different metabolites can be highly overlapped and subject to peak position variation due primarily to pH variations in the samples, and there are many small scale features (see Figure 3). Statistical methods of pre-processing NMR spectra for statistical analysis which address the problems outlined above, using, for example, dynamic time warping to achieve alignment of resonance peaks across replicate spectra as a form or spectral registration form part of the necessary holistic Bayesian framework.

Classical statistical methods for metabolic spectra include the following:

- **Principal Components Analysis (PCA) and Regression**: a linear data projection method for dimension reduction, feature extraction, and classification of samples in an unsupervised fashion, that is, without reference to labelled cases.
- **Partial Least Squares (PLS)**: a non-linear projection method similar to PCA, but implemented in a supervised setting for sample discrimination.

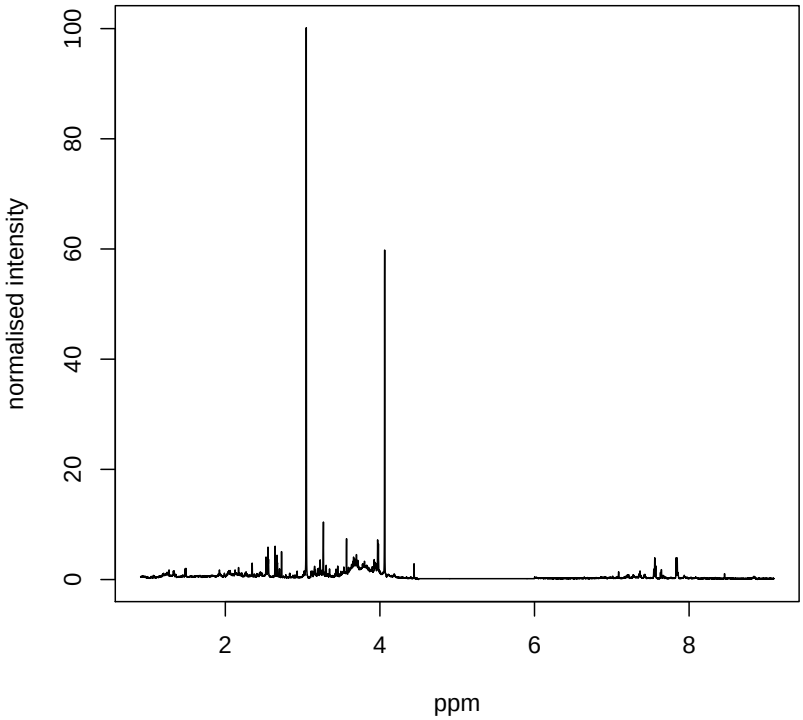


Figure 2: A normalised rat urine spectrum. The ordinate is parts per million, the abscissa is intensity after standardization.

- **Clustering** Clusters of spectra, or peaks within spectra, can be discovered using similar techniques to those described in section §3.3..
- **Neural Networks:** Flexible non-linear regression models constructed from simple mathematical functions that are learned from the observation of cases, that are ideal models for classification. The formulation of an neural involves three levels of interlinked variables; *outputs*, *inputs*, and *hidden variables*, interpreted as a collection of unobserved random variables that form the hidden link between inputs and outputs.

4.2. Bayesian Approaches

The Bayesian framework is a natural one for incorporating genuine biological prior knowledge into the signal reconstruction, and typically useful prior information (about fluid composition, peak location, peak multiplicity) is available. In addition, a hierarchical Bayesian model structure naturally allows construction of plausible models for the spectra across experiments or individuals.

- **Flexible Bayesian Models** The NMR spectrum can be represented as a noisy signal derived from some underlying and biologically important mechanism. Basis-function approaches (specifically, wavelets) have been much used to represent non-stationary time-varying signals [65, 72, 73, 71]. The sparse representation of the NMR spectrum in terms of wavelet coefficients makes them an excellent tool in data compression, yet these coefficients can still be easily transformed back to the spectral domain to give a natural interpretation in terms of the underlying metabolites.

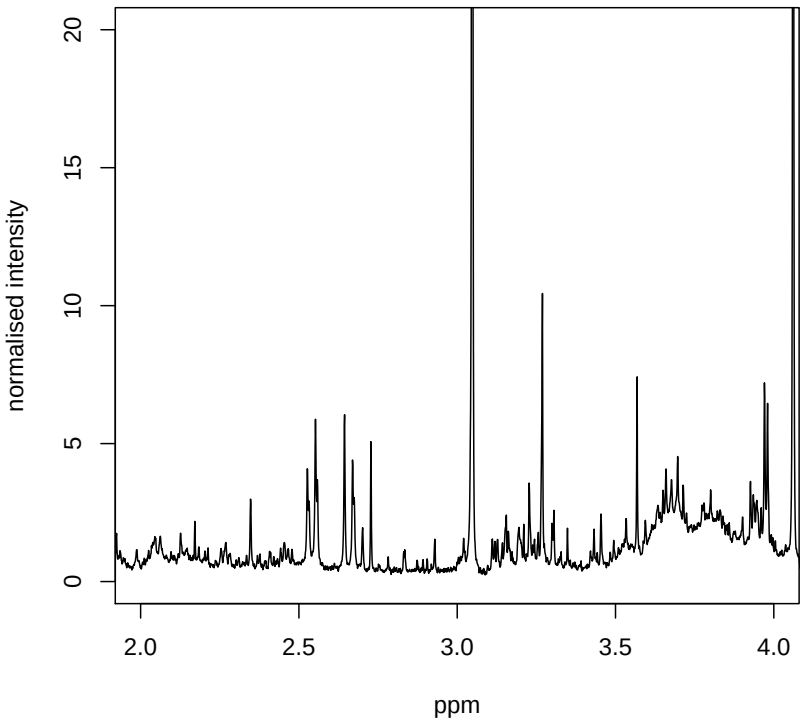


Figure 3: Magnified portion of the spectrum showing small scale features.

Figure 4 depicts the reconstruction of the rat urine spectrum in the region between 2.5 and 2.8 ppm using wavelet methods; see [71].

- **Bayesian Time Series Models for Complex Non-Stationary Signals:** see for example [74]. The duality between semi-parametric modelling of functions and latent time series models allows a view of the analysis of the underlying NMR spectrum not as a set of pointwise evaluations of a function, but rather as a (time-ordered) series of correlated observations with some identifiable latent structure. Time series models, computed using dynamic calculation (filtering), provide a method for representing the NMR spectra parsimoniously.

- **Bayesian Mixture Models:** A reasonable generative model for the spectra is one that constructs the spectra from a large number of symmetric peaks of varying size, corresponding to the contributions of different biochemical compounds. This can be approximated using a finite mixture model, where the number, magnitudes and locations, of the spectral contributions are unknown. Much recent research has focussed on the implementation of computational strategies for Bayesian mixtures, in particular Markov chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) have proved vital. The reconstruction of NMR spectra is a considerably more challenging area than those for which mixture modelling is conventionally used, as many more individual components are required. Flexible semi-parametric mixture models have been utilized in [75, 76], whilst fully non-parametric mixture models similar to those described in section §2.5.2 can also be used [73].

A major advantage of using the fully Bayesian framework is that, once again, all relevant information (the spectral data itself, knowledge of the measurement processes for different experimental platforms, the mechanisms via which multiple peaks and shifts are introduced) can be integrated in

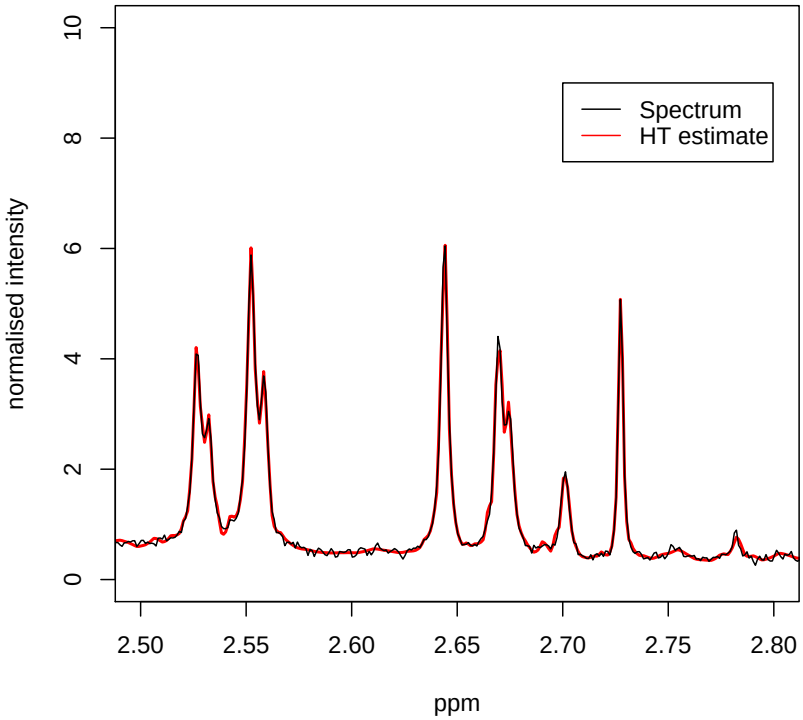


Figure 4: Wavelet reconstruction of a region of the spectrum results under the "Least Asymmetric wavelet" with four vanishing moments using the hard thresholding (HT).

a coherent fashion. In addition, prior knowledge about the chemical composition of the samples can be integrated via a prior distribution constructed by inspection of the profiles for training samples. At a higher level of synthesis, the Bayesian paradigm offers a method for integrating metabolomic data with other functional or structural data, such as gene expression or protein expression data. Finally, the metabolic content of tissue changes temporally, so dynamic modelling of the spectra could also be attempted.

5. Future Directions

Biological data relating to structure and function of genes, proteins and other biological substances are now available from a wide variety of platforms. Researchers are beginning to develop methods for coherent combination of data from different experimental processes to get an entire picture of biological cause and effect. For example, the effective combination of gene expression and metabonomic data will be of tremendous utility. A principal challenge is therefore the fusion of expression data derived from different experimental platforms, and seeking links with sequence and ontological information available. Such fusion will be critical in the future of statistical analysis of large scale systems biology and bioinformatics data sets.

In terms of public health impact of systems biology and statistical genomics, perhaps the most prominent is the study of common diseases through high-throughput genotyping of single nucleotide polymorphisms (SNPs). In genomewide association studies, SNP locations that correlate with disease status or quantitative trait value are sought. In such studies, the key statistical step

involves the selection of a informative predictors (SNP or genomic loci) from a large collection of candidates. Many such genomewide studies have been completed or are ongoing (see [77, 78, 42, 43]). Such studies represent huge challenges for statisticians and mathematical modellers, as the data contain many subtle structures but also as the amount of information is much greater than that available for typical statistical analyses.

Another major challenge to the quantitative analysis of biological data comes in the form of image analysis and extraction. Many high throughput technologies rely on the extraction of information from images, either in static form, or dynamically from a series of images. For example it is now possible to track the expression level of mRNA transcripts in real-time ([79, 80, 81]), and to observe mRNA transcripts moving from transcription sites to translation sites (see for example [82]). Imaging techniques can also offer insights into aspects of the dynamic organization of nuclear function by studying the positioning of nuclear compartments and how those compartments reposition themselves in relation to each other through time. The challenges for the statistician are to develop real-time analysis methods for tracking and quantifying the nature and content of such images, and tools from spatial modelling and time series analysis will be required.

Finally, *flow cytometry* can measure characteristics of millions of cells simultaneously, and is a technology that offers many promises for insights into biological organization and public health implications. However, quantitative measurement and analysis methods are only yet in the early stages of development, but offer much promise (see [83, 84]).

References

[1] H Kitano, editor. *Foundations of Systems Biology*. MIT Press, Cambridge, MA, USA, 2001.

[2] H Kitano. Computational systems biology. *Nature*, 420(6912):206–210, November 2002.

[3] U Alon. *An Introduction to Systems Biology*. Chapman and Hall (CRC), Boca Raton, FL, USA, 2006.

[4] A W F Edwards. *Foundations of mathematical genetics*. Cambridge University Press, Cambridge, UK, 2nd edition, 2000.

[5] GU Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London, Series B*, 213:21–87, 1924.

[6] R A Fisher. On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341, 1922.

[7] RA Fisher. *The genetical theory of natural selection.* Clarendon Press, Oxford, 1930.

[8] S Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[9] JD Murray. *Mathematical Biology: I. An Introduction*. Springer-Verlag, 2002.

[10] JD Murray. *Mathematical Biology: II. Spatial Models and Biomedical Applications*. Springer-Verlag, 2003.

[11] B Lewin. *Genes*. Jones & Bartlett Publishers, Boston, MA, USA, 9th edition, 2007.

[12] DL Spector. Nuclear domains. *Journal of Cell Science*, 114(16):2891–3, 2001.

[13] JM Bernardo and AFM Smith. *Bayesian Theory*. John Wiley & Sons, New York, NY, USA, 1994.

[14] JW Haefner, editor. *Modeling Biological Systems: Principles and Applications*. Springer, 2nd edition, 2005.

[15] JO Ramsay, G Hooker, D Campbell, and J Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Methodology)*, 69(5):741–796, 2007.

[16] S Donnet and A Samson. Estimation of parameters in incomplete data models defined by dynamical systems. *Journal of Statistical Planning and Inference*, 137(9):2815–2831, 2007.

[17] S Rogers, R Khanin, and M Girolami. Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, 8(Suppl 2.):doi:10.1186/1471–2105–8–S2–S2, 2007.

[18] DJ Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall (CRC), Boca Raton, FL, USA, 2006.

[19] EA Heron, B Finkenstädt, and DA Rand. Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. *Bioinformatics*, 23(19):2596–2603, 2007.

[20] EM Airoldi. Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3(12), 2007.

[21] D Husmeier, R Dybowski, and S Roberts, editors. *Probabilistic Modelling in Bioinformatics and Medical Informatics*. Springer, 2005.

[22] N Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

[23] R Opgen-Rhein and K Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1:37:1–10, 2007.

[24] AJ Butte, P Tamayo, D Slonim, TR Golub, and IS Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences of the USA.*, 97:12182–12186, 2000.

[25] N Friedman, M Linial, I Nachman, and D Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[26] M West, C Blanchette, H Dressman, E Huang, S Ishida, R Spang, H Zuzan, JA Olson Jr, JR Marks, and JR Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the USA.*, 98(11):462–467, 2001.

[27] A Dobra, C Hans, B Jones, J Nevins, G Yao, and M West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

[28] B Jones, C Carvalho, A Dobra, C Hans, C Carter, and M West. Experiments in stochastic computation for high dimensional graphical models. *Statistical Science*, 20:388–400, 2005.

[29] F Markowetz, J. Bloch, and R. Spang. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21:4026–4032, 2005.

[30] D Eaton and KP Murphy. Exact Bayesian structure learning from uncertain interventions. *Artificial Intelligence & Statistics*, 2007.

[31] CP Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Texts in Statistics. Springer, 2nd edition, 2007.

[32] A Doucet, N de Freitas, and NJ Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York, 2001.

[33] CP Robert and G Casella. *Monte Carlo Statistical Methods*. Texts in Statistics. Springer, 2nd edition, 2005.

[34] D Gamerman and HF Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. Chapman & Hall (CRC), 2nd edition, 2006.

[35] CE Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174, 1974.

[36] M D Escobar and M West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[37] DB Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. In KA Do, P Müller, and M Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 10. Cambridge University Press, Cambridge, UK, 2006.

[38] S Kim, MG Tadesse, and M Vannucci. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893, 2006.

[39] M. West and J. Harrison. *Bayesian Forecasting and Dynamic models*. Springer New York, 2nd edition, 1999.

[40] A Philipov and ME Glickman. Multivariate stochastic volatility via Wishart processes. *Journal of Business and Economic Statistics*, 24(3):313–328, 2006.

[41] D Gresham, MJ Dunham, and D Botstein. Comparing whole genomes using DNA microarrays. *Nature Reviews Genetics*, 9:291–302, 2008.

[42] Wellcome Trust Case Control Consortium. Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. *Nature Genetics*, 39:1329 – 1337, 2007.

[43] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[44] XS Liu. Getting started in tiling microarray analysis. *PloS Computational Biology*, 3(10):1842–1844, 2007.

[45] WE Johnson, W Li, CA Meyer, R Gottardo, JS Carroll, M Brown, and XS Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the USA.*, 103(33):12457–62, 2006.

[46] JL Freeman et al. Copy number variation: New insights in genome diversity. *Genome Res.*, 16(8):949–961, 2006.

[47] AE Urban et al. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA.*, 103(12):4534–4539, 2006.

[48] S Saha et al. Using the transcriptome to annotate the genome. *Nature Biotechnology*, 20:508–512, 2002.

[49] A Shadeo et al. Comprehensive serial analysis of gene expression of the cervical transcriptome. *BMC Genomics*, 8:142, 2007.

[50] SJ Robinson, JD Guenther, CT Lewis, MG Links, and IA Parkin. Reaping the benefits of SAGE. *Methods Molecular Biology*, 406:365–386, 2007.

[51] C Lu, SS Tej, S Luo, CD Haudenschild, BC Meyers, and PJ Green. Elucidation of the Small RNA Component of the Transcriptome. *Science*, 309(5740):1567–1569, 2005.

[52] H Weiner, J Glökler, C Hultschig, K Bssow, and G Walter. Protein, antibody and small molecule microarrays. In UR Müller and DV Nicolau, editors, *Microarray Technology and Its Applications*, Biological and Medical Physics, Biomedical Engineering, pages 279–295. Springer, Berlin, 2006.

[53] TP Speed, editor. *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall (CRC), Boca Raton, FL, USA, 2003.

[54] G Parmigiani, ES Garett, RA Irizarry, and SL Zeger, editors. *The Analysis of Gene Expression Data*. Statistics for Biology and Health. Springer, 2003.

[55] E Wit and J McClure. *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley, Chichester, UK, 2004.

[56] R Gentleman, V Carey, W Huber, R Irizarry, and S Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health. Springer, 2005.

[57] KA Do, P Müller, and Vannucci M. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, 2006.

[58] BS Everitt, S Landau, and M Leese. *Cluster Analysis*. Hodder Arnold, London, UK, 4th edition, 2001.

[59] L Kaufman and PJ Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, New York, NY, USA, 2nd edition, 2005.

[60] KY Yeung, C Fraley, A Murua, AE Raftery, and WL Ruzzo. Model-based clustering and data transformation for gene expression data. *Bioinformatics*, 17:977–987, 2001.

[61] GJ McLachlan, RW Bean, and D Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(413–422), 2002.

[62] M De Iorio, TMD Ebbels, and DA Stephens. Statistical techniques in metabolic profiling. In DJ Balding, M Bishop, and C Cannings, editors, *Handbook of Statistical Genetics*, chapter 11. John Wiley, Chichester, UK, 3rd edition, 2007.

[63] N A Heard, C C Holmes, D A Stephens, D J Hand, and G Dimopoulos. Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *Proceedings of the National Academy of Sciences of the United States*, 102(47):16939–16944, 2005.

[64] N A Heard, C C Holmes, and D A Stephens. 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*, 101(473):18–29, 2006.

[65] JS Morris, PJ Brown, KA Baggerly, and KR Coombes. Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. In KA Do, P Müller, and M Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 14, pages 269–292. Cambridge University Press, 2006.

[66] Z Bozdech, M Llinás, BL Pulliam, ED Wong, J Zhu, and DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biology*, 1(1):E5, 2003.

[67] RE Kass and AE Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(773–795), 1995.

[68] J K Nicholson, J Connelly, J C Lindon, and E Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 1:153–161, 2002.

[69] J C Lindon, J K Nicholson, E Holmes, H Antti, M E Bollard, H Keun, O Beckonert, T M Ebbels, M D Reily, and D Robertson. Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicology and Applied Pharmacology*, 187:137, 2003.

[70] J T Brindle, H Annti, E Holmes, G Tranter, J K Nicholson, H W L Bethell, S Clarke, S M Schofield, E McKilligin, D E Mosedale, and Graingerand D J. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nature Medicine*, 8:143, 2002.

[71] T-J Yen, TMD Ebbels, M De Iorio, DA Stephens, and S Richardson. Analysing real urine spectra with wavelet methods. In preparation, 2008.

[72] P J Brown, T Fearn, and M. Vannucci. Bayesian wavelet regression on curves with applications to a spectroscopic calibration problem. *Journal of the American Statistical Society*, 96:398–408, 2001.

[73] M A Clyde, L L House, and R L Wolpert. Nonparametric models for proteomic peak identification and quantification. In KA Do, P Müller, and M Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 15, pages 293–308. Cambridge University Press, 2006.

[74] M West, R Prado, and A Krystal. Evaluation and comparison of EEG traces: Latent structure in non-stationary time series. *Journal of the American Statistical Association*, 94:1083–1095, 1999.

[75] S Ghosh, DF Grant, DK Dey, and DW Hill. A semiparametric modeling approach for the development of metabonomic profile and bio-marker discovery. *BMC Bioinformatics*, 9:38, 2008.

[76] S Ghosh and DK Dey. A unified modeling framework for metabonomic profile development and covariate selection for acute trauma subjects. *Statistics in Medicine*, 30(27(19)):3776–88, 2008.

[77] RH Duerr et al. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science*, 314(5804):1461–1463, 2006.

[78] R Sladek et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[79] D Longo and J Hasty. Imaging gene expression: tiny signals make a big noise. *Nature Chemical Biology*, 2:181–182, 2006.

[80] D Longo and J Hasty. Dynamics of single-cell gene expression. *Molecular Systems Biology*, 2:64, 2006.

[81] AL Wells, JS Condeelis, RH Singer, and D Zenklusen. Imaging real-time gene expression in living systems with single-transcript resolution: Image analysis of single mRNA transcripts. *CSH Protocols*, 2007.

[82] AJ Rodriguez, JS Condeelis, RH Singer, and JB Dictenberg. Imaging mRNA movement from transcription sites to translation sites. *Seminars in Cell & Developmental Biology*, 18(2):202–208, 2007.

[83] G Lizard. Flow cytometry analyses and bioinformatics: Interest in new softwares to optimize novel technologies and to favor the emergence of innovative concepts in cell research. *Cytometry A*, 71A:646–647, 2007.

[84] K Lo, RR Brinkman, and R Gottardo. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A*, 73A(4):321–332, 2008.