Gilles Celeux, Sylvia Früwirth-Schnatter, and Christian P. Robert

Handbook of Mixture Analysis

Contents

Mixture notations

a	a_g^j most frequent response level in cluster g for a categorical variable j
b	
c	shape parameter for (inverted) Gamma or (inverted) Wishart distribution
c_0	shape parameter for a prior
c_g	shape parameter for posterior of the g-th component
c_n (or c_T)	shape parameter for posterior, if independent of g
c_g^X, $c_g^Y, \ldots$	distinguish between the posterior of X and Y
d	dimension of data space
e	Neperian constant
e	exposures
e_i	exposures associated with case i
f	$f_g(\cdot\|\cdot)$ = density of g-th component
g	index for groups/components in a finite (or infinite) mixture
h	index for states in a finite (or infinite) Markov mixture
i	index for cases/observations
j	index for variables
k	index for models
l	index for the categorical variable levels
m	index for MCMC iterations
$\theta^{(m)}$	m-th draw for parameter θ
n	number of cases/observations
n_g	number of cases/observations in g-th group
o	
p	probability density function (see P&S)
q	
r	dimension of the component-specific parameter θ_g
s	index for EM algorithm iteration
t	index for time
u	
u_j	column vector of loading matrix U
v	disturbances (state space models, etc)
v_t	disturbances at time t

w	regressor
w_i	row vector of regressors for i-th case
w_t	row vector of regressors at time t
w_{it}	row vector of regressors for i-th case at time t
x	independent observations
$\mathbf{x}$	data collection: $\mathbf{x} = (x_1, \ldots, x_n)$
x_i	row vector of regressors for i-th case
x_t	row vector of regressors at time t
x_{it}	row vector of regressors for i-th case at time t
y	data/outcome/dependent observation
$\mathbf{y}$	data collection: $\mathbf{y} = (y_1, \ldots, y_n)$
y_i	data (vector) for i-th case
y_t	data (vector) at time t
y^t	data (vector) up to time t; i.e. $y^t = (y_1, \ldots, y_t)$
y_{it}	data (vector) for i-th case at time t
z	latent indicator
$\mathbf{z}$	data collection: $\mathbf{z} = (z_1, \ldots, z_n)$
z_i	group to which i-th case belongs (takes values in $1, \ldots, G$)
$z_{ig} = 1$	if case $i \in$ group g, 0 otherwise
$\hat{z}_{ig}$	Estimate of z_{ig} at the MLE using the MAP operator
z^*_{ig}	estimate of z_{ig} from the CEM algorithm (= 0 or 1)

A	shape matrix
A_g	shape matrix for g-th component
B	
B_g	shape matrix for g-th component when the matrix of eigenvector is the identity matrix
C	scale parameter for a Gamma, inverted Gamma, Wishart, inverted Wishart distribution
C_0	scale parameter for a prior
C_g	scale parameter for posterior of the g-th component
C_n (or C_T)	scale parameter for posterior, if independent of g
$C_g^X, C_g^Y, \ldots$	distinguish between the posterior of X and Y
D	
D_g	matrix of eigenvectors for g-th component

E	expectation (see P&S)
F	
G	number of groups in a finite mixture
H	number of states in a hidden Markov model
I	state of a hidden Markov chain (takes values in $1, \ldots, H$)
I_t	state at time t
J	
K	number of models
L_j	number of categorical levels of variable j
M	number of MCMC iterations
M_0	burn-in of a MCMC chain
N	counts for hidden states
N_{gh}	number of transition from g to h
O	
P	probability (see P&S)
Q	covariance matrix in a random effects model
R	loss function in a Bayesian decision problem
T	number of time series observations
T_i	number of repeated measurements/observations per case (unit)
V	variance (see P&S)
W	regressor/design matrix in a regression model
X	regressor matrix in a regression model
Y	random variable
Z	

L	likelihood for mixture model
L_c	complete-data likelihood
L_o	observed-data likelihood
ℓ	log likelihood for mixture model
ℓ_c	complete-data log likelihood
ℓ_o	observed-data log likelihood
$\mathcal{M}_k$	k-th model considered

α	parameter vector of the multinonial dist. for the latent class model with categorical variables
β	regression parameter
β_g	regression parameter in group g
β_t	regression parameter at time t in a time varying model
β_i^s	individual regression parameter in random effects model
γ	regression parameters in MNL model for mixtures-of-experts model
γ_g	regression parameter in group g
δ	model indicator in a (Bayesian) variable selection problem
ϵ	vector of the scattering parameters for the latent class model with categorical variables
ε	error term in a regression model
ε_i	error term for case i
ε_t	error term at time t
ε_{it}	error term for case i at time t
ζ	
η	weight distribution in finite (or infinite) mixture
η_g	weight of the g-th mixture component
θ	vector of all parameters of a model
θ_g	parameter vector of g-th mixture component (not including η_g)
ι	
κ	
λ	
λ_g	Exponential mixtures: intensity parameter for g-th group
λ_g	Poisson mixtures: relative risk for g-th group
λ_g	Gaussian mixtures: volume parameter for g-th group
λ_{gj}	j-th eigenvalue of Σ_g
μ	mean/expectation of a random variable
μ_g	mean vector for g-th component (Gaussian, Poisson, Student-t, etc)
μ_t	time-varying level in a state space model
ν	degrees of freedom parameter (Student-t, χ^2-distribution)
ν_g	degrees of freedom parameter for the g-th component in a Student-t mixture
ξ	transition matrix of the hidden Markov chain in a hidden Markov model
ξ_{gh}	probability to move from state g to state h
π	$\pi = 3.14159$
π	success probability
π_g	binomial mixtures: success probability in the g-th component
π_g^j	binary latent class model: success probability for variable j in g-th component
π_{lg}^j	latent class model: success probability for category l of variable j in g-th component

ρ	permutation
ρ_g	correlation coefficient for the g-th component
σ	notation for the standard deviation of a univariate random variable
σ^2	notation for the variance of a univariate random variable
σ_g^2	Gaussian mixtures: variance for g-th component mixtures of regression model: variance of the error term for the g-th regression model
τ	
τ_{ig}	conditional probability that $z_{ig} = 1$
υ	
ϕ	$\phi_g(\cdot\|\cdot,\cdot)$ = pdf of Normal distribution for g-th component
χ	χ^2-distribution
ψ	
ω	scale parameter in data augmentation models (e.g. student-t distribution)
ω_i	scale parameter for i-th case/observation
ω_t	scale parameter at time t

Γ	
Δ	
Θ	parameter space
Λ	
Ξ	
Π	
Σ	covariance/scale matrix of a multivariate random variable
Σ_g	Gaussian mixtures: covariance matrix for g-th component Student-t mixtures: scale matrix for g-th component
Φ	AR coefficients
Φ_g	Markov switching model: AR coefficients for g-th state
$\Phi_g(\cdot\|\cdot,\cdot)$	cdf of the Normal distribution for g-th component
Ψ	
Ω	covariance of a multivariate random variable
Ω_g	covariance for g-th component, if different from Σ_g (e.g. Student-t, factor model)

Special notation Linear Algebra

determinate of a matrix A	$\|A\|$
diagonal matrix	$\operatorname{Diag}(d_1, \ldots, d_k)$
identity matrix	I
trace of a matrix A	$\operatorname{tr}(A)$
transpose of a matrix A	$A^\top$
vector of ones	$\mathbf{1}$

Special notation Probability & Statistics

Bernoulli distribution	$\mathcal{B}er(p)$
Beta distribution	$\mathcal{B}e(a,b)$
Density of the Beta distribution	$f_{\mathcal{B}e}(y\|a,b)$
Binomial distribution	$\mathcal{B}(n,p)$
χ^2-distribution	χ^2_ν
Dirichlet distribution	$\mathcal{D}(e_1,\dots,e_G)$
Dirichlet process	$\mathcal{DP}(\alpha,H_0)$
Exponential distribution	$\mathcal{E}(\lambda)$
Gamma distribution	$\mathcal{G}(a,b)$ (with expectation a/b)
Density of the Gamma distribution	$f_{\mathcal{G}}(y\|a,b)$
inverted Gamma distribution	$\mathcal{IG}(a,b)$
Multinomial distribution	$\mathcal{M}(M,p_1,\dots,p_G)$
Normal distribution	$\mathcal{N}(\mu,\sigma^2)$
d-variate Normal distribution	$\mathcal{N}_d(\mu,\Sigma)$
pdf of standard Normal distribution	$\phi(\cdot)$
cdf of standard Normal distribution	$\Phi(\cdot)$
pdf of Normal distribution	$\phi(\cdot\|\cdot,\cdot)$
cdf of Normal distribution	$\Phi(\cdot\|\cdot,\cdot)$
Poisson distribution	$\mathcal{P}(\mu)$
Negative Binomial	$\mathcal{NB}(n,p)$
Student-t distribution	$t_\nu(\mu,\sigma^2)$
Uniform distribution	$\mathcal{U}[0,1]$
Wishart distribution	$\mathcal{W}(c,C)$
inverted Wishart distribution	$\mathcal{IW}(c,C)$
y	
prior density	$p(\cdot)$
posterior densities	$p(\cdot\|\cdot)$
probability of an event A	$\mathrm{P}(A)$
expectation of r.v. Y	$\mathrm{E}(Y)$
variance of r.v. Y	$\mathrm{V}(Y)$
variance-covariance matrix of a r.v. Y	$\mathrm{Cov}(Y)$
correlation between two r.v. X and Y	$\mathrm{Corr}(X,Y)$
sample mean	$\overline{y}$
sample variance (divided by n)	s^2_y
sample covariance matrix	S_y

Further mathematical notations

Beta function	$B(\alpha, \beta)$
Gamma function	$\Gamma(\alpha)$
indicator function for event	$\mathbb{I}(x = a)$
indicator function	$\mathbb{I}_a(x)$

1

Applications in Genomics

Stephane Robin and Christophe Ambroise

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, France

Mixture models are intensively used in genetics and genomics either for identifying latent structures or for modeling densities. According the type of mixture component and the nature of the hypothesis about latent structures, mixture models may be relevant in numerous different frameworks.

In this chapter, use of mixture models in genetic and genomic are presented by increasing complexity of the latent structure. The first section considers applications with independent latent variable structures to genome and transcriptome analysis. The second section illustrates the use of Hidden Markov Models (HMM) in genomics, presenting a variety of problems with their associated translation in terms of emission distributions and hidden states. Eventually the last section introduces more complex dependency structures used in genomics such as Hidden Markov Random Field (HMRF) or Stochastic Block Model (SBM) with their associated parameter estimation difficulties.

1.1 Mixture Models in Transcriptome and Genome Analysis

1.1.1 Analyzing the genetic structure of a population

Identifying the underlying structure of populations is a recurrent task in genetics. It allows to correct population stratification in genetic association studies (13) or to study the evolutionary relationships between populations as well as to learn about their demographic histories (16).

In this context, mixture models emerge as a natural strategy to infer the structure of the population or the structure of the genome itself. Indeed there is a two level structure: each individual can be considered as belonging to a subpopulation but regions of the genome of a given individual can themselves be considered as having different origins. This last case is known as genetic admixture. It occurs when individuals from two or more previously separated populations begin interbreeding.

Many different parametric approaches exist, differing mainly in the estimation method but relying on the same basic mixture of multinomials. The so called Structure algorithm (64) proposes a mixture of multinomial distributions in a Bayesian framework with MCMC inference. FRAPPE (76) uses a maximum-likelihood approach associated with an EM algorithm and Admixture (4) computes the same estimates using a sequential quadratic programming algorithm with a quasi-Newton scheme. We also mention the Bayesian Analysis of Population Structure (BAPS) (18) which includes the number of subpopulations in the model.

Let us describe the reference Structure model. Structure uses Bayesian statistical inference to cluster individuals from genotype data or to determine admixture proportions (64). Different statistical models are associated with each endgame of the method. Both BAPS and Structure models assume the estimated subpopulations to be in Hardy-Weinberg equilibrium. The model without admixture assumes G subpopulations from which were sampled the n diploid individuals genotyped at L multi-allelic (but often bi-allelic) loci $(y_\ell^{i,1}, y_\ell^{i,2})_{\ell=1,\dots,L,i=1,\dots,n}$. The parameters of the mixture are the allele frequencies $\theta = (\theta_{gj\ell})$ where $\theta_{gj\ell} = p(y_\ell^{(i,a)} = j|z^i = g, \theta)$ is the frequency of allele ℓ at locus j in population g. The conditional distribution for individual i is

$$p(y^i = (y_1^i, \cdots, y_L^i)|z^i = g) = \prod_\ell \prod_{a=1,2} \theta_{g,y_l^{i,a},\ell}.$$

This resulting mixture has numerous parameters since each locus (often of the order of the million) of each class is described by a different vector of proportions. Bayesian inference is used to obtain the distribution of $p(z, \theta|x)$. In the original paper (64) the posterior distribution

$$p(z, \theta|y) \propto p(z)p(\theta)p(y|z, \theta),$$

considers a uniform prior for z and a Dirichlet prior for θ.

To account for admixture, a new parameter is introduced in the model. The parameter $\pi = (\pi_{ig})$ where π_{ig} represents the proportion of individual i's that originated from population g:

$$p(z^i = (z_1^i, \cdots, z_L^i)|\pi_i) = \prod_\ell \prod_{a=1,2} \pi^i_{z_\ell^{(i,a)}},$$

where $z_\ell^{(i,a)}$ is the population of origin of allele copy $y_\ell^{(i,a)}$. The conditional distribution then becomes

$$p(y^i|z^i, \pi^i) = \prod_\ell \prod_a \theta_{z_\ell^{(i,a)}, y_\ell^{(i,a)}, \ell}.$$

For each individual, the parameter π_i has a Dirichlet prior distribution as well. The estimated posterior distribution becomes $p(z, \pi, \theta|y)$.

Because of the high-dimensionnality of the problem, this type of inference can be really slow and variational inference offers a faster alternative in the Bayesian framework (66).

Other approaches may also consider mixture models associated with kernel methods. In that context the use of adapted kernel function enable to operate in a high-dimensional, implicit feature space via Gaussian Mixture Models. SHIPS (Spectral Hierarchical Clustering for the Inference of Population Structure in Genetic Studies) is an avatar of such approaches (14). It is based on a spectral clustering algorithm. The algorithm first uses a kernel based on the allele sharing distance (ASD) that has been previously used to identify genetic patterns among populations (54). In the implicit feature space, classical GMM are used to recursively split the individuals in subpopulations.

1.1.2 Finding sets of co-transcribed genes

Gene expression is modulated (up or down) depending on tissue (e.g., liver vs. brain), development stage (e.g., fetal vs. adult), disease status, genotype (e.g. mutant vs. wild) or dynamically as a response to environmental signals.

A DNA microarray consists of thousands of microscopic spots of DNA oligonucleotides from a specific DNA sequence, known as probes (or reporters) that are used to hybridize a cRNA sample (called target). The DNA microarray technology allows to measure the expression levels of thousands of genes across different conditions. These measurements provide a ”picture” of cells functioning at a given time. Such technology is thus of great importance in many applications such as functional genomic, medical and clinical diagnosis, drug discovery, targeting and monitoring . . .

The data resulting from this type of experiments is a gene expression matrix X whose columns describe the genes and rows describe the samples. Notice that each sample is also described by other variables such as the conditions of the experiment. Most of the time reasonable experiments involve a set of replicates for each condition. Gene expression data analysis aims at pointing to differences between conditions and giving insight into global gene patterns.

Such data has many features, few observations and is most of the time very noisy. Statistical analysis of microarray raise issues and challenges for statisticians. Analyzing such data usually requires a succession of steps (49): a normalization process to make all samples comparable, a differential analysis to pinpoint genes which have different expression across the conditions, and an exploratory data analysis step to enhance the understanding of the results. The analysis of microarray data relies on univariate and multivariate descriptive statistics at each step in order to control the process or gain insight into the data. Clustering approaches are often used and mixture model is a classical tool in this context (37). When dealing with DNA microarray, data is considered continuous and Gaussian mixture is the dominant model.

In gene expression analysis, clustering aims at finding a structure within the samples or/and within the genes. Both approaches bring different and relevant information about the data. The seminal paper of (23) proposes to cluster the genes by mean of a classical hierarchical agglomerative clustering using average-linkage and an initial metric based on euclidian distance. The author observe that genes of similar function cluster together. This observation justifies the use of clustering for searching hints about gene function guided by a ”guilty by association” principle.

When considering the clustering of samples, high dimension (i.e. the number of genes) represents a problem. Indeed, classical mixture models are not able to deal with samples where the number of variables is much greater than the number of samples, unless the variables are assumed to be uncorrelated within a cluster. This problem is caused by the estimation of the covariance matrices whose number of parameters grows quadratically with the number of variables. (82) exploit the representation of the covariance matrix in terms of its eigenvalue decomposition:

$$\Sigma_g = \lambda_g D_g A_g D_g^\top,$$

where D_g is the matrix of eigenvector, A_g is a diagonal matrix whose elements are proportional to the eigenvalues and λ_g a scalar. Reduction of the number of free parameters in the covariance matrix can also be achieved by mixture of factor analyzers (50) :

$$\Sigma_g = B_g B_g^\top + D_g,$$

where B_g is a matrix of loading factors and D_g a diagonal matrix.

Since the mid-2000, Next Generation Sequencing (NGS) has become the new standard tool for measuring gene expression. Compared to data produced with microarrays, NGS data are count-based measures, discrete, positive, and highly skewed. We introduce the main two alternatives that have been proposed to adapt model-based clustering approaches to such data: it is possible to change the normalization of the data to adapt to Gaussian mixture (40; 34) or develop model specifically for discrete positive skewed data.

Although a multivariate version of the Poisson distribution does exist, (68) assume variables to be independent conditionally on the components. Considering discrete gene expressions $y_{ij\ell}$ of gene i in condition j for replicate ℓ, the component distribution of the expression vector of gene i is a product of Poisson distributions:

$$y_i|g \sim \prod_{j=1}^{d}\prod_{\ell=1}^{r_j}\mathcal{P}(y_{ij\ell};\mu_{ij\ell g}).$$

The authors propose to further reduce the number of parameters using parametrization in the spirit of above mentioned Gaussian parametrization, assuming a common mean across the replicates

$$\mu_{ij\ell g} = w_i\lambda_{jg},$$

or adapting the mean in function of known library sizes $s_{j\ell}$

$$\mu_{ij\ell g} = w_i s_{j\ell}\lambda_{jg},$$

since the number of reads mapped to a gene is highly dependent on the gene size.

Biclustering is a technique in two way data analysis, which aims at finding a structure of both rows and columns of a data table. This approach is popular for exploring DNA microarray since there is often a structure both in samples and genes. Looking for a gene/sample block structure can obviously be achieved in two steps (one step for each dimension) or can be searched simultaneously in both dimension (8). A widespread graphical representation of this approach is the classical `heatmap` which is an false color image of the data table with reordering of the rows and columns according to some identified latent structure.

There are many different types of structure and algorithms in the field of biclustering. Block structure is a possibility and can be considered as a latent structure. So called Latent Block Model (LBM, mixture model with associated estimation procedure) have been proposed in this context by (30) for identifying a simultaneous partition of rows and columns. The density of y knowing the partition of the rows z and the partition of columns w is

$$p(y|z,w,\alpha) = \prod_{ijg\ell}\psi(y_{ij};\alpha_{g\ell})^{z_{ig}w_{j\ell}},$$

where ψ is a parametric distribution of parameter vector $\alpha_{g\ell}$ for block $g\ell$. In that context the likelihood is not tractable and variational EM (see Section 1.3.3) or Bayesian strategies have to been proposed for estimating the parameters of the mixture. Although model selection is a complex problem in this context since the likelihood is not tractable, Bayesian inference offers efficient alternative for designing criterion (38).

1.1.3 Variable selection for clustering with GMM

As mentioned above, Gaussian mixture models are not identifiable in high dimension setting (when the number of observations is small compared to the number of variables) and strategies have been developed for limiting the number of parameters of the model. Biological data typically falls into this high-dimension setting not only because of transcriptome but also with epigenome, proteome, metabolome, molecular pathways, molecular imaging, *etc.* Dimension reduction is thus a key issue in the field. It can be achieved via factor analysis, regularization other sort of constrained parametrization (15). But a simple alternative for reducing the dimensionality consists in selecting relevant variables. Variable selection is an important and old topic in supervised learning but has a more recent history in clustering. The difficulty of the problem plays undoubtedly a role in this difference of treatment. Nevertheless variable selection in discrimination (41) and clustering share common aspects. Variable selection may be performed in different ways. The so-called filter approach consists in selecting ”informative” variables beforehand. In the context of transcriptome the most widespread method is differential analysis (see Section 1.1.4). A second possibility consists in selecting (or ordering) the variables after clustering (if clustering is possible) (50). The last possibility consists selecting the variables while estimating the mixture model parameters.

In a Bayesian paradigm (75) propose a method using a reversible jump algorithm to simultaneously choose the number of mixture components and select relevant variables. The paper of (65) defines two different sets of variables: relevant and irrelevant variables. They do not assume independence between the relevant and irrelevant variables for the clustering, as considered in (75). The model of (65) considers a partition of the variables into two main subsets: the variables relevant for clustering S, and the irrelevant ones S_c. The model skilfully mixes clustering and regression. The integrated likelihood is decomposed into two multiplicative parts

$$p(y|g) = p_{clust}(y^S|g)p_{reg}(y^{S_c}|y^S),$$

$p_{clust}(y^S|g)$ being a classical mixture model and $p_{reg}(y^{S_c}|y^S)$ a multivariate regression model where the variable in S_c are explained as linear combination of the variable in S. (47) propose a refined version of this model avoiding non-parsimonious models by selecting the predictor variables in the linear regression part of the model in a two steps stepwise algorithm.

In supervised learning, penalized regression represents a popular approach for selecting variable while estimating the parameters. The same kind of approach has been explored in the context of mixture model. (59) propose for example a penalized log-likelihood criterion by assuming a Gaussian mixture model with common diagonal covariance matrices. The LASSO like penalty penalizes the sum of the absolute values of the component j of cluster means g:

$$p_\lambda(\theta) = \lambda \sum_{gj} |\mu_{gj}|.$$

There are many related works in this line of research borrowing and adapting the idea developed for supervised methods (81).

1.1.4 Mixture models in the specific case of multiple testing

Because of the dimensionality of most genomic data, multiple testing issues have become a common place in genomic analyses. The most emblematic case is this of the detection of differentially expressed genes, which can be summarized as follows. Consider all (known) genes from a given species and, for each of them, perform a sample comparison test with null hypothesis

$$H_{0i} = \text{'gene } i \text{ has the same expression level in all conditions'}.$$

In a frequentist setting, each gene is then associated with a test statistics, the distribution of which is known under H_{0i} and from which a p-value y_i is derived. Such a setting raises obviously a multiple testing problem about which a huge literature exists (see (19; 72) for reviews). Multiple testing procedures aim at controlling some multiple type I error rate such as family-wise error rate (FWER), false discovery rate (FDR), *etc.* Most of these procedures rely on the fact that, under H_{0i}, y_i has a uniform distribution $\mathcal{U}[0, 1]$.

Unsupervised classification point of view.

Efron et al. (21) rephrased this problem as a clustering problem, in which one wishes to classify genes according to the latent variable z_i defined as follows:

$$z_i = \begin{cases} 0 & \text{if } H_{0i} \text{ is true} \quad \text{('null' gene)}, \\ 1 & \text{if } H_{0i} \text{ is not} \quad \text{(differentially expressed gene)}. \end{cases}$$

The unsupervised classification task can then be achieved using the mixture model

$$y_i \sim \eta_0 f_0 + (1 - \eta_0) f_1 \tag{1.1}$$

where η_0 stands for the proportion of null genes, f_0 for the pdf of the uniform distribution $\mathcal{U}[0, 1]$ and f_1 for the pdf of the p-values under the alternative hypothesis H_{1i}, which is supposed to be common to all genes. This model provides an alternative view of the problem, in which the (estimated) conditional probability $\tau_{i0} = \eta_0 f_0(y_1)/[\eta_0 f_0(y_i) + (1 - \eta_0) f_1(y_i)]$ is interpreted as a *local* FDR (22). A natural classification rule then consists of classifying gene i as positive (i.e. non 'null') when τ_{i0} is below a given threshold t. Note that, in this setting, because f_0 is known, the proportion η_0 of null genes can be estimated under mild conditions on f_1 (see e.g. (74)).

Parametric mixture models.

A first parametric version of Model (1.1) was proposed by (5), were f_1 is supposed to be a Beta distribution $\mathcal{B}e(a, b)$. In the same vein, (48) considered a two-group mixture model similar to Model (1.1), but applied to the transformed $\widetilde{y}_i = \Phi^{-1}(y_i)$, so that the transformed null distribution $\widetilde{f}_0$ is a standard Gaussian distribution. In this approach, the alternative distribution $\widetilde{f}_1$ is supposed to be Gaussian $\mathcal{N}(\mu_1, \sigma_1^2)$. The probit transform Φ^{-1} turns out to be efficient, as it zooms into the region where p-values are close to 0, which improves the identification of the positive genes (see Figure 1.1). (48) further elaborate on the estimation of the FDR and suggest to consider the estimate

$$\widehat{FDR}(t) = \sum_i \tau_{i0} I_0(\tau_{i0} < t) \Big/ \sum_i I_0(\tau_{i0} < t) \ .$$

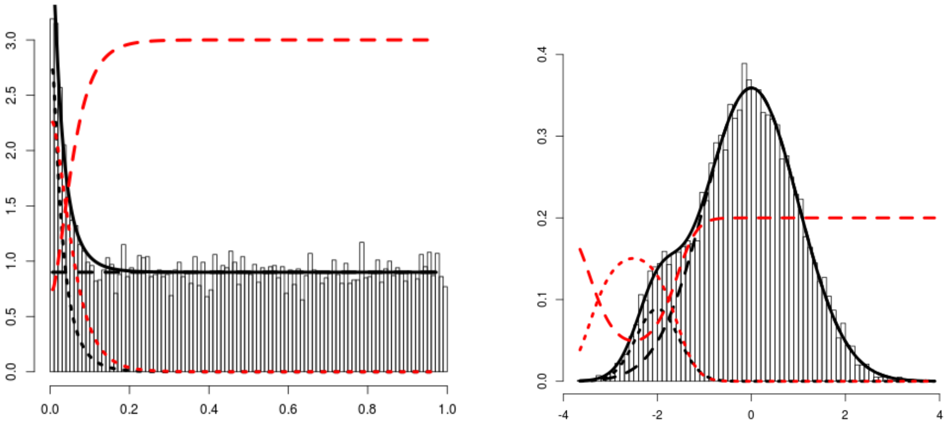


FIGURE 1.1
Left: Histograms of p-values with fitted Beta mixture (black lines, dashed=f_0, dotted=f_1) and rescaled conditional probabilities (gray lines). Right: Histograms of probit-transformed p-values with fitted Gaussian mixture, same legend.

Semi- and non-parametric mixture models.

One of the key feature in Model (1.1) is that the distribution f_0 is known, which allows for a higher flexibility for f_1. (5) propose a 'semi-parametric' extension of it, taking f_1 as a mixture of Beta distributions itself. Equation (1.1) then takes the standard form $y_i \sim \sum_{g=0}^{G} \eta_g f_g$ but, in terms of classification, one is only interested in the distinction between group 0 and the union of all other groups, so gene i is classified according to $\tau_{i+} := \sum_{g=1}^{G} \tau_{ig} = 1 - \tau_{i0}$.

In a non-parametric setting, Bordes at al. (12) prove that Model (1.1) is identifiable provided that $f_1(\cdot|\theta_1) = \psi(\cdot - \theta_1)$ where ψ is some even function. They consider both a symmetrization-based and a moment estimate for the location parameter θ_1. (70) consider a kernel density estimate of f_1, for the estimation of which a convergent algorithm is proposed, provided that η_0 is known.

1.2 Hidden Markov Models in Genomics: Some Specificities

Because many genomic data are collected at loci (probe, nucleotide) located along the genome, hidden Markov models (HMM) have become a standard tool in that field (20; 73). We remind that an HMM deals with data collected in a sequential manner and is similar to a mixture model, except that the $\{z_t\}_{1 \leq t \leq n}$ form a Markov chain on $\{1, \dots, G\}$ with transition matrix Π. The group z_t to which the observation collected at locus t belongs to is called its (hidden) *state*. In most HMM, the observations $\{y_t\}_{1 \leq t \leq n}$ are supposed to be independent conditionally on the hidden states, the conditional distribution being named the *emission* distribution. In this section, we first present a typical genomic problem where HMM can be used and then we introduce a series of special case where the genomic context requires to consider more sophisticated models in terms of hidden states, emission distribution and dependency structure.

1.2.1 A typical case: Copy number variations

Many diseases are associated with genomic alterations which consist of either the loss or the amplification of some regions of the genome (2). As a result, some regions of the genome are not present in two copies (as expected in a normal cell of a diploid species, such as human), but in less (zero or one copy, named 'loss') or in more (three, four or even more copies, named 'gain'). A series of technologies have be developed in the last decades to get a measure y_t that is related to the number of genomic copies at locus t. These technologies range from micro-arrays to sequencing technologies (NGS) (3). A typical example of the signal at hand is displayed in the left panel of Figure 1.2.

Hidden Markov models are especially well suited to address the task of both finding the location at which the number of copies changes and to classify each of the segments according to the number of copies or simply as 'normal', 'loss' or 'gain'. (28) first proposed to use an HMM with Gaussian emission for the detection of CNV based on micro-array data. The classification step, which consists of retrieving the hidden path (z_t) is referred to as 'CNV calling' and can be achieved using the Viterbi algorithm (77).

The detection of the loci where the copy number varies can by seen as a change-point detection problem, as proposed by (61), but this approach does not address the calling step. As a complement, (62) introduced a mixture model where each *region* belongs to a certain group g, the data y_t being independent with same distribution ϕ_g. The inference

Is it a Gaussian distribution? If not replaces ϕ_g by f_g

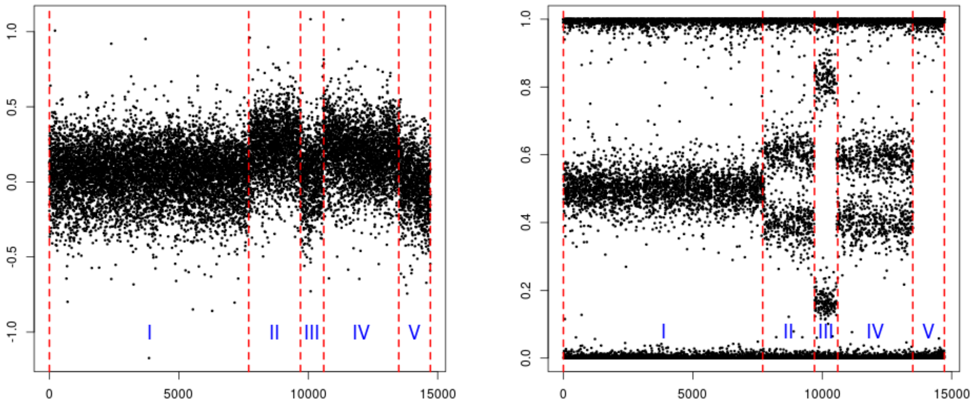


FIGURE 1.2
SNP data from (63), chromosome 11. Left: Log R ratio $\log_2[(a_t + b_t)/2]$ as a function of t. Right: B-allele frequency $= b_t/(a_t + b_t)$ as a function of t.

of such a model can be made using an specific EM algorithm where the M-step includes a segmentation step that can be efficiently solved using dynamic programming.

The size of the data varied a lot in the last decade, from few hundreds of probes per chromosome on original CGH-arrays (3) to the number of nucleotides per chromosome (i.e. about 10^8 for the longest human one) for sequencing technologies. Indeed the forward-backward recursion used in the E-step of the EM algorithm is linear, whereas the regular dynamic programming is quadratic. Still, the number of iterations of the EM algorithm is not known and recent advances have dramatically reduced the complexity of the segmentation algorithms (see e.g. (69; 39)). So many CNV analyses are done using segmentation approaches (using a post-processing for calling) for which dynamic programing-like approaches turn out to perform well (42; 32).

1.2.2 Complex emission distribution

As in any domain, HMM need to be adapted to the nature of the signal under study. Because of the variety of both the biological objects and technologies, a huge variety of emission distributions ψ have been considered in genomics. First ψ must accommodate to the fact that some technologies (e.g. microarrays) provide a continuous signal whereas others (e.g. NGS) a discrete signal (counts). As expected, Gaussian and Poisson emission distributions are respectively the most popular parametric distributions although, in the later case, the negative binomial seems preferable. Still, less easy-to-handle distribution such as the negative binomial have turned out to be more relevant for a series of application using NGS data (see e.g. (71; 67)).

Also, the efficiency of many molecular technologies depends on local properties of the genomic sequence such as the GC-content (proportion of `g` and `c` nucleotides). In order to correct this bias, the emission distribution can account for some local covariate x_t, so $f_g(y_t) = p(y_y|z_t = g)$ becomes $f(y_t; x_t) = p(y_t|z_t = g, x_t)$ (see e.g. (10; 67)).

DNA sequences.

HMM were early used for genome annotation, typically to determine the boundaries of isochores (regions with different nucleotide composition) or to distinguish between gene coding regions and non-coding regions. In such analyses, the observed vector $y = (y_1, y_2, \dots, y_n)$ is made of the DNA sequence itself, each y_t being one of the element of the nucleotides alphabet: $y_t \in \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

The most naive model consists of an HMM with multinomial emission distribution with parameter $\theta_g = (\theta_{ga})_a$ where θ_{ga} represents the frequency of nucleotide a in state g. However, such a simple model turns out to be much too poor to account for the local complexity of the DNA sequence. This model has been generalized in (55), who considered Markov chains as emission distributions to account for the local frequency of di-, tri-, or any oligo-nucleotide. Indeed, as the sufficient statistics for a Markov chain of order m (denoted Mm) are the frequencies of all sequences of $m + 1$-nucleotides. As consequence, a Markov model Mm with transition probability θ_g

$$\theta_g((a_m, \dots a_1); b) := \mathrm{P}(y_t = b | z_t = g, (y_{t-m}, \dots, y_{t-1}) = (a_m, \dots, a_1)),$$

accounts for the frequency of the $(m + 1)$-nucleotides in state g. Such a model is denoted M1-Mm in (55) as the hidden states (z_t) follow an M1 model and the observed sequence (y_t) conditionally arises from an Mm model. As an example, coding-regions are composed of triplets of nucleotides (*codons*) that are ultimately translated into amino-acids, which constitutes the building block of a protein. An M1-M2 model can typically account for this triplet structure (57).

Multivariate signal.

Several molecular technology are intrinsically comparative in the sense that, at each locus t, they provide a pair of measures. This holds for CGH-arrays, which compare the genomic material from a normal with a test sample to detect genomic alterations. This is also true for micro-arrays, which allow to compare the level of transcription of a given locus t in two different conditions. SNP-arrays, which will be discussed later, also yield in a bivariate signal (a_t, b_t). In some situation, one of the signal is simply considered as a covariate (10). In other cases, one may chose to consider a summary variables such as $y_t = b_t - a_t$, although this obviously yields a loss of information.

Copy number variation and loss of heterozigocity.

An interesting case is this of the joint detection of CNV and loss of heterozygocity (LOH) using SNP arrays. Single nucleotide polymorphism (SNP) refers to single nucleotide loci t spread along the genome where two alternative nucleotides are observed in the human population, whereas the neighborhood of the locus is very conserved. The most frequent allele is arbitrarily named A_t and the minor allele B_t. SNP arrays provide a signal (a_t, b_t) where a_t (resp. b_t) proportional to the abundance of A (resp. B) in the sample. At a normal locus, one should have $a_t + b_t \approx 2$ because two copies of each chromosome exist. This sum is often transformed into the Log R ratio $LRR_t = \log_2[(a_t + b_t)/2]$, which is close to 0 in the normal case (see region I in Figure 1.2 left). Furthermore, in a normal situation, the B-allele frequency ($BAF_t := b_t/(a_t + b_t)$) should be close to either 0, 1/2 or 1, corresponding to the three possible normal genotypes: AA, AB and BB (region I in Figure 1.2 right). Note that it is equivalent to observe (a_t, b_t) and (LRR_t, BAF_t).

As described in Section 1.2.1, CNV are revealed by the LRR profile. Still the joint analysis of the two profiles may reveal more complex patterns, such as region V of Figure 1.2 where LRR_t seems normal (left panel) but where no heterozigocity is observed (right panel). Such a pattern suggests that one copy of this region has been lost and has been then rebuild by copying the remaining copy, so that this region is now made of two identical copies, making all loci homozygous. Although such events do not affect the copy number, they may be of interested as their genomic diversity has been reduced of one half and as favorable alleles may have been lost. More complex patterns may arise (see regions II, III and IV of Figure 1.2) when the total number of copies is not 2 (normal case) but more (say 3) resulting in a BAF around 0, 1/3, 2/3 or 1. Note that these ratios remain theoretical as the sample is often contaminated with normal cells, which shrinks the empirical BAF towards 0, 1/2 or 1 (63).

The analysis of such data aims at classifying regions with respect to both the copy number and the heterozigocity status. The fully normal case corresponds to a copy number of two and three possible genotypes at each locus: AA, AB and BB. As a consequence, in this state, the observed signal $y_t = (a_t, b_t)$ is distributed according to a bivariate mixture with three components (corresponding to each possible genotypes). From a general point of view, this problem can be modeled with an HMM where emission distribution are themselves mixtures (see (78))

$$\phi_g(y_t) = \sum_k w_{gk}\psi(y_t;\gamma_k), \qquad \text{whith } \sum_k w_{gk} = 1$$

where ψ is some parametric distribution with parameter γ. Note that, in the present case, the parameter γ_k does not depend on g as several states g may involve the same component (see e.g. regions I and V from Figure 1.2). The reader is left to the determination of the number of components for any given number of copies and heterozogocity status. (31) and (17) provide a more extensive description of this problem.

Non-parametric emission distribution.

Indeed the classification performances of a HMM strongly rely on the choice of the emission distributions it involves. In the case of a bivariate continuous signal, bivariate Gaussian emission distribution are attractive. Still a careful modeling of the respective variance matrix in the vein of (25; 11) can dramatically improve the performances, as shown in (9).

Still, fully parametric emission distributions may not be flexible enough. As mentioned above, mixtures can be used as emission distributions (78; 43). The identifiability of such a model was not addressed in these references but has been proven since then in (29), who propose a non-parametric HMM, considering a kernel-based shape for each emission distribution:

$$f_g(y_t) = \sum_s w_{gs}\psi(y_t - y_s)$$

where ψ is some centered pdf and w_{gs} is the contribution of the data point s to the definition of f_g. Note that, as a counterpart of flexibility, non-parametric HMM may provide less interpretable results, as the estimate of f_g does not always tell to which biological regime the state g corresponds.

1.2.3 Complex hidden states

Most HMM used in genomics are used for classification purposes, for which the hidden space resumes to very few easily interpretable states. Still, the hidden space can be much refined in order to include prior knowledge. Genome annotation is an emblematic example: the aim is to detect coding regions in a genome, only based on the genomic sequence itself but also taking advantage of all the knowledge we have about the structure of genes.

Gene detection from genomic sequences

In prokaryote organisms, the sequence coding for a given gene is often divided into non-adjacent regions called *exons* (in-between the regions being calls *introns*). An HMM dedicated to the detection of gene-coding region must therefore have at least three states corresponding to non-coding (i.e. between gene) regions, exons and introns, respectively. In addition to this, all gene coding regions start with a so-called 'start' codon `atg`. So a fourth state should be added, through which the hidden Markov chain must necessarily transit when going from the non-coding state to the exon state. Figure 1.3 depicts the graph of the hidden Markov chain proposed by (35) to account for a series of such characteristics, including the fact that genes can coded in both directions of the sequence.

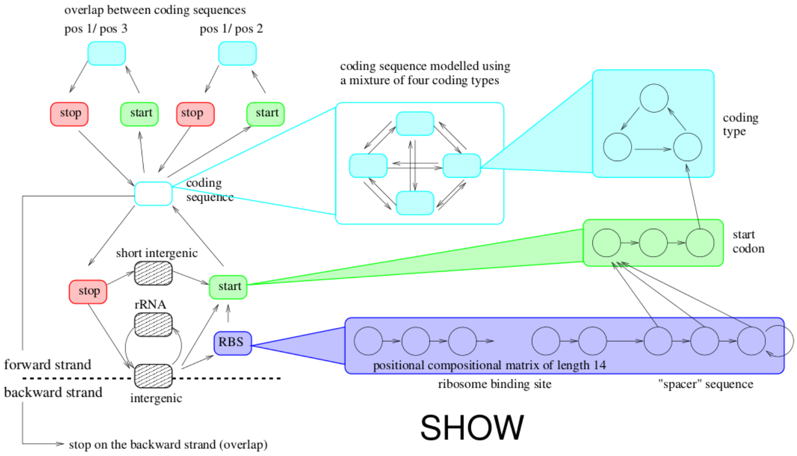


FIGURE 1.3
Structure of the hidden Markov chain used for gene detection (35).

A general property of Markov chains is that the sojourn time of the hidden chain in state g has a geometric distribution with failure probability Π_{gg}. Side information (e.g. the empirical distribution of the length of known exons) may suggest that this property is not desirable (51). Still, the geometric distribution is a side product of the Markovian assumption, which allows for the use of the forward-backward algorithm for the inference. A typical trick to keep the Markovian structure while modifying the sojourn time distribution consists of building 'macro-states', that is to split state g into sub-states $g_1, g_2, \dots g_b$ forcing the transition from sub-state g_{k-1} to sub-state g_k. As the result, the sojourn time has a negative binomial distribution with parameter b and Π_{gg}. The parameter b can be fitted (or manually tuned) to fit the distribution length of the exons known in species similar to this under study.

Gene expression profiles

In the same vein, sub-states can be used to distinguish between a main classification step and a more refined behavior of the signal. For example, (58; 53) are interested in understanding the transcriptionnal landscape, which means both detecting transcribed regions (main task) and the way the level of transcription varies within each of these regions (secondary task). To this aim, the 'transcribed' main state is divided into a series of secondary hidden states corresponding to different levels of transcription and among which Markovian

transition also occur. The secondary hidden structure allows to account for the dynamic dimension of the transcriptional process. Indeed a given gene is typically transcribed from the 'start' to the 'stop' so, at a given time, a fraction of the 'start' end has already started to be degraded (after translation) whereas a fraction of the 'stop' end has still not been synthesized (before the end of transcription).
The distribution of the hidden states can itself be modeled to account for exogenous information such as the annotation of the genome (9).

1.2.4 Non-standard hidden Markov structure

As mentioned above, HMM are quite popular in genomics because of the 1D structure that underlines many 'omic' data. Still more complex hidden structures can be encountered, two examples of which we give below.

Paired HMM for sequence alignment.

Sequence alignment is one of the oldest problem in bioinformatics, which aims at comparing genomics regions (e.g. genes) observed in two different species. Suppose we observe the sequences A = (gatctgaac) and B = (gacgtta), the first step to compare them is to align them, that is to make them match as well as possible. Such an alignment can be viewed as an HMM (xXxXxXx) with bivariate observed data $y_t = (a_t, b_t) \in \{-, \text{a, c, g, t}\}^2$, where a_t (resp. b_t) stands for the letter from sequence A (resp. B) observed at the aligned position t. Four hidden states are then typically considered, the corresponding emission distribution having disjoint support:

match: {(a, a), (c, c), (g, g), (t, t)};

mismatch: {(a, c), (a, g), (a, t), (c, a), (c, g), ..., (g, t)};

insertion: {(–, a), (–, c), (–, g), (–, t)};

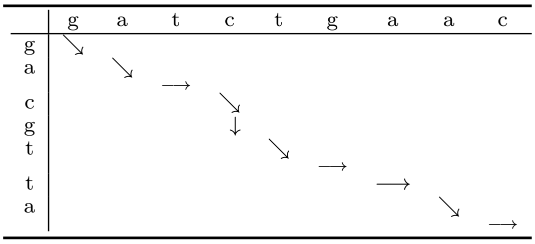
deletion: {(a, –), (c, –), (g, –), (t, –)};

Note that the aligned positions t are not observed in advance. In practice, the whole inference process is rarely carried out. Most often, the transition matrix is given in advance and its entries are interpreted as costs of each possible transitions, the emission distributions being uniform over their respective supports (except for mismatches). The alignment algorithm then simply consists of the Viterbi algorithm. Figure 1.4 (left) gives a representation of the most probable path (top), and the resulting alignment (bottom).

Tree-structured models.

Trees are often used to described the past evolution of a population, a trait or a genome. The tree structure is indeed consistent with many evolutionary scenarios. In many situations, only present observations are available, although we are interested in the past evolution. As a consequence, we are faced with situations as depicted in Figure 1.5 where the data at all past nodes (also referred to as ancestor nodes) are unobserved. In this framework, most model assume that the trait (or the genome sequence) evolves as a Markov process along the branches of a given phylogenetic tree. The aim is then to infer the parameters that governs this evolutionary process, which typically requires some insight about the value of the trait at the ancestor nodes.

EM can be used to infer the value of the trait in an ancestor node. As often, the critical step is the E step where moments of the conditional distribution $p(z|y)$ need to be computed. The case of the tree is actually not far from this of HMM where the forward-backward



$t:$	1	2	3	4	5	6	7	8	9	10
$A:$	g	a	t	c	–	t	g	a	a	c
$B:$	g	a	–	c	g	t	–	t	a	–

FIGURE 1.4
Label: Example of paired-HMM for sequence alignment. Top left: most probable hidden path: ↘= match, ⟶= mismatch, ↓= insertion, , ⟶= deletion. Bottom left: resulting alignment.

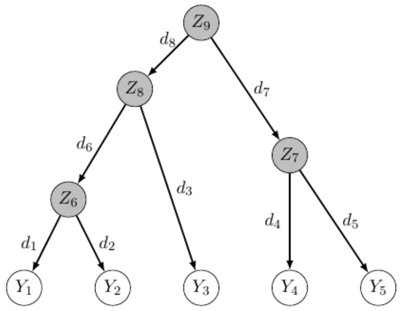


FIGURE 1.5
Example of an evolutionary tree. Only present nodes are observed (y_i) whereas ancestor nodes (z_j) are hidden.

recursion enables us to compute these moments. Indeed an 'upward-downward' recursion can derived in a similar way (24) to get the conditional distribution $p(z_j|y)$ for each internal node z_j.

We only give a flavor of the upward recursion, based on the example of Figure 1.5. The upward recursion goes from the leafs (y_i) to the root and consists of computing the conditional distribution of each ancestor node given its offsprings. $p(z_7|y_4, y_5)$ and $p(z_6|y_1, y_2)$ are first computed directly. The remaining conditional distributions are then obtained as

$$\begin{aligned} p(z_8|y_1, y_2, y_3) &= \int p(z_8|z_6, y_3)\ p(z_6|y_1, y_2)\ \mathrm{d}z_6 \\ p(z_9|y_1, y_2, y_3, y_4, y_5) &= \iint p(z_9|z_8, y_7)\ p(z_8|y_1, y_2, y_3)\ p(z_7|y_4, y_5)\ \mathrm{d}z_7\ \mathrm{d}z_8 \end{aligned}$$

which ends the upward recursion. We only refer to (44) and (6) for two applications of this type of modeling.

1.3 Complex dependency structure

In the preceding section, we presented models for which a genuine EM can be applied because the dependency structure is either sequential or tree structured. More complex structures may be required in some applications.

1.3.1 Markov Random Fields

Considering for example the genetic structure of a population (see Section 1.1.1), geo-referenced individuals may be considered. Spatial and genetic information are somehow related and two individuals living close by are more likely to be genetically close. To take this prior information into account, it is relevant to consider a Hidden Markov Random Field (HRMF) as prior distribution for the latent variable describing the population structure (26).

Markov fields extend the dependence structure to stochastic processes with indexes belonging to a multidimensional space, rather than simply to a subset of $\mathbb{R}$. There are two types of Markov Random field (MRF): Markov fields with continuous indexes, commonly used in theoretical physics and Markov fields with discrete indexes, used, among other things, as models for statistics of a spatial nature. In applications to genomics we are mainly concerned with the second category.

When the domain of the index is a subset of $\mathbb{R}^d$ rather than a subset of $\mathbb{R}$, the idea of left and right in relation to an index no longer applies, and MRF must revert to a more general concept of neighborhood. The neighborhood system can be modeled via a contiguity graph where each node corresponds to an index and each vertex to a neighborhood relationship.

In the population structure example, nodes are not distributed regularly and Markov modeling requires that relations of contiguity are explicitly defined. One solution is to draw a Voronoï tessellation and to specify that two sites are contiguous if their respective Voronoï tiles have an edge in common. The Strauss model represents a natural prior distribution for the latent cluster variable. It can be considered as a generalization of the Ising model, in the case where the variables take G discrete values ($G \geq 2$). In the isotropic case — i.e. where there is no particular spatial direction — the Gibbs distribution is defined by the energy function:

$$U(z) = -\beta \sum_{r,s \in S : r \sim s} \mathbb{I}_{\{z_s = z_r\}} = -\beta \sum_{r,s \in S : r \sim s} z_s \cdot z_r \tag{1.2}$$

where the binary vector z_s denotes the class of node s ($z_{sk} = 1$ if node s belongs to class g) and $r \sim s$ means that r and s are neighbors. This energy function therefore counts the number of pairs of contiguous nodes which have the same value, and is maximized when the variables of the entire set of nodes are identical. In physics this model is referred to as the Potts model.

The same Potts model has been used as prior for the spatial normalization of array-CGH data (56). In this latter case the hidden variable was coding for a type of experimental artifact. Identifying this spatial experimental bias allowed to tune the bias adjustment for each area of the micro-array

This type of prior about graph neighborhood can also be found in differential analysis. The classical approach for differential analysis relies on univariate test statistics for selecting a short list of genes. Then links from selected genes to known biological pathways through gene set enrichment analysis can be performed in order to identify involved pathways (80). In that context directly assuming a prior model considering that neighboring genes in the network are more likely to have a joint effect may improve the relevance of the differential analysis.

1.3.2 Stochastic block-model

The analysis of biological networks have become a common place in bioinformatics. Such networks typically describe the interactions between a set of entities such as genes, proteins or, at a higher level, bacterial species. Such data typically consists of a set of n nodes, each corresponding to an entity, and in the value y_{ij} of the edge between nodes i and j. Because of the diversity of the interactions, the form of y_{ij} ranges from binary (simple presence or absence of the edge), to continuous uni- or multivariate.

Understanding the topology of such a network has been one of the primary task. The clustering point-of-view can be used to this aim, assigning each node to a specific class with a typical role in the network. This results in the well-known stochastic block-model (SBM) (27; 33), in which hidden classes $\{z_i\}$'s are drawn independently for each node and edges are drawn independently conditionally on the z_i's. Importantly, the distribution of the edge y_{ij} is conditional on the class of both nodes i and j:

$$p(y_{ij}|z_i, z_j) = \phi_{z_i,z_j}(y_{ij}).$$

In the binary case, conditional on $z_i = g$ and $z_j = g'$, edge y_{ij} is present with probability $\gamma_{gg'}$. (60) give a series of examples of use of SBM for various biological networks.

1.3.3 Inference issues

Both HMRF and SBM raise inference issue. Indeed, in both cases the hidden labels z_i's are not independent conditionally on the observed data y. In HRMF, their conditional dependency structure is given by the graph G and in SBM it turns out to be the complete graph (46). The latent block-model introduced in Section 1.1.2 raises similar issues. In such cases, no factorization can be hoped to compute efficiently the moments of the conditional distribution $p(z|y)$ that a required in a regular EM algorithm.

Because of the size of the data, variational approximation are often used in this field, as they result in deterministic and reasonably fast algorithms. The general principle of these techniques is to replace the hard-to-compute distribution $p(z|y)$ with an approximate distribution q easier to handle. So, the regular E step is replaced with an approximation step

$$q^{(h+1)} = \arg\min_{q\in\mathcal{Q}} D[q^*(\cdot)||p_{\theta^{(h)}}(\cdot|y)],$$

where $\mathcal{Q}$ is a restricted class of distribution (typically factorisable) and D is a divergence measure. In the specific case where D is chosen to be the Kullback-Leibler divergence, it can be seen that the variational EM (VEM) algorithm aims at maximizing a lower bound of the likelihood of the data simply because, denoting E_q the expectation wit respect to q,

$$\begin{aligned}\log p_\theta(y) &= E_{p_\theta(\cdot|y)}[\log p_\theta(y,z)] - E_{p_\theta(\cdot|y)}[\log q(z)] \\ &\geq \log p_\theta(y) - KL[q(\cdot)||p(\cdot|y)] = E_q[\log p_\theta(y,z)] - E_q[\log q(z)].\end{aligned}$$

A huge literature exist on these techniques. We refer to (36) for a tutorial, to (52) for a discussion on the choice of D and to (79) for a complete tour.

This approach can be extended to a Bayesian setting, resulting in so-called variational Bayes inference. In this case, an approximation of the joint conditional distribution $p(\theta, z|y)$ is looked for, see e.g. (7) for an introduction and (45; 1) for applications to SBM.

Bibliography

[1] C. Aicher, A.Z. Jacobs, and A. Clauset. Adapting the stochastic block model to edge-weighted networks. ICML Workshop on Structured Learning (SLG), 2013.

[2] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray. Chromosome aberrations in solid tumors. *Nature genetics*, 34(4):369–376, 2003.

[3] Donna G Albertson and Daniel Pinkel. Genomic microarrays in human genetic disease and cancer. *Human molecular genetics*, 12(suppl 2):R145–R152, 2003.

[4] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[5] D. B. Allison, G. Gadbury, M. Heo, J. Fernandez, C.-K. Lee, T. A. Prolla, and R. A Weindruch. Mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39:1–20, 2002.

[6] P. Bastide, M. Mariadassou, and S. Robin. Detection of adaptive shifts on phylogenies using shifted stochastic processes on a tree. Technical report, arXiv:1508.00225, 2015.

[7] J. Beal, M. and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayes. Statist.*, 7:543–52, 2003.

[8] Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology*, 10(3-4):373–384, 2003.

[9] C. Bérard, M.-L. Martin-Magniette, V. Brunaud, S. Aubourg, S. Robin, et al. Unsupervised classification for tiling arrays: ChIP-chip and transcriptome. *Statistical applications in genetics and molecular biology*, 10(1):1–22, 2011.

[10] Caroline Bérard, Michael Seifert, Tristan Mary-Huard, and Marie-Laure Martin-Magniette. MultiChIPmixHMM: an R package for chip-chip data analysis modeling spatial dependencies and multiple replicates. *BMC bioinformatics*, 14(1):271, 2013.

[11] Christophe Biernacki, Gilles Celeux, Gérard Govaert, and Florent Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, 51(2):587–600, 2006.

[12] Laurent Bordes, Céline Delmas, and Pierre Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian journal of statistics*, 33(4):733–752, 2006.

[13] Matthieu Bouaziz, Christophe Ambroise, and Mickael Guedj. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. *PLoS One*, 6(12):e28845, 2011.

[14] Matthieu Bouaziz, Caroline Paccard, Mickael Guedj, and Christophe Ambroise. Ships: spectral hierarchical clustering for the inference of population structure in genetic studies. *PLoS One*, 7(10), 2012.

[15] Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

[16] Luigi Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. *The history and geography of human genes*. Princeton university press, 1994.

[17] Hao Chen, Haipeng Xing, and Nancy R Zhang. Estimation of parent specific dna copy number in tumors using high-density genotyping arrays. *PLoS Comput Biol*, 7(1):e1001060, 2011.

[18] Jukka Corander, Patrik Waldmann, and Mikko J Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374, 2003.

[19] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[20] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.

[21] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[22] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

[23] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[24] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

[25] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[26] Olivier François, Sophie Ancelet, and Gilles Guillot. Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2):805–816, 2006.

[27] O. Frank and F. Harary. Cluster inference by using transitivity indices in empirical graphs. *J. Amer. Statist. Assoc.*, 77(380):835–840, 1982.

[28] Jane Fridlyand, Antoine M Snijders, Dan Pinkel, Donna G Albertson, and Ajay N Jain. Hidden markov models approach to the analysis of array CGH data. *Journal of multivariate analysis*, 90(1):132–153, 2004.

[29] E Gassiat, A Cleynen, and S Robin. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, pages 1–11, 2015.

[30] Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473, 2003.

[31] Chris D Greenman, Graham Bignell, Adam Butler, Sarah Edkins, Jon Hinton, Dave Beare, Sajani Swamy, Thomas Santarius, Lina Chen, Sara Widaa, et al. Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, 11(1):164–175, 2010.

[32] Toby Dylan Hocking. *Learning algorithms and statistical software, with applications to bioinformatics*. PhD thesis, École normale supérieure de Cachan-ENS Cachan, 2012.

[33] P. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: some first steps. *Social networks*, 5:109–137, 1983.

[34] Haiyan Huang, Li Cai, and Wing H Wong. Clustering analysis of sage transcription profiles using a poisson approach. In *Serial Analysis of Gene Expression (SAGE)*, pages 185–198. Springer, 2008.

[35] M. Ibrahim, P. Nicolas, Ph. Bessières, A. Bolotin, V. Monnet, and R. Gardan. A genome-wide survey of short coding sequences in streptococci. *Microbiology*, 153(11):3631–3644, 2007.

[36] T. Jaakkola. *Advanced mean field methods: theory and practice*, chapter Tutorial on variational approximation methods. MIT Press, 2000.

[37] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[38] Christine Keribin, Vincent Brault, Gilles Celeux, and Gérard Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014.

[39] Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[40] Kyungpil Kim, Shibo Zhang, Keni Jiang, Li Cai, In-Beum Lee, Lewis J Feldman, and Haiyan Huang. Measuring similarities between gene expression profiles through new data transformations. *BMC bioinformatics*, 8(1):29, 2007.

[41] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[42] Weil R Lai, Mark D Johnson, Raju Kucherlapati, and Peter J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21(19):3763–3770, 2005.

[43] Heike Lange, Helene Zuber, François M Sement, Johana Chicher, Lauriane Kuhn, Philippe Hammann, Véronique Brunaud, Caroline Berard, Nathalie Bouteiller, Sandrine Balzergue, et al. The rna helicases atmtr4 and hen2 target specific subsets of nuclear transcripts for degradation by the nuclear exosome in arabidopsis thaliana. 2014.

[44] Nicolas Lartillot. A Phylogenetic Kalman Filter for Ancestral Trait Reconstruction Using Molecular Data. *Bioinformatics*, 30(4):488–496, February 2014.

[45] P. Latouche, E. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statis. Model.*, 12(1):93–115, 2012.

[46] Matias, C. and Robin, S. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proc.*, 47:55–74, 2014.

[47] Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

[48] G. McLachlan, R.W. Bean, and L. Ben-Tovim Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608–1615, 2006.

[49] Geoffrey McLachlan, Kim-Anh Do, and Christophe Ambroise. *Analyzing microarray gene expression data*, volume 422. John Wiley & Sons, 2005.

[50] Geoffrey J. McLachlan, RW Bean, and David Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

[51] Christelle Melodelima, Laurent Guéguen, Didier Piau, and Christian Gautier. A computational prediction of isochores based on hidden markov models. *Gene*, 385:41–49, 2006.

[52] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005. `ftp://ftp.research.microsoft.com/pub/tr/TR-2005-173.pdf`.

[53] Bogdan Mirauta, Pierre Nicolas, and Hugues Richard. Parseq: reconstruction of microbial transcription landscape from rna-seq read counts using state-space models. *Bioinformatics*, page btu042, 2014.

[54] Joanna L Mountain and L Luca Cavalli-Sforza. Multilocus genotypes, a tree of individuals, and human evolutionary history. *The American Journal of Human Genetics*, 61(3):705–718, 1997.

[55] Florence Muri. Modelling bacterial genomes using hidden markov models. In *Compstat*, pages 89–100. Springer, 1998.

[56] P. Neuvial, P. Hupé, I. Brito, S. Liva, É. Manié, Ca. Brennetot, F. Radvanyi, A. Aurias, and E. Barillot. Spatial normalization of array-cgh data. *BMC bioinformatics*, 7(1):264, 2006.

[57] P. Nicolas, L. Bize, F. Muri, M. Hoebeke, S.D. Rodolphe, F.and Ehrlich, B. Prum, and Ph. Bessières. Mining bacillus subtilis chromosome heterogeneities using hidden markov models. *Nucleic Acids Research*, 30(6):1418–1426, 2002.

[58] P. Nicolas, A. Leduc, S. Robin, S. Rasmussen, H. Jarmer, and P. Bessieres. Transcriptional landscape estimation from tiling array data using a model of signal shift and drift. *Bioinformatics*, 25:2341–2347, Sep 2009.

[59] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8:1145–1164, 2007.

[60] F. Picard, V. Miele, J.-J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using mixnet. *BMC Bioinformatics*, Suppl 6:S17, 2009. `doi:10.1186/1471-2105-10-S6-S17`.

[61] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 6(27):1, 2005. `www.biomedcentral.com/1471-2105/6/27`.

[62] F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3):758–766, 2007. doi: 10.1111/j.1541-0420.2006.00729.x.

[63] T. Popova, E. Manié, D. Stoppa-Lyonnet, G. Rigaill, E. Barillot, M.-H. Stern, et al. Genome alteration print (gap): a tool to visualize and mine complex cancer genomic profiles obtained by snp arrays. *Genome Biol*, 10(11):R128–R128, 2009.

[64] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[65] A. E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[66] A.l Raj, M. Stephens, and J.K. Pritchard. faststructure: variational inference of population structure in large snp data sets. *Genetics*, 197(2):573–589, 2014.

[67] Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol*, 12(7):R67, 2011.

[68] A. Rau, C. Maugis-Rabusseau, M.-L. Martin-Magniette, and G. Celeux. Co-expression analysis of high-throughput transcriptome sequencing data with poisson mixture models. *Bioinformatics*, 31:1420–1427, 2015.

[69] G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. Technical report, arXiv:1004.0887, 2010.

[70] S. Robin, A. Bar-Hen, J.-J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational Statistics & Data Analysis*, 51(12):5483–93, 2007.

[71] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[72] Etienne Roquain. Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique*, 152(2):3–38, 2011.

[73] M Seifert. Hidden markov models with applications in computational biology. *Saarbrücken, Germany: SVH-Verlag*, 2013.

[74] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[75] Mahlet G Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617, 2005.

[76] Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, 28(4):289–301, 2005.

[77] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.

[78] Stevenn Volant, Caroline Bérard, Marie-Laure Martin-Magniette, and Stéphane Robin. Hidden markov models with mixtures as emission distributions. *Statistics and Computing*, pages 1–12, 2013.

[79] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008. http:/dx.doi.org/10.1561/2200000001.

[80] Z. Wei and H. Li. A markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, 2007.

[81] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 2010.

[82] Ka Yee Yeung, Chris Fraley, Alejandro Murua, Adrian E Raftery, and Walter L Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.