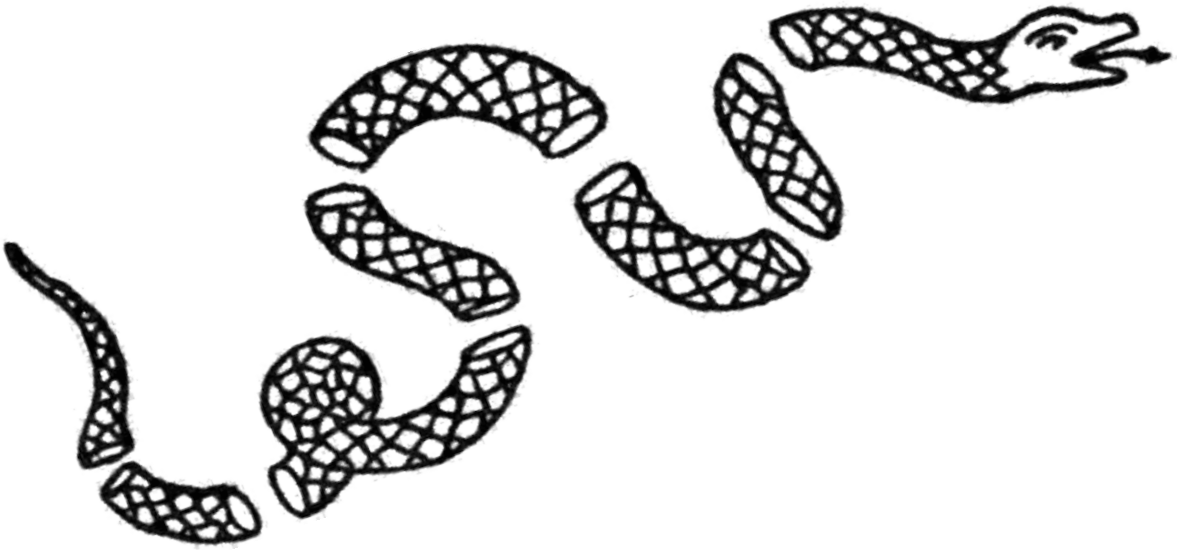
Version 1.1



Programming with Big Data in R

Speaking Serial R with a Parallel Accent

Package Examples and Demonstrations

pbdR Core Team

Speaking Serial R with a Parallel Accent

pbdR Package Examples and Demonstrations

Drew Schmidt
Remote Data Analysis and Visualization Center,
University of Tennessee, Knoxville

Wei-Chen Chen
Computer Science and Mathematics Division,
Oak Ridge National Laboratory

Pragneskumar Patel
Remote Data Analysis and Visualization Center,
University of Tennessee, Knoxville

George Ostrouchov
Computer Science and Mathematics Division,
Oak Ridge National Laboratory

Version 1.1

June 12, 2013

This publication was typeset using LaTeX. Illustrations were created using the **ggplot2** package (Wickham, 2009), native R functions, and Microsoft Powerpoint.

Contents

List of Figures

List of Tables

Preface

A common question is "Who is the master?" and the proper reply is "There is no master!"
– George Ostrouchov

Acknowledgements

Schmidt, Ostrouchov, and Patel were supported in part by the project "NICS Remote Data Analysis and Visualization Center" funded by the Office of Cyberinfrastructure of the U.S. National Science Foundation under Award No. ARRA-NSF-OCI-0906324 for NICS-RDAV center. Chen and Ostrouchov were supported in part by the project "Visual Data Exploration and Analysis of Ultra-large Climate Data" funded by U.S. DOE Office of Science under Contract No. DE-AC05-00OR22725.

This work used resources of National Institute for Computational Sciences at the University of Tennessee, Knoxville, which is supported by the Office of Cyberinfrastructure of the U.S. National Science Foundation under Award No. ARRA-NSF-OCI-0906324 for NICS-RDAV center. This work also used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work used resources of the Newton HPC Program at the University of Tennessee, Knoxville.

We also thank Brian D. Ripley, Kurt Hornik, Uwe Ligges, and Simon Urbanek from the R Core Team for discussing package release issues and helping us solve portability problems on different platforms.

Disclaimer

Warning: The findings and conclusions in this article have not been formally disseminated by the U.S. Department of Energy and should not be construed to represent any determination or policy of University, Agency and National Laboratory.

This document is written to explain the main functions of **pbdDEMO** (Schmidt *et al.*, 2013), version 0.1-1. Every effort will be made to ensure future versions are consistent with these instructions, but features in later versions may not be explained in this document.

Information about the functionality of this package, and any changes in future versions can be found on website: "Programming with Big Data in R" at `http://r-pbd.org/`.

Part I

Preliminaries

1 Likelihood

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."
– Sherlock Holmes

1.1 Introduction

This is a preamble chapter for Chapters **??** and 2 where applications are heavily relied on likelihood functions which one of the most important modern Statistical technique. The "likelihood" was popularized in mathematical statistics by R.A. Fisher in 1922: "On the mathematical foundations of theoretical statistics."(Fisher, 1922) In short, the likelihood is a way based on data to build up theoretical inference for the facts.

We introduce general notations for likelihood function which is a standard method for parametric statistics and is useful for statistical inference (Casella and Berger, 2001). Two useful distributions are introduced. The normal distribution additional to linear model is applied to the example in Section **??**. The multivariate normal distribution is also popular to model high dimensional data such as model-based clustering in Chapter **??**.

Suppose $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ is a random sample, which means independent identically distributed (i.i.d.), from a population which has a distribution $\mathcal{F}(\boldsymbol{\theta})$ with unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ is the parameter space. Suppose further $\mathcal{F}$ has a probability density function (pdf) $f(\boldsymbol{X}_n; \boldsymbol{\theta})$ provided an appropriate support. The goal is to estimate $\boldsymbol{\theta}$ based on the observed data $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$. Ideally, we want to infer what is the best candidate of $\boldsymbol{\theta}$ where $\boldsymbol{x}$ is observed from. Unlike in Mathematics, $\boldsymbol{x}$ is known, but $\boldsymbol{\theta}$ is unknown to be determined in Statistics.

Typically, a fancy way to estimate $\boldsymbol{\theta}$ is based on the likelihood function for the data $\boldsymbol{x}$

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n; \boldsymbol{\theta}) \tag{1.1}$$

or the log likelihood function

$$\log L(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{n=1}^{N} \log f(\boldsymbol{x}_n; \boldsymbol{\theta}). \tag{1.2}$$

The product of Equation (1.1) is due to the independent assumption of $\boldsymbol{X}$, but the $L(\boldsymbol{\theta}; \boldsymbol{x})$ may blow to infinity or negative infinity quickly as sample size N increased. While, Equation (1.2) has some better properties for some distribution families and is more numerically stable than Equation (1.1). We then either analytically or numerically maximize Equation (1.2) over $\boldsymbol{\Theta}$ to obtain a maximum likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}}_{ML} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \log L(\boldsymbol{\theta}; \boldsymbol{x})$$

in fairly-wide assumptions such as regularity conditions of parameter space and parameter does not depend on support, see Casella and Berger (2001) for details.

1.2 Normal Distribution

Section **??** is one way to find $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$ for a linear model without parametric assumption via ordinary least square estimator $\hat{\boldsymbol{\theta}}_{ols} = \{\hat{\boldsymbol{\beta}}_{ols}, \hat{\sigma}^2_{ols}\}$. Beside Gauss-Markov Theorem, an alternative way is based on likelihood approach by assuming an identical normal distribution with mean zero and variance σ^2 to the independent error terms of Equation (**??**). This implies a normal distribution to the response y_n for $n = 1, 2, \ldots, N$ that

$$y_n \overset{i.i.d}{\sim} N(\boldsymbol{x}_n^\top \boldsymbol{\beta}, \sigma^2) \tag{1.3}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$, and $\boldsymbol{\beta}$ and $\boldsymbol{x}_n$ has dimension $p \times 1$.

One may construct a log likelihood based on the normal density function as

$$\log L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \sum_{n=1}^{N} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_n - \boldsymbol{x}_n^\top \boldsymbol{\beta})^2}{2\sigma^2} \right]. \tag{1.4}$$

The MLEs $\hat{\boldsymbol{\theta}}_{ML} = \{\hat{\beta}_{ML}, \hat{\sigma}^2_{ML}\}$ can be obtained analytically for this case by taking the first derivatives of Equation (1.4), setting them to zero, and solving the equations. The implementations for numerical solutions (from analytical solutions) or numerical optimization of Equation (1.4) is not difficult and leaved in Exercise 1-7.

The assumption of Statement (1.3) limit the modeling capability. We introduce a more general approach next for better modeling. Since the independent assumption and Multivariate Statistics, the Statement (1.3) implies a multivariate normal distribution (MVN) (introduced in Section 1.4) to response variables $\boldsymbol{y}$ with dimension $N \times 1$ that

$$\boldsymbol{y} \sim MVN_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1.5}$$

where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ with length N, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{I}$ is an $N \times N$ identity matrix. In this case, the $\boldsymbol{y}$ has a density function as

$$\phi_N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})}$$

and the log likelihood can reduce to Equation (1.4). The MLEs are $\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ and $\sigma^2_{ML} = \frac{1}{N}(\boldsymbol{y} - \bar{y}\boldsymbol{1})^\top (\boldsymbol{y} - \bar{y}\boldsymbol{1})$ where $\bar{y}$ is the average of $\boldsymbol{y}$, and $\boldsymbol{1}$ is an one vector with length N.

1.3 Likelihood Ratio Test

An advance statistical inference tool is based on likelihood ratio test (LRT) provided suitable assumption holds. Suppose we have data $\boldsymbol{X}$ and want to test a hypothesis

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \text{ v.s. } H_a : \boldsymbol{\theta} \in \boldsymbol{\Theta}_a$$

where $\boldsymbol{\Theta}_0 \neq \boldsymbol{\Theta}_a$ that two spaces are not equivalent. The LRT says

$$-2\log\Lambda(\boldsymbol{\theta}_0, \boldsymbol{\theta}_a; \boldsymbol{X}) := -2\log\frac{\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L(\boldsymbol{\theta}; \boldsymbol{X})}{\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0\cup\boldsymbol{\Theta}_a} L(\boldsymbol{\theta}; \boldsymbol{X})} \sim \chi^2_p \tag{1.6}$$

where $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_a$ are parameters that has the maximum likelihoods on spaces $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a$, and χ^2_p is a chi-squared distribution with p degrees of freedom. In some case, the p can be simplified as the number of dimension difference of $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_0$.

For example, in the least square case, such as Statement (1.5), we may want to know

$$H_0 : \boldsymbol{\sigma}^2 = 1 \text{ v.s. } H_a : \boldsymbol{\sigma}^2 > 0$$

which means $\boldsymbol{\Theta}_0 = \{\boldsymbol{\beta}\}$ and $\boldsymbol{\Theta}_a = \{\boldsymbol{\beta}, \sigma^2\}$. Note that $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}_a$, so $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a = \boldsymbol{\Theta}_a$. Given the MLEs $\hat{\boldsymbol{\theta}}_{0\,ML}$ and $\hat{\boldsymbol{\theta}}_{a\,ML}$ on two spaces $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_a$, respectively. The LRT will be

$$-2\log\hat{\Lambda}(\hat{\boldsymbol{\theta}}_{0\,ML}, \hat{\boldsymbol{\theta}}_{a\,ML}; \boldsymbol{X}) := -2\log\frac{L(\hat{\boldsymbol{\theta}}_{0\,ML}; \boldsymbol{X})}{L(\hat{\boldsymbol{\theta}}_{a\,ML}; \boldsymbol{X})} \sim \chi^2_1.$$

For type I error $\alpha = 0.05$, if the value

$$-2\log\hat{\Lambda}(\hat{\boldsymbol{\theta}}_{0\,ML}, \hat{\boldsymbol{\theta}}_{a\,ML}; \boldsymbol{X}) > q_{\chi^2_1}(0.95) \approx 3.84$$

where $q_{\chi^2_1}(0.95)$ is the 95% quantile of chi-squared distribution with 1 degree of freedom, then we may reject $H_0 : \boldsymbol{\sigma}^2 = 1$ provided type I error is no greater than 0.05 level.

Note that the LRT introduced here is not dependent on the types of distributions, but has nested parameter space restriction and some regular conditions of parameter space. See Casella and Berger (2001); Ferguson (1996) for more details of LRTs.

1.4 Multivariate Normal Distribution

Suppose $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$ is a random sample from multivariate normal distribution (MVN)

$$\boldsymbol{X}_n \overset{i.i.d}{\sim} MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1.7}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $\boldsymbol{\mu}$ is a center with dimension $p \times 1$, and $\boldsymbol{\Sigma}$ is an $p \times p$ dispersion matrix. The $\boldsymbol{X}_n$ has a density function as

$$\phi_p(\boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})}$$

In general, $\boldsymbol{\Sigma}$ could be an unstructured dispersion and positive definite. Excepting over fitting problem, an unstructured dispersion $\boldsymbol{\Sigma}$ is desirable to characterize correlation of dimensions since the estimation of $\boldsymbol{\Sigma}$ is completely supported by observed data.

Let $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_N^\top)^\top$ be an observed data matrix with dimension $N \times p$. The log likelihood function for N observations is

$$\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \sum_{n=1}^{N} -\frac{1}{2}\left[p \log(2\pi) + \log|\boldsymbol{\Sigma}| + (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right].$$

The problem is the computing time grows as N and p increased. In some numerically cases, such as model-based clustering in Chapter **??**, the total log likelihood is repeated computed in each iteration for all samples and all components.

Suppose $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known, and no over- and under-flow, an efficient way is given in next

R Code

```
U <- chol(SIGMA)
logdet <- sum(log(abs(diag(U)))) * 2
B <- sweep(X.spmd, 2, MU) %*% backsolve(U, diag(1, p))

# The over- and under-flow need extral care after this step.
distval.spmd <- rowSums(B * B)

distval <- allreduce(sum(distval.spmd))
total.logL <- -(p * log(2 *pi) + logdet + distval) * 0.5
```

where `X.spmd` is a SPMD row-major matrix with dimension `N.spmd` by `p`, `MU` is a vector with length `p`, and `SIGMA` is a `p` by `p` positive definite matrix. The sample size N will be the sum of `N.spmd` across all processors. Note that this trick of computing log likelihood is an one-pass implementation of `X.spmd`, `MU`, and `SIGMA`. See HPSC (Chen and Ostrouchov, 2011) or Golub and Van Loan (1996) for more details.

1.5 Exercises

1-1 What is the definition of "independent identical distributed"?

1-2 What is the definition of "probability density function"?

1-3 Suppose $g(\cdot)$ is a continuous function provided appropriate support, argue that $g\left(\hat{\theta}_{ML}\right)$ is still a maximum likelihood estimator of $g(\theta)$.

1-4 Derive MLEs from Equation (1.4).

1-5 As Exercise **??**, argue that $\hat{\boldsymbol{\beta}}_{ML}$ of Equation (1.4) is also an unbiased estimator of $\boldsymbol{\beta}$.

1-6 Argue that $\hat{\sigma}^2_{ML}$ of Equation (1.4) is a biased estimator, but it is an asymptotic unbiased estimator of σ^2.

1-7 Assume data are stored in SPMD row-major matrix format, implement an optimization function for Equation (1.4), numerically optimized via `optim()` in R. Verify the results with the analytical solution.

1-8 Argue that Statement (1.3) implies Statement (1.5) provided appropriated assumption hold.

1-9 Give an example that X and Y are both have a normal distribution but (X, Y) is not a multivariate normal distribution.

1-10 Give an example that (X, Y) has a multivariate normal distribution, but X and Y do not have an independent normal distribution.

1-11 Prove Statement (1.6). Hint: Ferguson (1996).

1-12 Implement similar trick of Section 1.4 to Principal Component Analysis (PCA). Hint: HPSC (Chen and Ostrouchov, 2011).

2

Phylogenetic Clustering (Phyloclustering)

"What one man can invent another can discover."
– Sherlock Holmes

2.1 Introduction

Phylogenetic Clustering (Phyloclustering) discovers population structure based on information of DNA/RNA sequences by combining two inventions: model-based clustering with evolutionary models (Chen, 2011). Note that what speaking here, regarding to "evolutionary", is a mathematical/statistical model to interpret biological targets. Neither religion nor theology is involved.

In an over simplified case, suppose a sequence is composed by four nucleotides $\mathcal{S} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$. Assume a sequence $\boldsymbol{x}_n = \{x_{n1}, x_{n2}, \ldots, x_{nL}\} \in \mathcal{S}$ has L loci (positions ordered) and is observed from a population, but may have K subpopulations that similar sequence patterns are expected within each common subpopulation. Each subpopulation is represented by a common center sequence $\boldsymbol{\mu}_k = \{\mu_{k1}, \mu_{k2}, \ldots, \mu_{kL}\} \in \mathcal{S}$ which may or may not hypothetically exit in population and has to be determined. Therefore, each sequence has a probability mutated/evolved from any center sequence. The higher the probability, the closer to the center sequence.

The evolutionary model is based on a continuous time Markov chain (CTMC) on a state space $\mathcal{S}$. where the mutation process is characterized by an instantaneously rate matrix $\boldsymbol{Q}$ with dimension 4×4, i.e. rate at scale of tiny mutation time $t \to 0$. We use the following steps to construct the likelihood function as introduced in Chapter 1:

1. Given the above setting, the mutation chance from a nucleotide x to a nucleotide y in time t is
$$\mathbb{P}_{x,y}(t) = e^{\boldsymbol{Q}_{x,y}t} \tag{2.1}$$
for all $x, y \in \mathcal{S}$.

2. Assume each locus is mutated independently, then the mutation chance (the transition

probability) from $\boldsymbol{\mu}_k$ to $\boldsymbol{x}_n$ in time t is

$$p_{\boldsymbol{\mu}_k,\boldsymbol{x}_n}(t) = \prod_{l=1}^{L} \mathbb{P}_{\mu_{kl},\boldsymbol{x}_{nl}}(t)$$

for all $\boldsymbol{\mu}_{kl}, \boldsymbol{x}_{nl} \in \mathcal{S}$.

3. Suppose there are K subpopulations with mixing proportion η_k's, then the mutation chance from a sequence $\boldsymbol{\mu}_k$ to a sequence $\boldsymbol{x}_n$ is

$$f(\boldsymbol{x}_n; \boldsymbol{\theta}_K) = \sum_{k=1}^{K} \eta_k p_{\boldsymbol{\mu}_k,\boldsymbol{x}_n}(t) \tag{2.2}$$

where $\boldsymbol{\theta}_K = \{\eta_1, \eta_2, \ldots, \eta_{K-1}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K, \boldsymbol{Q}, t\}$ are unknown and to be determined. For simplicity, assume $\boldsymbol{Q}$ and t are identical across K subpopulations. Denote the distribution $\mathcal{F}(\boldsymbol{\theta}_K)$ of the density function $f(\boldsymbol{x}_n; \boldsymbol{\theta}_K)$ for $\boldsymbol{x}_n$.

4. Suppose observed N sequences $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ (each has L loci) independently and identically selected from unknown K subpopulations with mixing proportion $\boldsymbol{\eta}$ to be estimated, then the likelihood is

$$L(\boldsymbol{\theta}_K; \boldsymbol{x}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n; \boldsymbol{\theta}_K).$$

See Section 1.1 for construction.

5. In short, the log likelihood is

$$\begin{aligned}
\log L(\boldsymbol{\theta}_K; \boldsymbol{x}) &= \sum_{k=1}^{K} \log f(\boldsymbol{x}_n; \boldsymbol{\theta}_K) \\
&= \sum_{k=1}^{K} \log \left[\sum_{k=1}^{K} \eta_k p_{\boldsymbol{\mu}_k,\boldsymbol{x}_n}(t) \right] \\
&= \sum_{k=1}^{K} \log \left[\sum_{k=1}^{K} \eta_k \left(\prod_{l=1}^{L} \mathbb{P}_{\mu_{kl},\boldsymbol{x}_{nl}}(t) \right) \right] \\
&= \sum_{k=1}^{K} \log \left[\sum_{k=1}^{K} \eta_k \left(\prod_{l=1}^{L} e^{\boldsymbol{Q}_{\mu_{kl},\boldsymbol{x}_{nl}} t} \right) \right].
\end{aligned} \tag{2.3}$$

Equation (2.2) has similar structure as Equation (**??**). Therefore, the EM algorithm (Dempster *et al.*, 1977) can be applied to maximize Equation (2.3) as maximize Equation (**??**). Except the parameter space $\boldsymbol{\Theta}_K$ of Equation (2.3) where $\boldsymbol{\theta}_K$ belongs to is neither continuous nor discrete space since $\boldsymbol{x}_n$ and $\boldsymbol{\mu}_k$ are in a categorical space which yields a very different E- and M-steps.

2.1.1 The phyclust Package

The **phyclust** (Chen, 2011) is an R package fully implements phyloclustering with different configurations, EM algorithms, and incorporating several useful tools such as **ms** (Hudson, 2002)

for simulating phylogeny and **seq-gen** (Rambaut and Grassly, 1997) for simulating sequence with vary mutations based on phylogenies. The **phyclust** also provides functions for re-sampling sequences from predicted models for determining an appropriate number of subpopulations. Those functions are particular useful for Sections 2.2 and 2.3.

An example of **phyclust** performing phyloclustering is given in Figure 2.1. It plots an example data set, Pony 524 (Carpenter *et al.*, 2011), where 146 Equine Infectious Anemia Virus (EIAV) sequences (Leroux *et al.*, 2004) are in y-axis and 405 loci in x-axis. The top row is the consensus sequence, and only mutation sites are spotted for 146 sequences. Colors represent A, C, G, and T nucleotides. Three clusters fitted by a CTMC model are shown and common mutation locations and types are grouped.

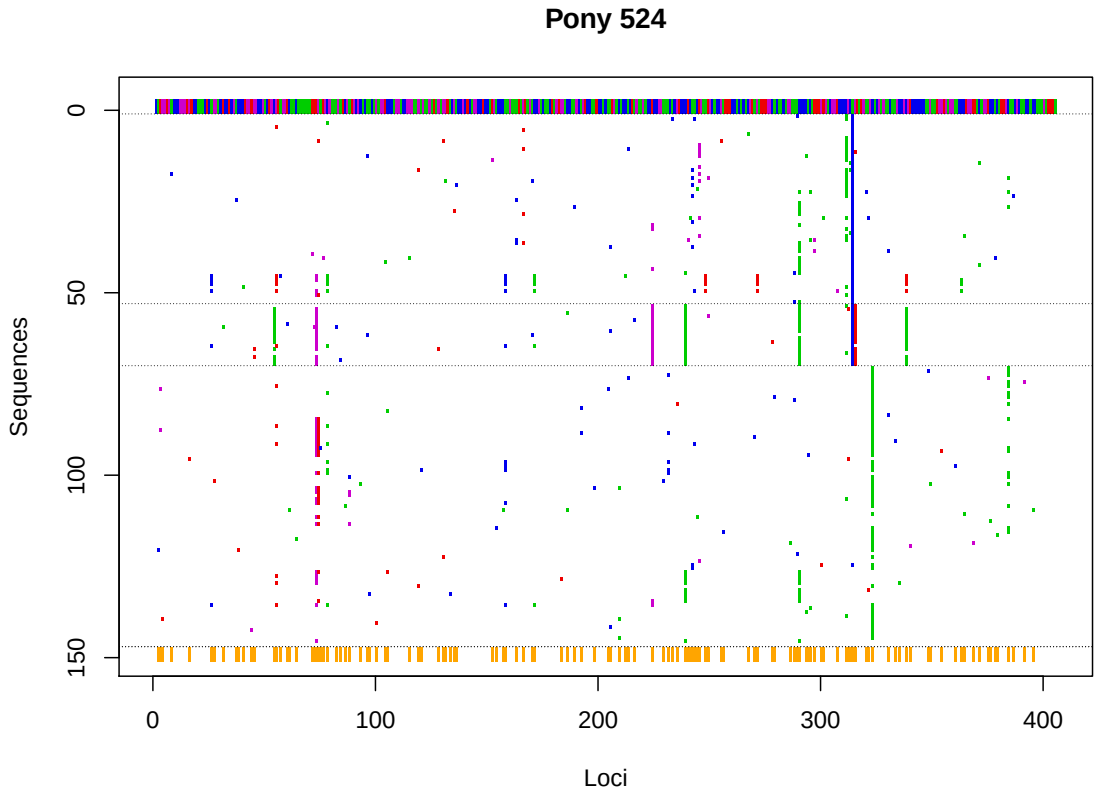


Figure 2.1: 146 EIAV sequences of Pony 524 in three clusters.

2.2 Bootstrap Method

"How many clusters are appropriate for the data?" is a typical question to any good scientists. There are several ways trying to infer this from data in Statistics via hypothesis testing. For example, $H_0 : K = 2$ v.s. $H_a : K = 3$ or more generally $H_0 : K = K'$ v.s. $H_a : K = K^*$ for any $K' \neq K^*$. In mixture models, the nested parameter space is inappropriate, hence, the LRT

introduced in Section 1.3 may not appropriate.

Bootstrap methods **??** may provide an adequate solution to rebuild an asymptotic distribution for the likelihood ratio (LR). The bootstrap methods is a re-sampling technique either from data (non-parametric) or from model (parametric) to form a distribution for (LR). Therefore, we may obtain a p-value by comparing LR to this distribution rather than deriving an asymptotic distribution from LRT.

Phyloclustering which uses a mixture models with unusual parameter space which is also particular suitable to apply the bootstrap methods to determine an appropriate number of subpopulations. For given data $\boldsymbol{X}$ and hypothetical K' and K^*, we may perform parametric bootstrap as the next.

Step 1: Based on $\boldsymbol{X}$, obtain MLEs $\hat{\boldsymbol{\theta}}_{K' ML}$ and $\hat{\boldsymbol{\theta}}_{K^* ML}$ under $\boldsymbol{\Theta}_{K'}$ and $\boldsymbol{\Theta}_{K^*}$, respectively.

Step 2: Compute and let $\hat{\lambda} := -2 \log \hat{\Lambda}(\hat{\boldsymbol{\theta}}_{K' ML}, \hat{\boldsymbol{\theta}}_{K^* ML}; \boldsymbol{X})$.

Step 3: Sample new data $\boldsymbol{X}^{(b)}$ from $\mathcal{F}(\hat{\boldsymbol{\theta}}_{K^* ML})$.

Step 4: Based on $\boldsymbol{X}^{(b)}$, obtain MLEs $\hat{\boldsymbol{\theta}}^{(b)}_{K' ML}$ and $\hat{\boldsymbol{\theta}}^{(b)}_{K^* ML}$ under $\boldsymbol{\Theta}_{K'}$ and $\boldsymbol{\Theta}_{K^*}$, respectively, via the EM algorithm.

Step 5: Compute and let $\lambda^{(b)} := -2 \log \hat{\Lambda}(\hat{\boldsymbol{\theta}}^{(b)}_{K' ML}, \hat{\boldsymbol{\theta}}^{(b)}_{K^* ML}; \boldsymbol{X}^{(b)})$.

Step 6: Repeat Steps 3 to 5 for B times, collect and let $\mathcal{F}^{(B)}(\lambda) := \{\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(B)}\}$ which is an approximation distribution to $\mathcal{F}(\lambda)$, the distribution of λ, as B large enough.

Step 7: If $\hat{\lambda}$ greater than $q_{\mathcal{F}^{(B)}(\lambda)}(0.95)$, then we reject the K' model under 0.05 level of type I error.

Unlike LRT of Section 1.3, note that $\hat{\boldsymbol{\theta}}_{K^* ML}$ is under $\boldsymbol{\Theta}_{K^*}$ rather than $\boldsymbol{\Theta}_{K'} \cup \boldsymbol{\Theta}_{K^*}$ nor $\boldsymbol{\Theta}_{K'+K^*}$.

2.3 Task Pull Parallelism

Obviously, Step 4 will be computationaly intensive as B increased, and no guarantee that each of $b = 1, 2, \ldots, B$ bootstrap sample will take similar time at obtaining MLEs. It may be possible to parallelize the EM algorithm fully in SPMD such as Section **??**, however, in general this step is still a bottleneck of whole computation.

The task parallelism as mention in Exercise **??** is one way to solve the problem by simply divided jobs equally likely to all processors. This is probably an optimal solution for equal loading jobs in homogeneous computing environment. However, it will be a terrible solution for unbalance loading jobs or in-homogeneous computing environment, such as bootstrap methods introduced in Section 2.2. Note that there are also some drawbacks for task parallelism:

- it requires a processor to handle job controls as the role of master in master/workers programming paradigm, and
- the code is not obviously and difficult to debug or generalize.

The website http://math.acadiau.ca/ACMMaC/Rmpi/examples.html has a general view of task parallelism and examples in **Rmpi**. Among three task parallel methods, task pull is the best performance and suit for bootstrap methods. We then introduce a simplified example of task pull in SPMD next.

2.4 Exercises

2-1 Argue that the instantaneously rate matrix $\boldsymbol{Q}$ of Equation 2.1 is positive definite. Therefore, argue that the eigenvlaues decomposition of $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ exists. Prove that $e^{\boldsymbol{Q}t} = \boldsymbol{U}e^{\boldsymbol{D}t}\boldsymbol{U}^{-1}$. Hence, this an easy way for computing transition probability $\mathbb{P}_{x,y}(t)$.

2-2 Argue that $\mathcal{F}^{(B)}(\lambda)$ is a good approximation to $\mathcal{F}(\lambda)$ in Step 4 of Section 2.2.

3

Bayesian MCMC

"No: I am not tired. I have a curious constitution. I never remember feeling tired by work, though idleness exhausts me completely."
– Sherlock Holmes

3.1 Introduction

To be done when WCC have time.

3.2 Gibbs Sampling

3.3 Metropolis-Hasting Algorithm

3.4 Quantum Monte Carlo

3.5 Exercises

Part II

Miscellany

References

Carpenter S, Chen WC, Dorman K (2011). "Rev Variation during Persistent Lentivirus Infection." *Viruses*, **3**, 1–11.

Casella G, Berger R (2001). *Statistical Inference*. 2nd edition. Cengage Learning.

Chen WC (2011). "Overlapping Codon model, Phylogenetic Cluetering, and Alternative Partial Expectation Conditional Maximization Algorithm." *Ph.D. Diss., Iowa Stat University*.

Chen WC, Ostrouchov G (2011). "HPSC – High Performance Statistical Computing for Data Intensive Research." URL http://thirteen-01.stat.iastate.edu/snoweye/hpsc/.

Dempster A, Laird N, Rubin D (1977). "Maximum Likelihood for Incomplete Data via the EM Algorithm (with discussion)." *Jounal of the Royal Statistical Society, Series B*, **39**, 1–38.

Ferguson TS (1996). *A Course in Large Sample Theory*. Chapman & Hall/CRC. ISBN 978-0412043710.

Fisher RA (1922). "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **222**, 309–368.

Golub GH, Van Loan CF (1996). *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. 3rd edition. The Johns Hopkins University Press.

Hudson R (2002). "Generating Samples under a Wright-Fisher Neutral Model of Genetic Variation." *Bioinformatics*, **18**, 337–338.

Leroux C, Cadoré JJ, Montelaro R (2004). "Equine Infectious Anemia Virus (EIAV): What has HIV's Country Cousin Got to Tell Us?" *Veterinary Research*, **35**, 485–512.

Rambaut A, Grassly N (1997). "Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution along Phylogenetic Trees." *Comput Appl Biosci*, **13**(3), 235–238.

Schmidt D, Chen WC, Ostrouchov G, Patel P (2013). "pbdDEMO: Programming with Big

Data – Demonstrations of pbd Packages." R Package, URL http://cran.r-project.org/package=pbdDEMO.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Index