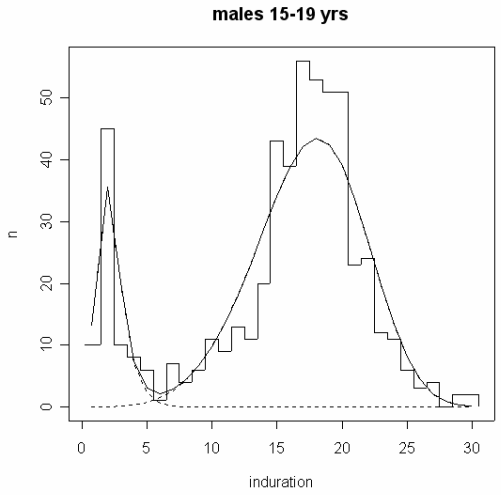
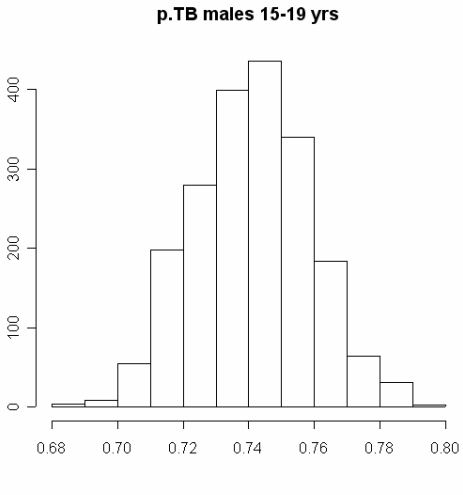
Bayesian Mixture Analysis for Tuberculin Induration Data

Beat Neuenschwander, PhD
July 2007
For the International Union against Tuberculosis and Lung Disease (The Union)

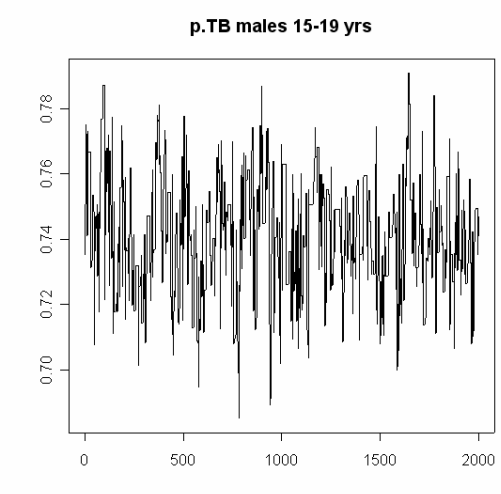
Data and Model Fit



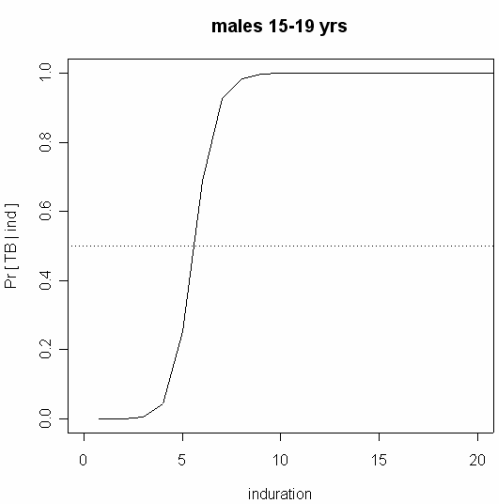
Prevalence of Infection



Sample Path



Probability of Infection



List of tables

List of figures

Preface

The mixture approach to the analysis of tuberculin induration data was introduced by Neuenschwander et al. (2000, 2002). This document provides the necessary information to perform mixture analyses based on an implementation of the Bayesian Markov Chain Monte Carlo approach in the programming environment R. The approach is exemplified with nine applications of increasing complexity.

For a first and basic understanding and outline of the statistical approach and implementation, the reader is referred to Sections 1 to 4. The mixture approach to analyzing induration data using the basic model is presented in Sections 5 and 7 (for unvaccinated subjects only), and 12 and 13 (for both unvaccinated and vaccinated subjects).

The more advanced (starred) Sections are Sections 8 and 9 (for the statistical aspects of mixture analysis and its MCMC implementation), and Sections 10 and 14 with extensions of the basic models.

More detailed information about the applications and the program are provided in the Appendix.

1 Mixture Analysis

1.1 Introduction

Mixture analysis provides a framework for analyzing data arising from different sub-groups. A characteristic of mixture analysis is the fact that it is generally not known to which sub-group an individual (item, data point) belongs. Questions of interest, depending on the context, are

1. What is the probability that an individual belongs to a specific sub-group (prevalence of sub-group)?
2. What do the distributions of the sub-groups look like?
3. Given the data on a specific individual, what is the probability that this individual belongs to a particular sub-group?
4. How many sub-groups are there?

Often, the number of sub-groups is known from the scientific context (typically a small number, e.g. 2). Moreover, the type of distribution for the sub-groups can be approximated by some well-known distribution (e.g. the normal distribution). If the data reflect these assumptions (showing, e.g., a bimodal distribution), estimation of mixture models, i.e. answering questions 1 to 4, is generally feasible.

1.2 Terminology for mixture models

We introduce some basic terminology for mixture models. Let us assume that individuals belong to one of two groups (group 1 or 2), and let the probability of belonging to these groups be

$$\Pr[Y=1] = p_1, \qquad \Pr[Y=2] = p_2 = 1- p_1,$$

where Y is a variable denoting (unobserved) group-membership. If an individual is from sub-group 1, his or her measurement X is distributed according to a distribution f_1

$$\Pr[X=x|Y=1] = f_1(x).$$

Note that in this notation everything to the right of the vertical bar is assumed known, and $\Pr(X=x|Y=1)$ is read as “the probability that X takes on the value x given Y has value 1”. In the same way, for an individual from sub-group 2,

$$\Pr[X=x|Y=2] = f_2(x).$$

Note that $f_1(x)$ and $f_2(x)$ are conditional distributions (conditional on group-membership Y), they are called *mixture component* distributions. The unconditional distribution of individual measurements, the *mixture distribution*, can be calculated according to the above quantities as follows:

$$\Pr[X=x] = m(x) = p_1 f_1(x) + (1-p_1) f_2(x).$$

Hence, m(x) is a mixture of f_1 and f_2 with mixture weights p_1 and $1\text{-}p_1$. Individual probabilities of being in sub-group 1 (given measurement x) follow according to

$$\Pr[Y=1|X=x] = p_1 f_1(x) / (p_1 f_1(x) + p_2 f_2(x)) = p_1 f_1(x) / m(x),$$

and $\Pr[Y=2|X=x] = 1\text{-}\Pr[Y=1|X=x]$. Note that all these equations follow from the basic probability calculus.

Statistical issues arise due to the fact that some of these quantities are unknown, i.e., the mixture weight p_1 and the parameters of the mixture component distributions f_1 and f_2 (e.g. mean and standard deviation of a normal distribution) have to be estimated. Therefore, for a two-component normal mixture model, the model has 5 unkwown parameters: mean and standard deviations from the 1st and 2nd group, μ_1,σ_1, μ_2,σ_2, and the mixture probability p_1. If the data (displayed as a frequency distribution) do not indicate bimodality, or if one of the sub-groups has a small mixture weight, estimation of mixture model parameters can be difficult.

1.3 Mixture analysis of tuberculin induration data

Tuberculin induration data fulfill the essential characteristics of the mixture analysis set-up. Subjects are either not reacting, reacting due to infection with *Mycobacterium tuberculosis*, or they react due to cross-reactions arising from environmental mycobacteria. The basic components of the mixture model for tuberculin induration data are:

$$\Pr[Y=0] = p_0 , \qquad \Pr[Y=1] = p_1 , \qquad \Pr[Y=2] = p_2 = 1 - p_0 - p_1 .$$

$$\Pr[X=x|Y=1] = f_1(x) , \qquad \Pr[X=x|Y=2] = f_2(x) ,$$

where Y=0, Y=1, and Y=2 refer to no reaction (zero induration), infection with *Mycobacterium tuberculosis*, and cross-reactions due to environmental mycobacteria, respectively. The corresponding prevalences are p_0 , p_1 , and p_2, respectively. Note that although there are three components present, since the first component is the group with zero induration, the mixture problem is essentially reduced to a two-component mixture problem.

2 Bayesian Statistics

The statistical analysis of the mixture models will proceed along the Bayesian approach. Bayesianism relies on the postulate that probabilities should be used to describe uncertain or unknown quantities, whether or not they are fixed or random. Based on observed quantities (the data D), statements about unobserved or unobservable quantities, θ, are derived according to a three-step program, its ingredients being

1. A *statistical model* is needed that relates the observed and unobserved part, i.e.,

   $$\Pr[D \mid \theta].$$

   This is the *likelihood function*, specifying the likelihood of the data for given θ.

2. Information (uncertainty) about the unknown parameter θ before observing the data D is specified by the *prior distribution* of θ, i.e.,

   $$\Pr[\theta].$$

3. Information (uncertainty) about θ after taking the data D into account is fully captured by the conditional distribution of θ given D, i.e.,

   $$\Pr[\theta|D].$$

   This is the *posterior distribution* of θ. The posterior distribution of each parameter (marginal distribution) is usually summarized by its mean (or median), standard deviation, and 95%- or 90%-credibility interval.

Once the statistical model and prior distribution have been specified, the Bayesian analysis is fully automatic due to the fact that the posterior distribution follows from Bayes theorem:

$$\Pr[\theta|D] = \Pr[D|\theta]\ \Pr[\theta] / \Pr[D].$$

Therefore, the Bayesian approach is straightforward, at least in principle. There is one major complication though, and this is the computation of Pr[D]. This (marginal) probability of the data is given by

$$\Pr[D] = \int \Pr[D|\theta]\ \Pr[\theta]\ d\theta.$$

Typically, this integration cannot be solved analytically. Hence, alternative (approximate) solutions have been developed, such as numerical integration, asymptotic approximations, and simulation (Monte Carlo) methods. Among the latter, the most recent Markov Chain Monte Carlo (MCMC) methods are the most promising ones, made possible by methodological breakthroughs and a parallel rapid progress in computing power since the mid 1980s.

Our analyses will be based on the MCMC approach. A Metropolis sampler will be used to simulate from the posterior distribution of mixture model parameters.

3 Software

3.1 A brief introduction to R

R is a general software for numerical and statistical computing. It is almost identical to the S and commercial S-Plus software. What follows is an absolute minimum the user needs to know when using R.

3.1.1 Installation

1. The software R can be downloaded for free from the R homepage:
   http:/lib.stat.cmu.edu/R/CRAN
   Detailed descriptions on how to download R for various platforms are given. First go to **R binaries**. If you use Windows, go to **Windows**, then to **base**, and download the latest **.exe** file.
2. For installation of R, start the **.exe** file and specify a directory of your choice.

3. The R icon should then appear on your desktop: double-click on it to start R. If everything works properly you should now be in R:. The R prompt is represented by the symbol >.
4. An example: type **rnorm(5,10,2)** at the R prompt. This should generate 5 normal random numbers with mean 10 and standard deviation 2.

Type y <- 2 and enter. "<-" is used to assign values to a variable. Type y*y and enter.

Sometimes the cursor is not at the R prompt, or you can't see the R prompt, or there is already something written on the command line you want to get rid of. Use the **Esc** key to get at the R prompt with an empty command line. The **Esc** key can also be used to interrupt a running program.

3.1.2 Setting a working directory

It is not necessary to set a working (project) directory, but it might be convenient to do so. To set a working directory, right click on your R shortcut, and choose a directory of your choice under **Properties**. There are advantages in doing this: you do not have to specify the absolute path when working in R. To summarize, it might be convenient to specify various R shortcuts on your desktop (or in in the project folder) for different projects.

3.1.3 Editing files

If you want to edit a file within R, e.g. the file `MyProgram.r` (assuming it exists), you can do this by typing

```
edit(file="c:\\….\\MyProgram.r")
```

on the command line. Be careful to specify the path correctly (note that that the double-backslash is used!), and note that R is case-sensitive. By default, the **Notepad** editor will be used. Of course, you can edit your files using your favourite editor by opening another window.
If you are using **Microsoft Word** for your R scripts, it is a good idea to change the fonts to **Courier New**.
If you need help on a special R command, e.g. `rnorm`, type

```
help(rnorm)
```

3.1.4 Running R interactively

For simple tasks it is convenient to run R interactively, i.e. from the command line. In the following example we will generate induration data for 200 subjects infected with Mycobacterium tuberculosis and 100 subjects with cross-reactions. For the TB data we type

```
TB <- rnorm(100,18,5)
```

on the command line. To see the data, type

```
TB
```

on the command line. Note that typing the name of an object is equivalent to printing it to the screen, i.e., `print(TB)` achieves the same as `TB`. To obtain the mean, median, and standard deviation of the TB data, use the command `mean`, `median`, and `sd`. For example,

```
mean(TB)
```

will print the mean.

Next, we generate the cross-reaction data by

```
CR <- rnorm(200,8,3)
```

Then, we combine the 300 observations into a data vector `Induration` as follows:

```
Indurations <- c(TB,CR)
```

Finally, we create a histogram:

```
hist(Indurations,breaks=seq(0,35))
```

The argument `breaks` is a vector of numbers from 0,1,2,…35 defining the intervals for the histogram. The histogram is shown in Figure 3-1.

The remaining Sections discuss a statistical approach (mixture analysis) that deals with the inverse problem, i.e., finding the underlying prevalence of TB for a given data set (or several data sets from various groups of subjects).

3.1.5 Running scripts

A script is a program file. Running (executing) such a program file (e.g. the file `MyProgram.r`) from the R prompt requires the following statement:

```
source("c:\\…\\MyProgram.r")
```

For example, you might want to put the previous commands

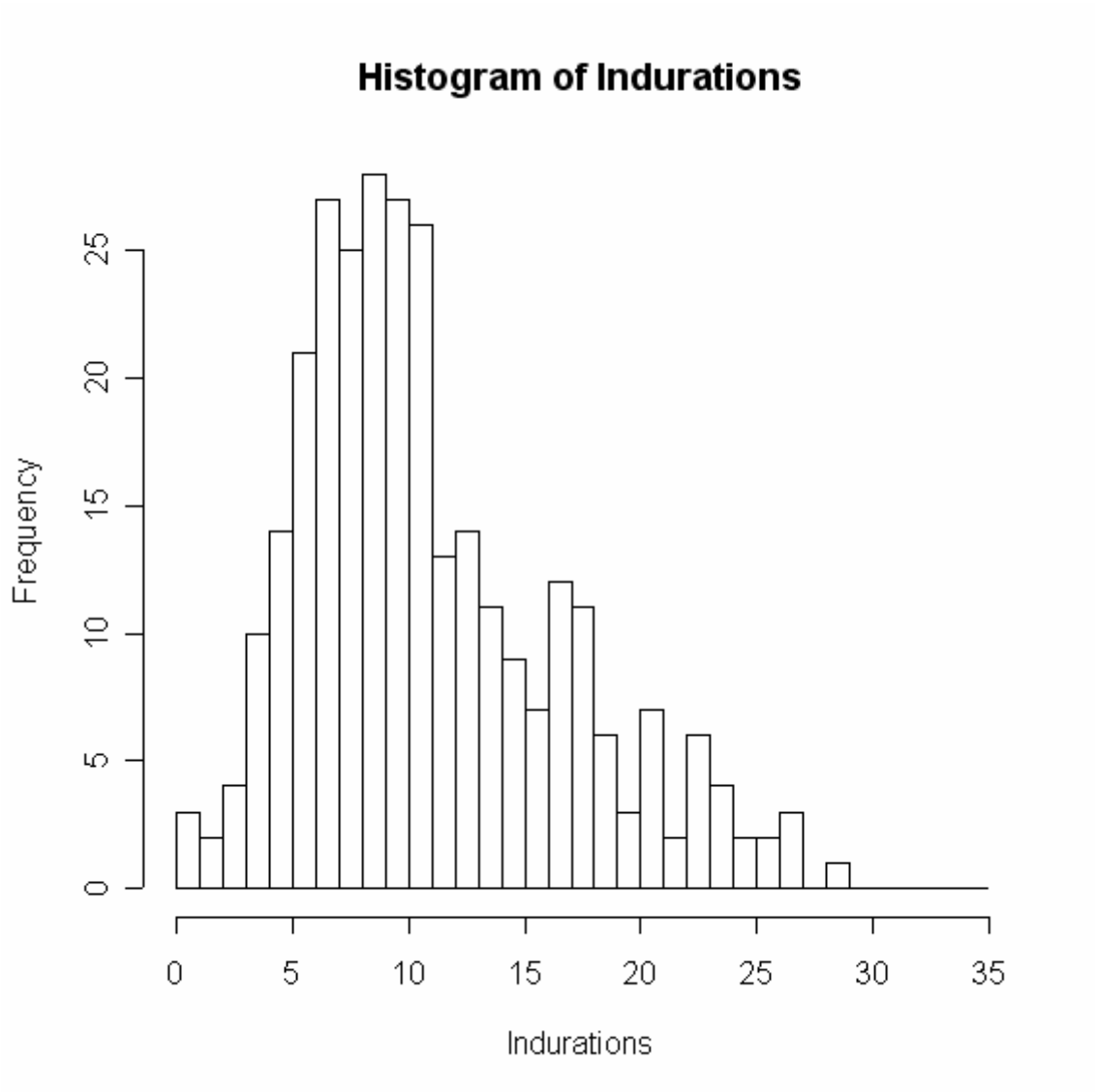
```
TB <- rnorm(100,18,5)
print(TB)
CR <- rnorm(200,8,3)
Indurations <- c(TB,CR)
hist(Indurations,breaks=seq(0,35))
```

into the file `MyProgram.r`, and execute the program via the `source` command given above.

If you edit a file within R, the file will be executed after you exit the file. If the file is not a program file; this will result in error messages (just ignore them).

The cursor-up and cursor-down keys can be used to obtain previously typed statements on the command line. So you don't have to retype previously used commands.

Figure 3-1 Histogram of simulated induation data



3.1.6 Output to screen

If you expect something to be written to the screen but nothing happens, check whether the **buffered output** option in the **Misc** menu (at top of screen) is disabled. If you still have nothing written to the screen, enter and execute **sink()** on the command line (this helps in case a program has been interrupted while writing results to an output file).

3.1.7 Getting help

If you know the name of a function you would like to use, but do not remember the exact syntax, use the help function. For example to get help on the normal random generator function, type

```
help(rnorm)
```

on the command line. You can also use the args function to get information on the arguments of the function

```
args(rnorm)
```

Quite often one does not remember the exact name of a function one would like to use. In these cases the help.search and apropos function might be of help. For example,

```
help.search("norm")
```

and

```
apropos("norm")
```

will be helpful to trace down the functions related to the normal distribution.

3.1.8 Quitting R

You can quit the program by either using the **File** menu or by typing

```
q()
```

on the command line. You have the option to save the current workspace image. If you do that, all commands and programs will be kept in memory, so you don't have to retype everything the next time you use R. The workspace can be saved at any time (see the **File** menu) during your R session.

3.1.9 R-homepage

For more information (including introductions to R, and other application packages) the user should visit the R-homepage:

http:/lib.stat.cmu.edu/R/CRAN

3.2 Mixture Programs: MS.r, NONBCG.r, BCG.r, IndurationPlot.r

All programs that are used for the different mixture models rely on the program **MS.r**, **NONBCG.r**, **BCG.r**, and **IndurationPlot.r**.

Note: Do not change theses files!

The use and meaning of these programs will become clear in later Sections. At this moment, just make sure that these files are available, either in your working directory or in another directory of your choice.

4 Getting Started: Application 1

We start with a first application using mixture analysis for a single frequency distribution. In this application only the main parts of the analysis will be shown. More details and further analyses will be given in the upcoming Sections, and in the Appendix.

Induration data from six age groups are shown in Figure 4-1. For the moment, let's assume that we are only interested in the analysis of males of age 5 to 9 years, shown in the second panel of the Figure. To perform the mixture analysis for this age group, some basic information needs to be put to an **Input Program**, e.g.

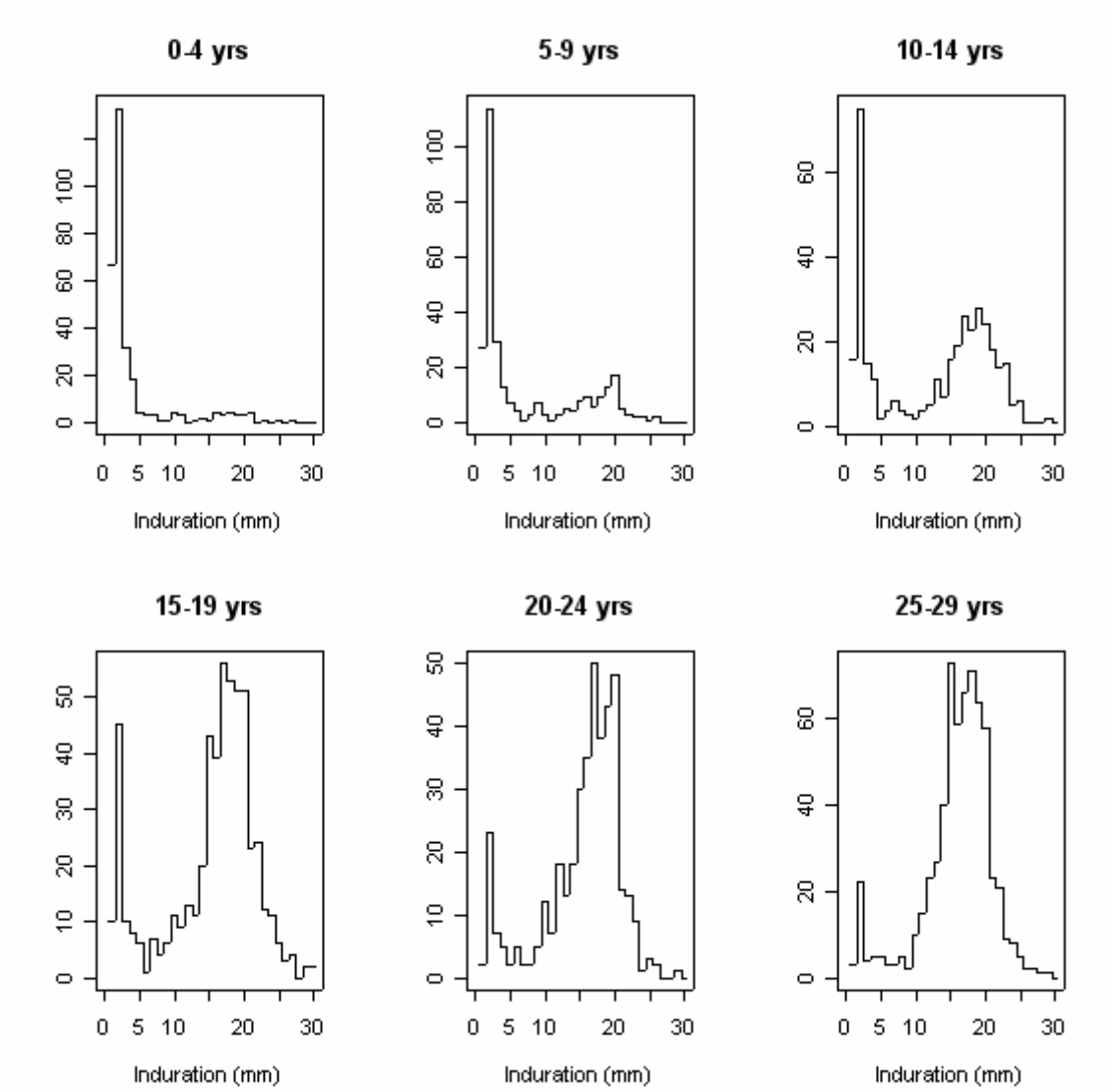
- the name of the input data file
- the name of the output file
- the type of mixture component distributions (model specification),
- some basic information for the estimation algorithm (the Metropolis sampling program).

4.1 Input program

For the analysis of induration coming from a single group, induration data must be a vector of frequencies for indurations 0mm, 1mm, …, 30mm. The vector consists of 31 entries. For example, the induration frequencies from the 3rd column of the Korean data set **korea75m.asc** (see Appendix 16.1.1), corresponding to males of age 5-9 years, are as follows:

```
282 27 114 29 13 7 4 1 3 7 3 1 3 5 4 8 9 6 9 13 17 5 3 2 2 1 2 0 0 0 0
```

Figure 4-1 Tuberculin induration data: males (Korea, 1975)



The following code needs to be put into an input program file (see **App1.r,** Appendix 16.2.1). Any editor can be used to do this.

```
infile <- "korea75m.asc"
freq.column <- 3

outfile <- "App1Out"      # optional argument

MS.run <- T               # optional argument
MS.results <- T           # optional argument
MS.check <- T             # optional argument
MS.graph <- T             # optional argument

MS.size <- c(2000,2000,1) # optional argument

distTB <- "Wb"
distCR <- "LN"

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\NONBCG.r")
```

Here are some explanations:

- `infile <- "korea75m.asc"`

A string (character variable) defining the input data file: `infile` is a simple ASCI or text file with 31 rows (corresponding to indurations 0mm, 1mm, …, 30mm), and an arbitrary number of columns (with frequencies for various groups of individuals). If there is only one group, the input file consists of 31 induration frequencies.

- `freq.column`
  the column number specifying the data (frequencies) to be analyzed.
- `outfile`
  a string (character variable) defining the output file containing the main results of the analysis (parameter estimates, credibility intervals, model checks, etc). Three output files will be created with
  - the main results of the analysis in output file 1
  - information of induration frequencies and fitted frequencies in output file 2 (this information can be used for creating graphical output)
  - information regarding model checks in output file 3

  In our example the three output files will be labeled **App1Out1.txt, App1Out2.txt, App1Out3.txt.**
- `MS.run, MS.results, MS.check, MS.graph`

  These variables can be either T (TRUE) of F (FALSE).
  `MS.run:` running the estimation algorithm (simulation program). This can be set to F if the simulation has been done already and you want to redo some of the other parts, like creating graphs.
  `MS.results:` if T, this will process the simulated values into a summary sent to the output file 1**.**
  `MS.check:` if T, posterior predictive model checks will be performed (see Section 8)
  `MS.graph:` if T, various figures will be produced
- `MS.size`
  Specifies the length of the Metropolis sampler (simulation). **MS.size** will be discussed later (Section 9)
- `distTB, distCR`

the parametric distribution representing the distribution of TB infections and cross-reactions. The options are:

- **Wb** for Weibull distribution,
- **LN** for log-normal distribution,
- **N** for normal distribution.
  It is recommended not to use **N** for the distribution of crossreactions due to the fact that the support of the normal distributions covers negative values.

The final two lines of the input program invoke the Metropolis Sampling program MS.r, and the mixture analysis program NONBCG.r. Note that the correct path of these files needs to be specified in the source command. If the lines of code shown above are in the file **App1.r**, say, the analysis can then be executed via

```
source(file="App1.r")
```

on the R-command line. This assumes that this file is in the current working directory. If not, the full path must be given.
Make sure that the **buffered output** option (in the **Misc** menu) is disabled, so that information about the sampling is sent to the screen. Note that it takes some time to run the simulation program.

4.2 Results

The main results of the analysis for males of age 5 to 9 years are presented in the first output file (Table 4-1). The results are as follows:

Table 4-1 Parameter estimates for Application 1

```
Prevalence of TB infections (p.TB)
mean st.dev  2.5%   50% 97.5%
0.186 0.0162 0.156 0.185 0.221

Prevalence of no reaction (p.zero)
mean st.dev  2.5%   50% 97.5%
0.485 0.0194 0.446 0.484 0.525

TB distribution (1st quantile, TB.qnt1)
mean st.dev 2.5%  50% 97.5%
16.6  0.593 15.3 16.7  17.8

TB distribution (2nd quantile, TB.qnt2)
mean st.dev 2.5%  50% 97.5%
23.8  0.729 22.6 23.8  25.3

CR distribution (1st quantile, CR.qnt1)
mean st.dev 2.5%  50% 97.5%
2.13 0.0677 2.01 2.13  2.28

CR distribution (2nd quantile, CR.qnt2)
mean st.dev 2.5%  50% 97.5%
3.97  0.243 3.53 3.96  4.52
```

The prevalence estimate of infection with *Mycobacterium tuberculosis* is 0.186 (posterior mean), with a 95% confidence interval ranging from 0.156 to 0.221. Note that the posterior median is 0.185, very close to the posterior mean. These two quantities are the main point

estimates considered in Bayesian analyses. The posterior standard deviation is 0.016 and can be seen as an analogue to the classical standard error of an estimate.

Additional information regarding the two quantiles of mixture component distributions are presented as well. For example, the estimated median (1st quantile) of the TB component distribution is 16.6, with a confidence interval ranging from 15.3 to 17.8.

5 Mixture Analysis for Induration Data

In this Section we will discuss some of the potential issues arising in mixture analyis.

5.1 The need for visualization

An important first step is to visualize the data, for example by a histogram representing the frequencies of indurations. The function **IndurationPlot** can be used to do that. It has to be loaded with

```
source(file=``c:\\...\\IndurationPlot.r'')
```

For example, to plot the histograms for the 6 groups in the file `korea75m.asc`, the following function call is needed:

```
IndurationPlot(infile="korea75m.asc",freq.column=2:7)
```

The arguments of the function `IndurationPlot` are as follows

- `infile:` the file name of induration frequencies
- `freq.column:` a vector of column indices denoting the selection of groups to be plotted.
- `header:` a logical (`TRUE` or `FALSE`) indicating whether the data file (infile) contains a header or not. The default is no header.
- `group.labels`: a vector of group labels (character strings) for the groups selected in `freq.column`.
- `page.layout:` a vector of two numbers with the number of rows and columns for the layout. For example, `page.layout = c(3,2)` will plot 6 histograms in 3 rows and 2 columns.
- `perc:` a logical (`TRUE` or `FALSE`) indicating the scale on the y-axis (percentages or absolute frequencies). The default is `perc=F`, i.e., absolute frequencies will be displayed.
- `xlab, ylab:` the labels on the x- and y-axis. The defaults are `xlab="Induration (mm)"` and no label for the y-axis.
- `zeromm:` three different options are available for displaying the number of zero indurations:

- "no": zero indurations are not included in the graph. This is the default.
- "yes": zero indurations are included in the graph
- "text": The number of zero indurations is not shown in the histogram but given as a text message in the figure.

- `plot.type`. Three options are available; "histogram" (the default), "polygon", or "polygon.points".
- `pch`: the symbol type (only if `plot.type="polygon.points"`). The default is `pch=16`, a solid circle.

5.2 Important aspects of mixture analysis for induration data

The following aspects are important when considering mixture analysis for induration data:

1. the sample size (number of subjects in the analysis). A relatively large (N=2000) and small (N=200) sample size will be considered.
2. the prevalence of infection with *Mycobacterium Tuberculosis*. Prevalences of 5%, 20%, 40% and 80% will be considered.
3. the amount of overlap for the mixture component distributions (TB and CR distribution). The median of the TB distribution will be 17mm, whereas for the distribution of cross-reactions, we will choose a median of 4mm and 8mm, implying small and large overlap with the TB distribution, respectively.

Table 5-1 Induration Data: Simulated Examples

	TB(50%) = 17mm, TB(95%) = 25mm	
	CR(50%) = 4mm, CR(95%) = 8mm	CR(50%) = 8mm, CR(95%) = 16mm
N = 2000	Example 1a	Example 1b
N = 200	Example 2a	Example 2b

1. Example 1a: Large sample size, little overlap of distributions.
   From Figure 5-1 we see that discrimination of the two mixture component distributions is relatively easy, if the prevalence of TB infections is not too small.
2. Example 1b. Large sample size, considerable overlap of distributions.
   If the overlap of the TB and CR distribution is considerable, discrimination of the two is more difficult. If we have data from different groups with both small and large prevalences of TB infection, estimation will become easier since mixture component distributions can be inferred for cross-reactions (if TB prevalence is small) and TB infections (if TB prevalence is high). Otherwise, due to the overlap of the distributions inference is more difficult.

3. Examples 2a and 2b. Small sample sizes.
   The above comments also apply for the smaller sample size, but the problems are becoming more difficult due to the fact that less information is available, and the figures become noisier. Another reason for noisy frequency distributions might be the result of poor induration measurements (e.g. digit preference).

To conclude, mixture analysis will become difficult if

- CR and TB component distributions overlap considerably
- neither the CR nor the TB distribution can be inferred from the data due to considerable mixture of the two components (intermediate prevalences of TB infection)
- small sample sizes resulting in noisy frequency distributions
- poor indurations measurements

These factors will usually not make mixture analysis infeasible, but estimates will typically be less precise.

Figure 5-1 Induration Data: Example 1a (n=2000)

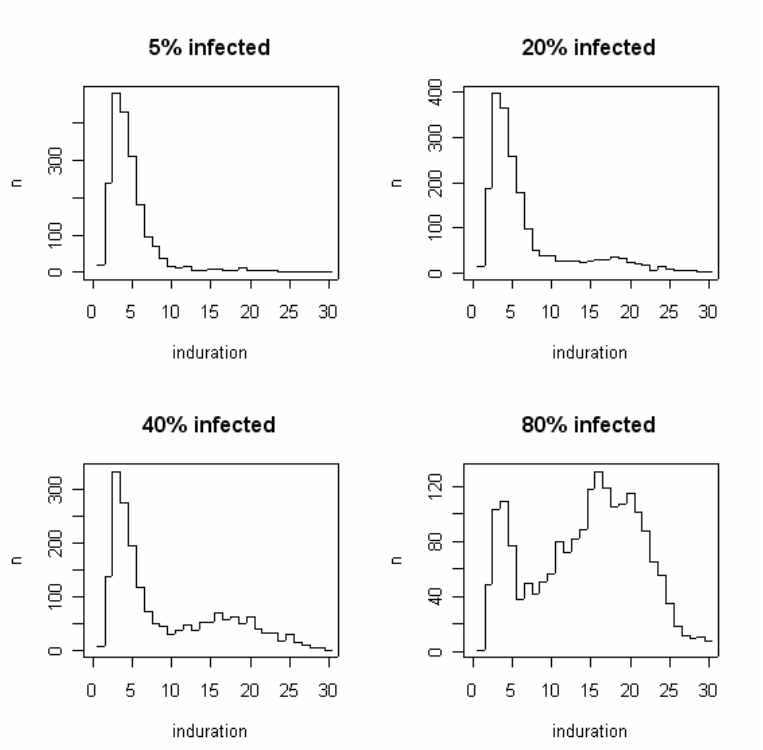


Figure 5-2 Induration Data: Example 1b (n=2000)

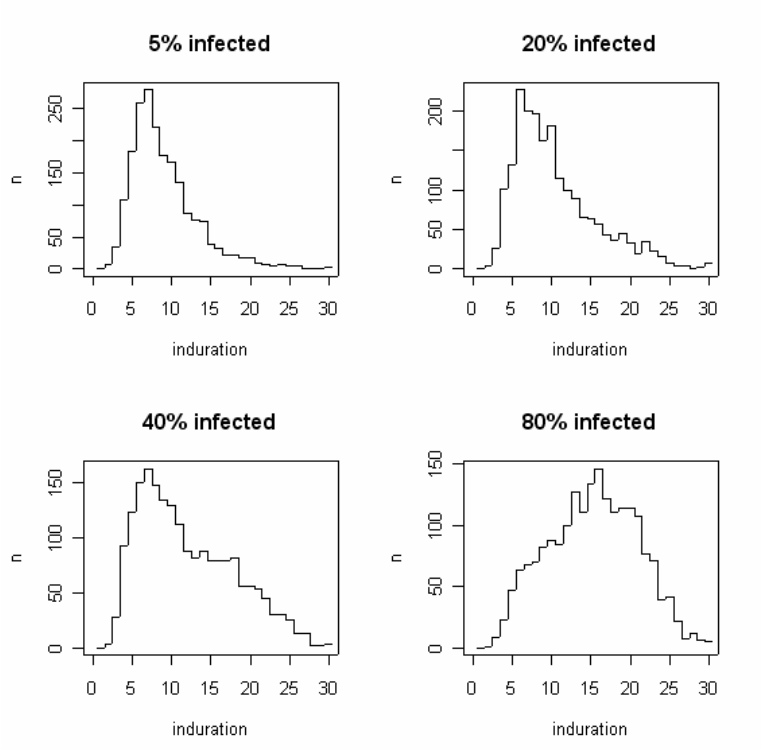


Figure 5-3 Induration Data: Example 2a (n=200)

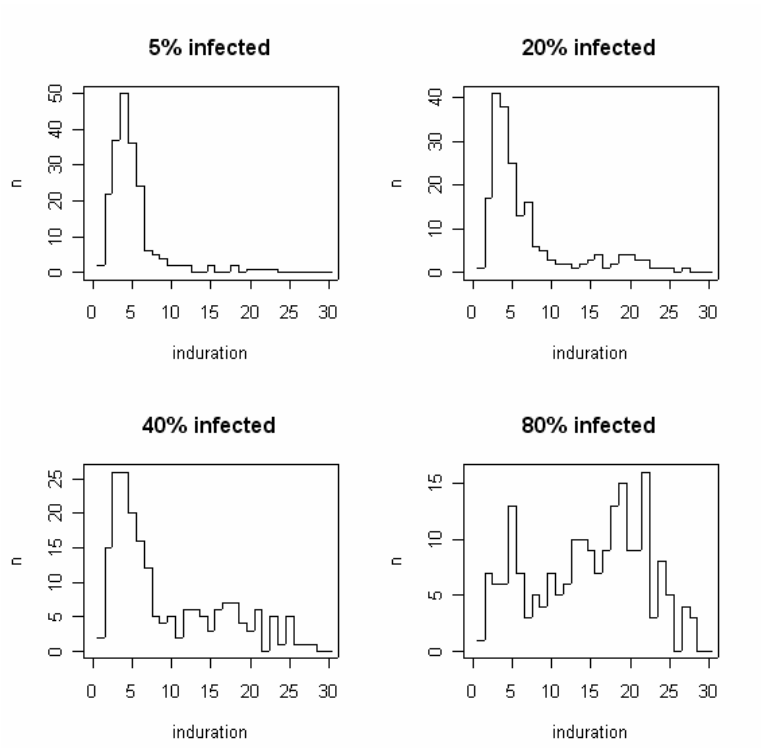
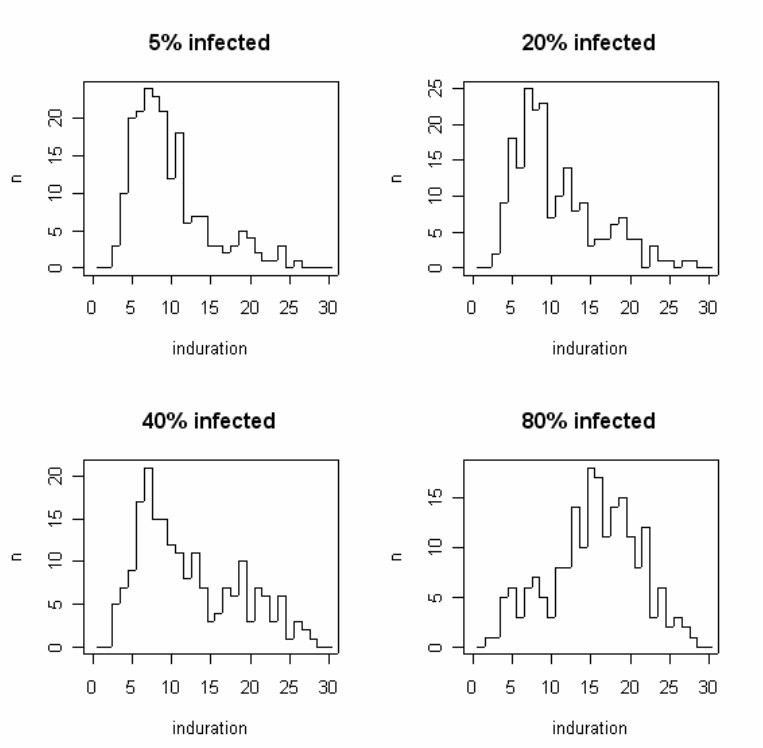


Figure 5-4 Induration Data: Example 2b (n=200)



6 Program Output

The program will produce different outputs

- on-line output while the program is running
- a log-file
- 3 output files containing results from mixture analysis
- graphical output
- an R-object containing all relevant information from the analysis

6.1 On-line output to screen

While the estimation algorithm is running, information is sent to the screen. This is helpful in case the algorithm is running for a considerable amount of time which is the case for more complex models (see later Sections). The information sent to the screen is the following:

- The actual iteration number. If this number is negative, the algorithm is still in the so-called burn-in phase. This means that the algorithm is still adapting and preparing for the so-called sampling phase from which results will be taken for the final inference.
- The number of iterations saved so far. Note that while the algorithm is in the burn-in phase, iterations are not saved.
- The approximate remaining time (in minutes) until the end of the algorithm
- Acceptance rates of the sampler. These values will usually be around 0.2 to 0.5, except in the very beginning of the algorithm. In case of convergence problems of the algorithm, some of these values might get close to 0 or 1. This should be considered as a warning that the algorithm behaves poorly.
- Log-likelihood, log-posterior values. These values usually increase during the burn-in phase of the algorithm, and stabilize when the sampling algorithm has reached its stationary phase.
- Current parameter values

Two examples are shown in Tables 6-1 and 6-2. In Table 6-1, the sampler is still in the burn-in phase (negative iteration, no saved values so far). In Table 6-2, iteration number 673 (out of 2000) has been reached, and the corresponding number of iterations has been saved.

Table 6-1 Output to screen during burn-in phase

```
iteration -151 / 2000 ,    save no. 0 of 2000 ,     remaining time [min]   0.5
acceptance rate(s) per block 0.391 0.382 0.341 0.368 0.35 0.355
log-likelihood, log-posterior, log-prior:
   actual:    -26.58358 -26.58358 0
   maxima:    -24.73345 -24.73345
$p.TB
```

```
[1] 0.182

$p.zero
[1] 0.479

$TB.qnt1
[1] 16.6

$TB.qnt2
[1] 23.4

$CR.qnt1
[1] 2.22

$CR.qnt2
[1] 4.41
```

Table 6-2 Output to screen after burn-in phase

```
iteration 673 / 2000 ,    save no. 672 of 2000 ,     remaining time [min]   0.3
acceptance rate(s) per block 0.308 0.334 0.263 0.285 0.259 0.279
log-likelihood, log-posterior, log-prior:
   actual:    -25.63721 -25.63721 0
   maxima:    -24.66511 -24.66511
$p.TB
[1] 0.197

$p.zero
[1] 0.49

$TB.qnt1
[1] 16.8

$TB.qnt2
[1] 23.6

$CR.qnt1
[1] 2.13

$CR.qnt2
[1] 4.11
```

6.2 Log-file

The log-file contains information about the steps performed in the analysis. This can be helpful for diagnosing program errors. A typical example is shown in Table 6-3.

Table 6-3 Output log-file

```
Program NONBCG (version 3.1, December 2004)
Tue Jan 11 08:56:22 2005
Preparatory steps
 ...Load and process data
 ...Set tuning parameters for Metropolis Sampler
 ...Load information from former analysis
 ...Set values for parameter constraints
 ...Load functions
 ...Create grouping information
 ...Prepare checks for component distributions
 ...Create initial values for model parameters
 ...Check initial model parameters
 ...Boundaries for parameters
 ...Create labels
MS.run: Start Metropolis Sampler
MS.results: Processing results
 ...Loading sampled values
```

```
 ...Basic input for analysis
 ...Parameter constraints
 ...Posterior summaries
 ...Maximum posterior estimates
 ...Total prevalences
 ...Fitted frequencies
MS.check: Predictive model checks
 ...Predictive summaries
MS.graph: Creating graphs
 ...MCMC plots
 ...Histograms for parameter estimates
 ...Graphs for probability of infection
 ...Data histograms and fits
Clearing Memory
Tue Jan 11 09:08:38 2005
```

6.3 Output files containing results from mixture analysis

Three output files with results will be produced

1. a file with the main results of the mixture analysises
2. a file with the estimated mixture and mixture component values
3. a file with model diagnostics

The files are names as follows: if **outfile** has name `xyz`, the output files will be names `xyzOut1.txt`, `xyzOut2.txt`, `xyzOut2.txt`, respectively.

6.4 Graphical Output

Graphical output is produced automatically (see the various examples in the document).

6.5 R-object

At the end of the program an R-object is created including all relevant information of the analysis. See Appendix 16.6 for details.

7 The Basic Model

7.1 Introduction

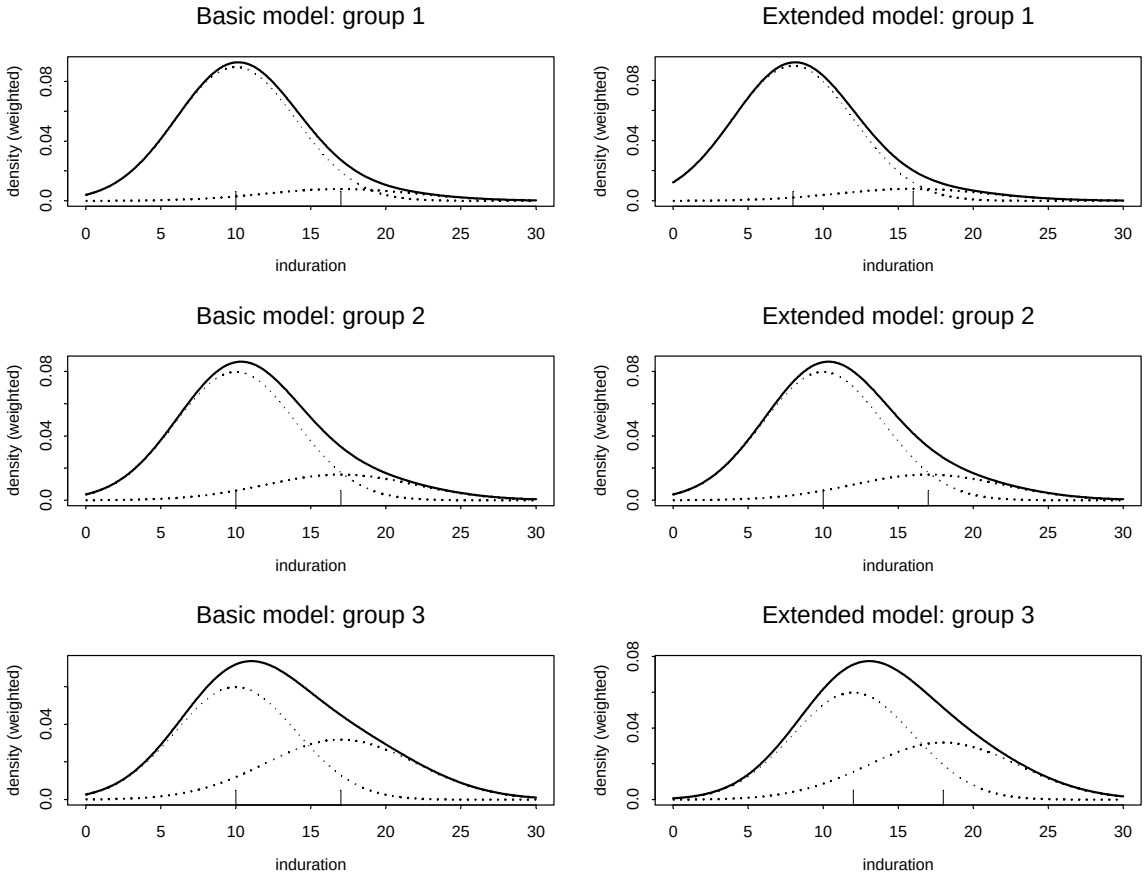
Induration from a single group has been analyzed in Application 1. Often, data from more than one group are available, so a generalization of the mixture analysis approach to several groups is needed, if one does not want to rely on stratified (by group) analyses. The basic mixture model assumes data from several groups with the assumption that mixture component distributions (TB and CR) are the same for all groups and that only prevalences differ among groups.

If the number of groups is *k*, there are *2*k+4* parameters to be estimated in this model: *k* prevalences of TB infection, *k* prevalences of non-reactors, and 4 parameters representing the two mixture component distributions.

The basic mixture distributions are displayed in Figure 7-1 (left panel), assuming 3 groups with prevalences of TB of 10%, 20%, and 40%, and normal component distributions with a common mean of 17mm (standard deviation 5), and 10mm (standard devation 4). Note that the (inside) tick marks refer to the means of the component distributions.

An extended model allowing for different mixture component distributions over the groups is displayed in the right panels of Figure 7-1. Here, the means increase (16, 17, 18mm for TB, and 8,10,12mm for CR) over the 3 groups. The extended model will be discussed in later Sections.

Figure 7-1 Basic and extended mixture model



7.2 Program input

The input for the analyses of several groups is very similar to the analysis for one group (Section 4). The differences are as follows:

- `freq.column` is now a vector of columns indices to be analyzed: e.g., if the data for the different groups are in columns 2 to 7, the following statements can be used:

  ```
  freq.column <- seq(2,7), or freq.column <- c(2,3,4,5,6,7)
  ```

  The command `seq(a,b)` creates a vector of integers from a to b. The `c()` command is the general command to specify a vector of values, e.g. **c(2,5,6)** is a 3-dimensional vector with values 2, 5, and 6.

- `group.names` is a vector of labels corresponding to the *k* groups, e.g.

  ```
  group.names <- c("males", "females")
  ```

  would be used for an analysis of two frequency distributions with data from males and females.

If no group names are specified, the groups will be labeled “Group 1”, “Group 2”, etc.

If the grouping structure is more complicated (more than one grouping variable), a different specification of `group.names` should be used.
For example, the groups might be set up by

- year of survey (1990, 1995),
- sex (males, females),
- age group (5-9 yr, 10-14 yr, 15-19 yr),

leading to a total of 12 groups. The specification of group.names should be set up as a list as follows:

```
group.names  <- list( c(“survey”, ”sex”, ”age”), c(“1990”, ”1995”), c(“males”, ”females”), c(“5-9”, “10-14”, “15-19”) )
```

where the first element gives the grouping variables, and the remaining elements specify the levels of each group. The columns in the data set have to be in the correct order, i.e.,

1990, males, 5-9
1990, males, 10-14
1990, males, 15-19
1990, females, 5-9
1990, females, 10-14
1990, females, 15-19
1995, males, 5-9
1995, males, 10-14
1995, males, 15-19
1995, females, 5-9
1995, females, 10-14
1995, females, 15-19

The order follows the order given in c(“survey”,”sex”,”age”), where the first element specifies the grouping variable that is changing the slowest, the last element being the one changing the fastest.

Due to the fact that the number of parameters can be considerable (depending on the number of groups to be analysed), the number of iterations (given by `MS.size`) required to obtain convergence of the sampler can be large. Therefore, the analysis can be time-consuming. In any case, convergence of the sampler needs to be inspected carefully (for details see Section 9). Moreover, the length of the burn-in period depends on the number of groups to be analysed.

7.3 Specification of initital parameter values

The estimation procedure starts with initial values for all model parameters. In the examples we have seen so far initial values were generated automatically. The user has the option to set initial values for model parameters by specifying values for

- `p.zero.init, p.TB.init:` initial values for prevalences
- `TB.qnt1.init, TB.qnt2.init, CR.qnt1.init, TB.qnt2.init:` initial values for the two quantiles of the mixture component distributions. Note that the default quantiles are the 50% and 95%-quantile which means that reasonable initial values for the TB distribution are (for example) 18mm and 25mm.

The initial values can be either scalars or vectors. In the latter case the vector must have length equal to the number of groups given in freq.column. The initial values can be vectors, but a scalar (a single number) can be used as well. In the latter case, the same initial value for each group will be used.

7.4 Application 2: Korea, males, all age groups

Application 2 provides the analysis of the six age groups (males) from the Korean data set. The basic model is assumed here, i.e., mixture components are assumed common across groups, only prevalences may be different.Note that the induration data for the six groups korea75m.asc are located in columns 2 to 7.

7.4.1 Input file

```
infile <- "korea75m.asc"
freq.column <- seq(2,7)

outfile <- "App2Out"

MS.run <- T
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247
Ms.size <- c(2000,2000,10)

distTB <- "Wb"
distCR <- "LN"

group.names <- c("0-4","5-9","10-14","15-19","20-24","25-29")

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\NONBCG.r")
```

Note the changes compared to application 1 in Section 4 (analysis of age group 5-9 years) are as follows:

- `freq.column` is a vector of group (column) indices in the data file
- a vector of group names has been specified (this is optional)
- `MS.size` has been changed: the burn-in is 2000 iterations. Then, from the 20000 sampled values, 2000 are saved (i.e., the 20000 iterations are thinned, thinning = 10); see Section 9.

7.4.2 Results

The TB prevalence estimates and the estimated medians of the TB component distribution are presented in Table 7-1 and Figure 7-2. Note that that the medians are the same for all groups (the basic model assumption). The TB prevalences increase form 0.040 (age group 0-4) to 0.922 (age group 25-29).

The estimated TB prevalence for age group 5-9 is 0.181 (95%-CI 0.149 to 0.217) which is slightly reduced compared to the estimate obtained in Application 1 (Section 4). This can be explained by the fact that in the present analysis, information regarding the mixture component distributions is shared across all age groups. This can lead to improved estimates (if the underlying assumption of common components distributions is valid).

The formal model check (see Table 7-2) reveals that 13.9% of the 180 fitted frequencies are predictive failures, i.e., observed frequencies lie outside the 95% prediction intervals. If the model would reflect the truth, one would expect about 5% predictive failures. Note that some of the predictive failures occur at 10mm (prediction too small), 14mm (too large), 15mm (too small), 20mm (too small), 21mm (too large), and 30mm (too small). This is an indication of digit preference for multiples of 5mm. Finally, from Figure 7-3 it can be seen that model estimates and data are in good agreement.

Table 7-1 Parameter estimates for Application 2

```
Prevalence of TB infections (p.TB)
mean  st.dev   2.5%    50%  97.5%
0-4   0.0396 0.00616 0.0287 0.0392 0.0529
5-9   0.1813 0.01689 0.1490 0.1805 0.2173
10-14 0.4981 0.02240 0.4587 0.4963 0.5473
15-19 0.7420 0.01604 0.7108 0.7426 0.7710
20-24 0.8629 0.01784 0.8288 0.8637 0.8969
25-29 0.9223 0.01093 0.9001 0.9231 0.9431

TB distribution (1st quantile, TB.qnt1)
mean st.dev 2.5%  50% 97.5%
0-4   17.3  0.105 17.1 17.3  17.5
5-9   17.3  0.105 17.1 17.3  17.5
10-14 17.3  0.105 17.1 17.3  17.5
15-19 17.3  0.105 17.1 17.3  17.5
20-24 17.3  0.105 17.1 17.3  17.5
25-29 17.3  0.105 17.1 17.3  17.5
```

Table 7-2 Model checks for Application 2

```
Summary of predictive checks (+ predictions too large, - predictions too small)
    13.9 % predictive failures

1: 0-4 2: 5-9 3: 10-14 4: 15-19 5: 20-24 6: 25-29

      1 2 3 4 5 6 total
```

1mm				+			1
2mm		-	-				2
3mm	+	+	+				3
4mm							0
5mm							0
6mm				-			1
7mm			-				1
8mm							0
9mm		-			+		2
10mm	-						1
11mm							0
12mm							0
13mm				+			1
14mm			+	+			2
15mm					-		1
16mm							0
17mm				-			1
18mm							0
19mm							0
20mm		-		-			2
21mm				+	+		2
22mm							0
23mm					+		1
24mm				+			1
25mm							0
26mm							0
27mm							0
28mm							0
29mm			-	-			2
30mm				-			1
total	2	4	5	4	6	4	25

Figure 7-2 Application 2: Estimates of prevalence of *Mycobacterium Tuberculosis*, and prevalence of zero reaction (males, 6 age groups)

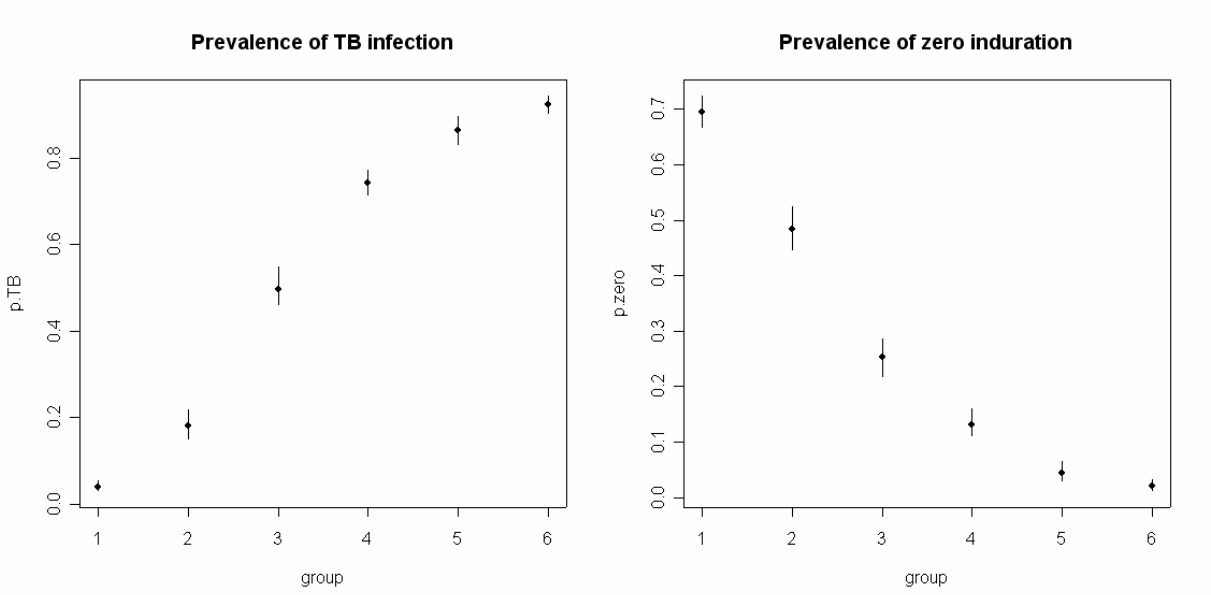
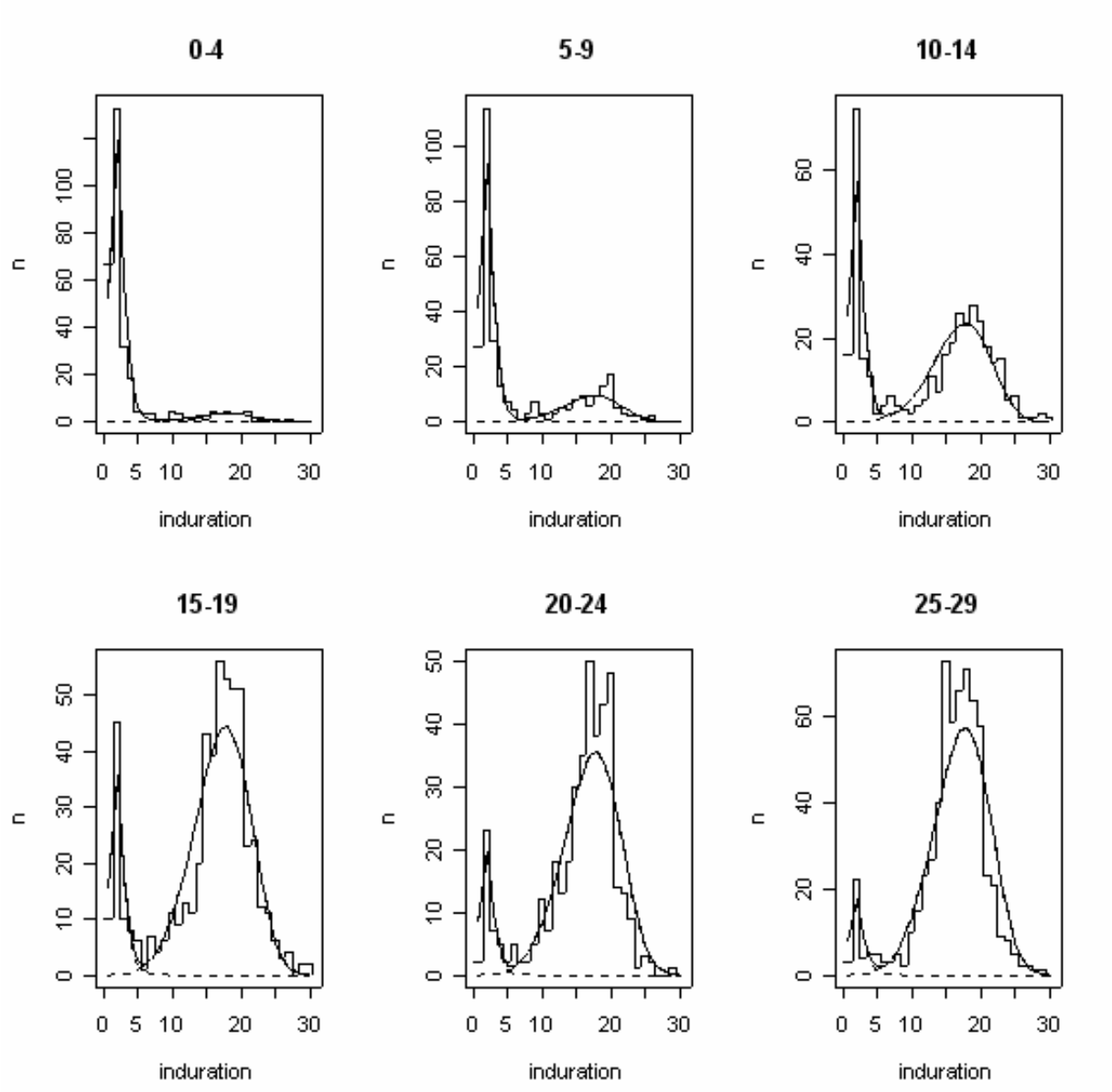


Figure 7-3 Application 2: Induration data and model fit



8 Model Selection and Model Checks*

Table 8-1 shows a summary of results for different models arising from Weibull, log-normal, and normal mixture component distributions. There are two models that stand out as good candidates for modeling the frequency distribution of males of age 5 to 9 years, with a Weibull or normal distribution for the distribution of TB infections, and a log-normal distribution for the distribution of cross-reactions. Both models are similar with respect to the approximate maximum of the log-likelihood function (and clearly better compared to the other models), and they perform fairly well with regard to the posterior predictive checks.

Table 8-1 Model selection: males of age 5-9 years (Korea, 1975)

Distributions: TB / CR	*Maximum of Log-likelihood*	*Number of predictive failures [%]*
LN / LN	-38.2	5 [16.7]
LN / Wb	-57.3	7 [23.3]
Wb / Wb	-40.3	4 [13.3]
Wb / LN	-24.7	2 [6.7]
N / LN	-26.5	2 [6.7]
N / Wb	-40.2	3 [10.0]

LN: log-normal, Wb: Weibull, N: Normal
Predictive failure if observed frequency is outside 95%-prediction interval (for indurations 1mm to 30mm).

It is important not to uncritically rely on one of these models. The optimal choice of distributions depends on the data to be analyzed. The maximum of the log-likelihood function can be used as a guide to choosing a suitable model. This does not guarantee that the fit of the chosen model is sufficiently good.

The quality of the fit should be assessed by comparing predicted and observed frequencies via posterior predictive model checks. For the models considered in Table 8-1, the Wb/LN and N/LN models show the best results with regard to posterior model checks: two out of the 30 frequencies are predictive failures. This corresponds to a predictive failure rate of 6.7%, which is close to the one expected if the model were true. The Wb/LN is slightly superior to the N/LN model with regard to the maximum of the log-likelihood. To conclude, the Wb/LN is the best model and provides a reasonably good fit to the data.

9 Tuning the Metropolis Sampler*

Metropolis sampling is a general iterative methodology (Markov Chain Monte Carlo) for doing Bayesian analyses. The main difficulty with this approach is that the algorithm needs to run for a sufficiently long time (burn-in period) before valid inferences can be drawn from the sampled values, i.e., convergence of the sampler has to be reached.

If the user is not familiar with Bayesian statistics and Markov Chain Monte Carlo methods, help from a professional statistician may be required.

Warning: if the Metropolis sampler has not converged, results will be unreliable!

MS.size consists of 3 numbers.

1. the length of the burn-in period
2. the length of the sample (number of values stored)
3. the thinning of the sample.

Ideas will be made clear with a simple example with only one frequency distribution. For this case, four examples will be considered

1. A short burn-in period (100 iterations), and a reasonably large sample size (1000). No thinning will be applied, i.e., all values from the sample will be used for the analysis.

```
MS.size <- c( 100, 1000, 1)
```

2. A reasonably large burn-in period (1000 iterations), and a small sample size (100). No thinning will be applied, i.e., all values from the sample will be used for the analysis. The total number of iterations is 1100.

```
MS.size <- c( 1000, 100, 1).
```

3. A reasonably large burn-in period (1000 iterations), and a reasonably large sample size (1000). No thinning will be applied, i.e., all values from the sample will be used for the analysis. The total number of iterations is 2000.

```
MS.size <- c( 1000, 1000, 1)
```

4. A reasonably large burn-in period (1000 iterations), and a reasonably large sample size (1000). A thinning of 10 will be applied, i.e., only values from every 10^{th} iteration will be used for the analysis. The total number of iterations is 11000.

```
MS.size <- c( 1000, 1000, 10).
```

Results from the analyses using the different choices for MS.size are displayed in Figures 9-1 to 9-4.

Example 1: the burn-in period is clearly too short. The algorithm has not yet converged, the deviance is still decreasing, and sampled parameter values have not stabilized yet.

Example 2: A burn-in of 1000 is clearly sufficient. The deviance has stabilized after approximately 300 iterations. However, note that a sample of size 100 is too small to draw valid conclusions about the parameter values. A small sample size is usually good enough for getting an approximate point estimate, but for standard deviations and confidence intervals, larger samples are needed. Also note that the sampled parameter values stay at the same value for several iterations: this is not a bug of the sampling program but rather a characteristic of the Metropolis algorithm. As a conclusion, the sample size needs to be increased considerably to obtain valid inferences for the model parameters.

Example 3: The same burn-in (1000) as in example 2 is used, and the sample size is increased to 1000. The results look okay now, and inferences could be drawn from this analysis.

Example 4: The sampling period is prolonged to 10000 iterations,

To conclude, assessment of convergence is critical to assure the validity of results. History plots of sampled values are a good diagnostic tool to assess the performance of the sampling algorithm. Ideally, the plots should look similar to the ones shown in examples 3 and 4.

Figure 9-1 Assessing convergence (Example 1): MS.size=c(100,1000,1)

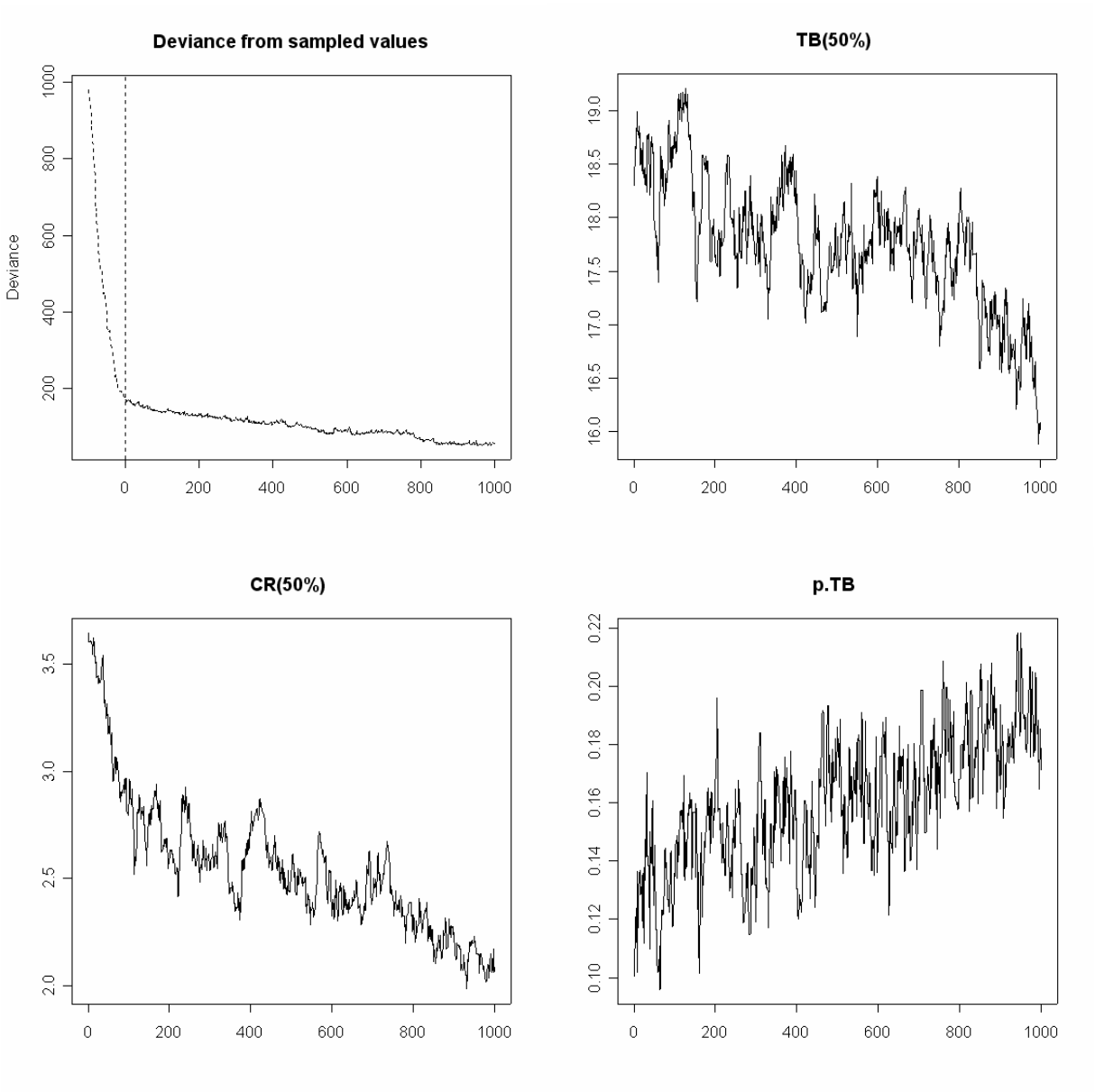


Figure 9-2 Assessing convergence (Example 2): MS.size=c(1000,100,1)

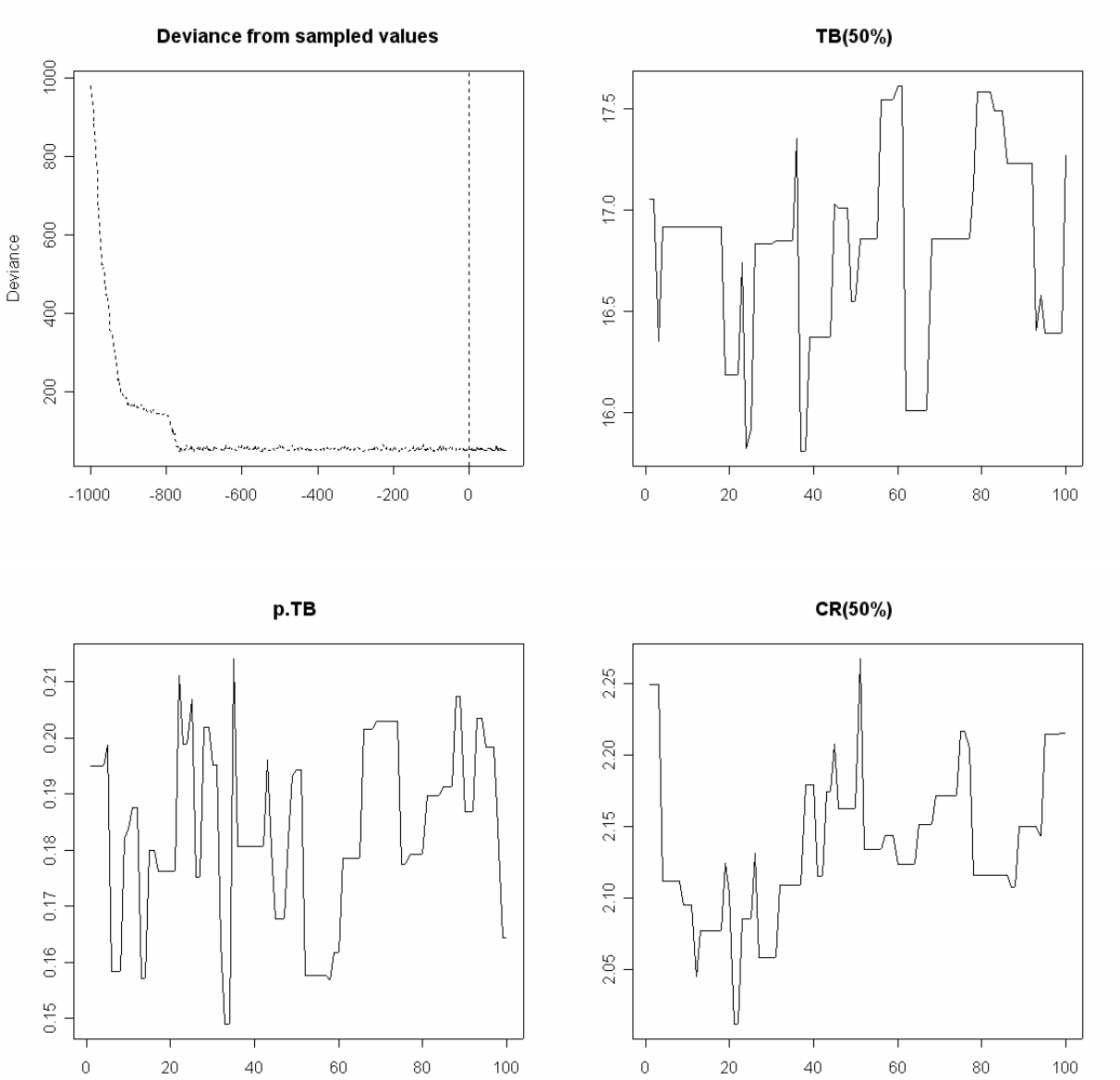


Figure 9-3 Assessing convergence (Example 3): MS.size=c(1000,1000,1)

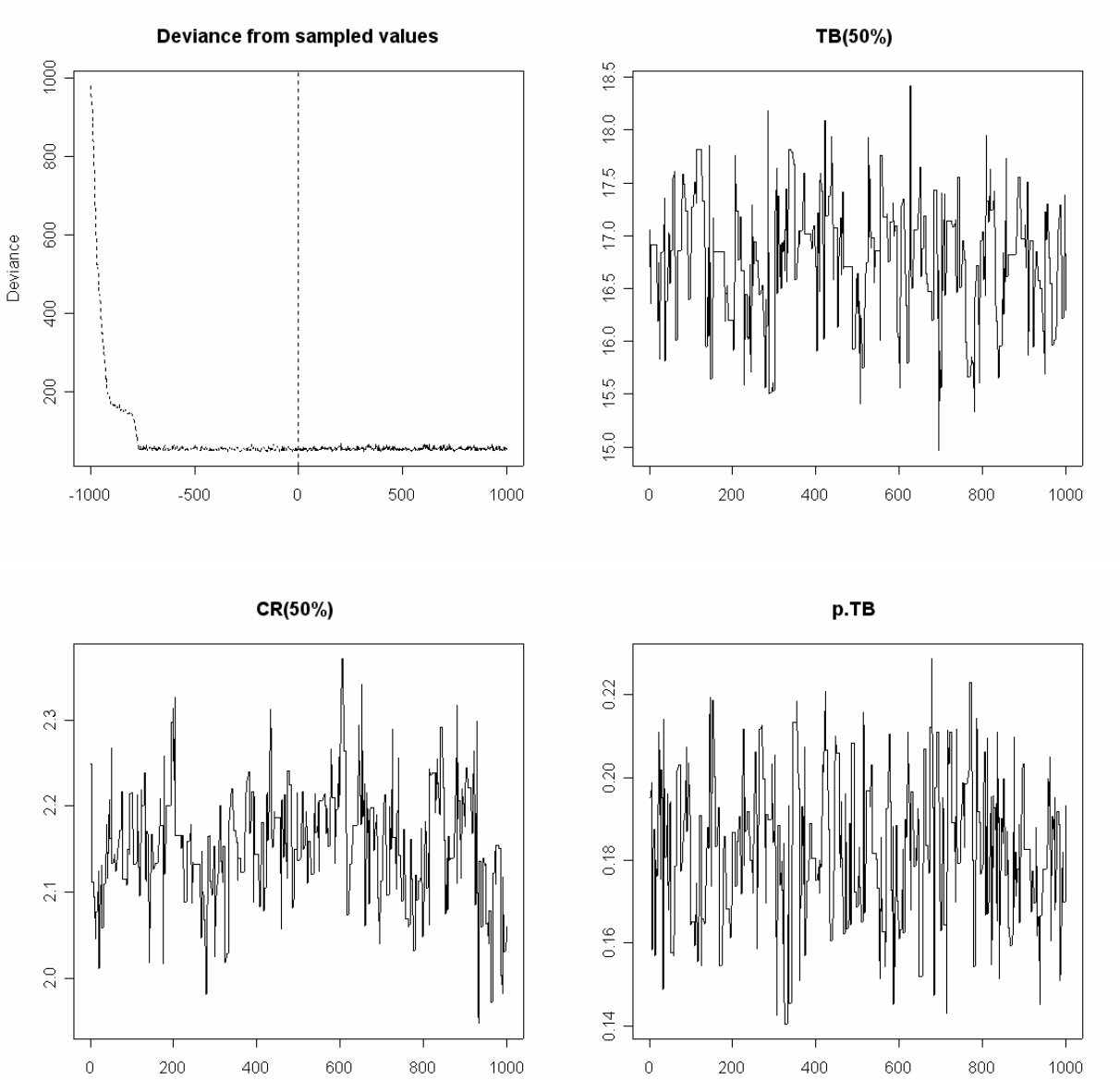
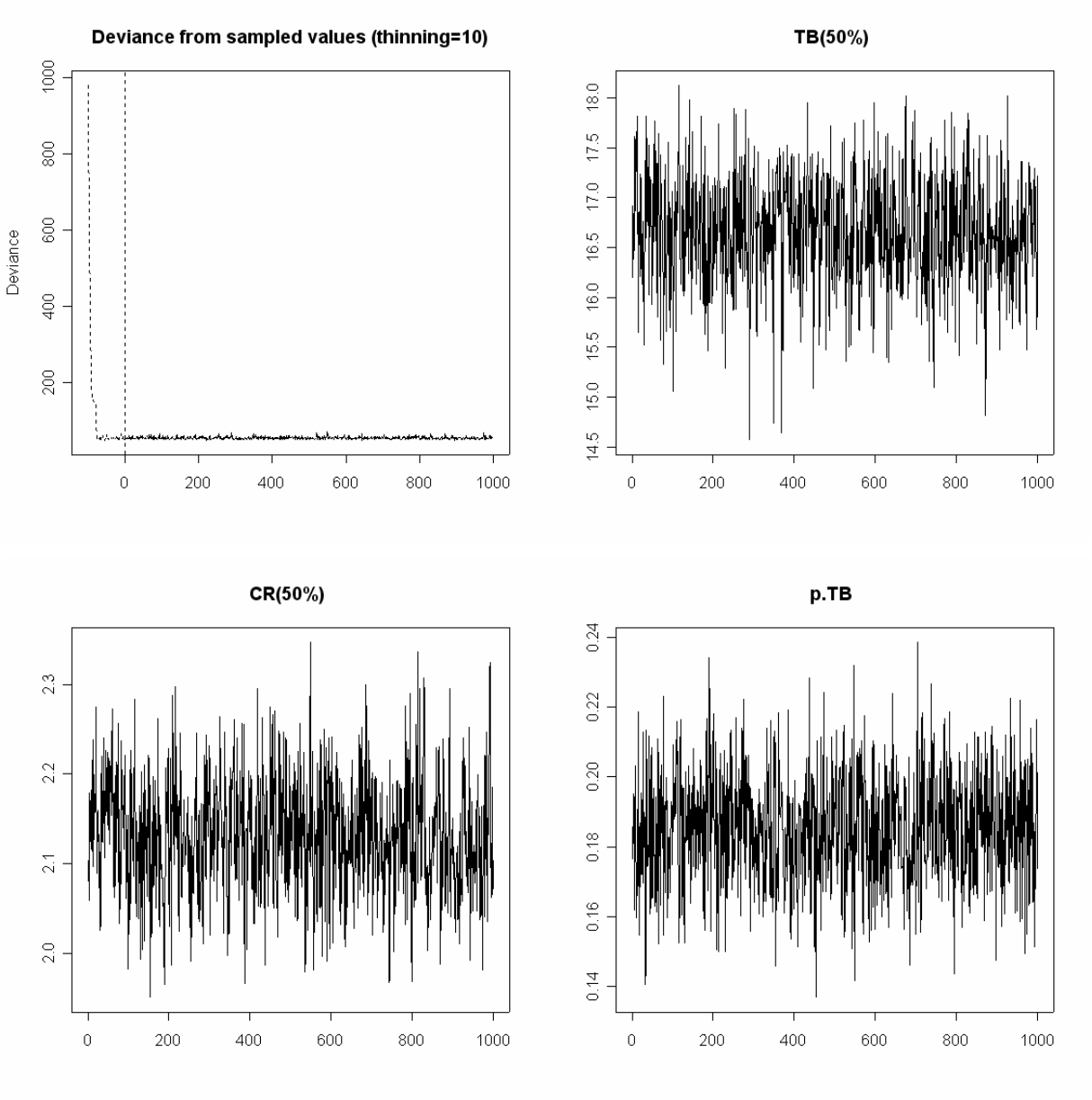


Figure 9-4 Assessing convergence (Example 4): MS.size=c(1000,1000,10)



10 Extending the Basic Model*

The program **NonBCG.r** can be used to do mixture analysis of induration data for more than one frequency distribution (e.g. for different age groups), even if the basic assumption of common mixture component distributions among groups is relaxed. Note that extending the model is only necessary if the basic model performs poorly.

The differences are the new parameters

```
TB.qnt1.diff, TB.qnt2.diff, CR.qnt1.diff, CR.qnt2.diff
```

that specify the maximum differences over all groups for the parameters specifying the mixture component distributions. For example, the specifications

```
TB.qnt1.diff <- 2
TB.qnt2.diff <- 2
CR.qnt1.diff <- 1
CR.qnt2.diff <- 1
```

allow the quantiles of the infection distribution to differ by at most 1mm, and the quantiles of the cross-reaction distribution by at most 1mm.

Note: the starting values of the mixture component quantiles must be consistent with the above specifications!

Due to the fact that the number of parameters can be considerable (depending on the number of groups to be analysed), the number of iterations (given by `MS.size`) required to obtain convergence of the sampler can be large. Therefore, the analysis can be time-consuming. In any case, convergence of the sampler needs to be inspected. Moreover, the length of the burn-in period depends on the number of groups to be analysed.

10.1 Application 3: Korea, males, all age groups

Application 3 provides the analysis of the six age groups (males) from the Korean data set based on the extended model. Thus, instead of assuming identical mixture component distributions (for TB infection as well as cross-reactions) for the six age groups, the model allows for more flexibility. As an example, it will be assumed that the quantiles of the TB distributions differ by at most 2mm over the six age groups. For the quantiles of the CR distributions a difference of 1mm will be assumed.

10.1.1 Program input

```
infile <- "korea75m.asc"
freq.column <- seq(2,7)

outfile <- "App3Out"

MS.run <- T
```

```
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247
MS.size <- c(5000,2000,10)

distTB <- "Wb"
distCR <- "LN"

group.names <- c("0-4","5-9","10-14","15-19","20-24","25-29")

TB.qnt1.diff <- 2
TB.qnt2.diff <- 2

CR.qnt1.diff <- 1
CR.qnt2.diff <- 1

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\NONBCG.r")
```

Note that the only change needed compared to the basic model of Section 7 are the specifications for `TB.qnt1.diff, TB.qnt2.diff, CR.qnt1.diff, CR.qnt2.diff.`

10.1.2 Results

Table 110-1 Parameter estimates for Application 3

```
Prevalence of TB infections (p.TB)
        mean  st.dev   2.5%    50%  97.5%
0-4   0.0411 0.00635 0.0295 0.0408 0.0545
5-9   0.1839 0.01633 0.1527 0.1842 0.2171
10-14 0.4967 0.02201 0.4545 0.4971 0.5409
15-19 0.7414 0.01713 0.7068 0.7419 0.7739
20-24 0.8599 0.01776 0.8239 0.8610 0.8917
25-29 0.9207 0.01093 0.8971 0.9217 0.9403

 TB distribution (1st quantile, TB.qnt1)
      mean st.dev 2.5%  50% 97.5%
0-4   16.3  0.455 15.6 16.3  17.5
5-9   16.8  0.441 16.0 16.8  17.6
10-14 17.7  0.277 17.1 17.7  18.2
15-19 17.6  0.196 17.2 17.6  18.0
20-24 17.1  0.207 16.7 17.1  17.5
25-29 17.1  0.161 16.8 17.1  17.4
```

Table 110-2 Model checks for Application 3

```
Summary of predictive checks (+ predictions too large, - predictions too small)
    10 % predictive failures

1: 0-4 2: 5-9 3: 10-14 4: 15-19 5: 20-24 6: 25-29

      1 2 3 4 5 6 total
1mm                   0
2mm     -             1
3mm   + + +           3
4mm                   0
5mm                   0
6mm           -       1
7mm       -           1
8mm                   0
9mm     -       +     2
10mm                  0
11mm                  0
12mm                  0
13mm        +         1
```

```
14mm       +            1
15mm              -     1
16mm                    0
17mm         -          1
18mm                    0
19mm                    0
20mm     -       -      2
21mm             + +    2
22mm                    0
23mm                    0
24mm             +      1
25mm                    0
26mm                    0
27mm                    0
28mm                    0
29mm                    0
30mm         -          1
total 1 4 3 3 4 3      18
```

Note that the analysis using the extended model results in 18 (out of 180) predictive failures, a 10% failure rate. This is an improvement over the 25 failures seen for the basic model (Section 7). It should be noted that some of the predictive failures are probably due to digit preference (at 20 and 21mm), i.e., the actual predictive failure rate for the extended model would be below 10%.

The comparison the prevalence estimates for TB infections for the basic and extended model (Table 10-1) are in good agreement. It can therefore be concluded that the findings are relatively robust with regard to model specifications.

Figure 110-1 Application 3: Estimates of prevalence of *Mycobacterium Tuberculosis*, and prevalence of zero reaction (males, 6 age groups)

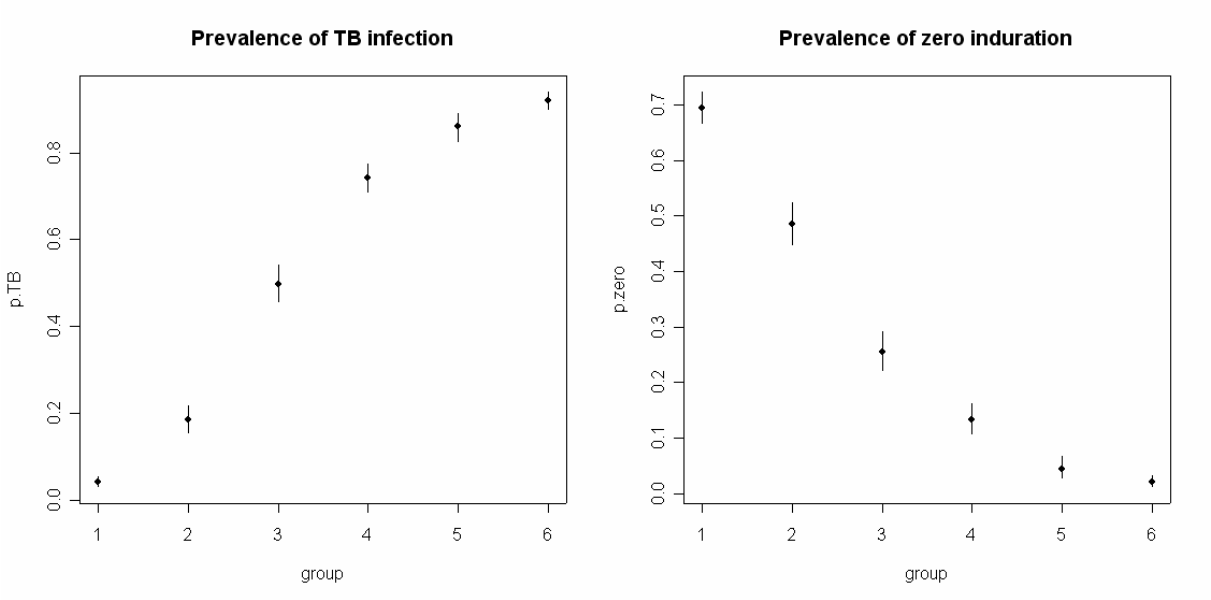
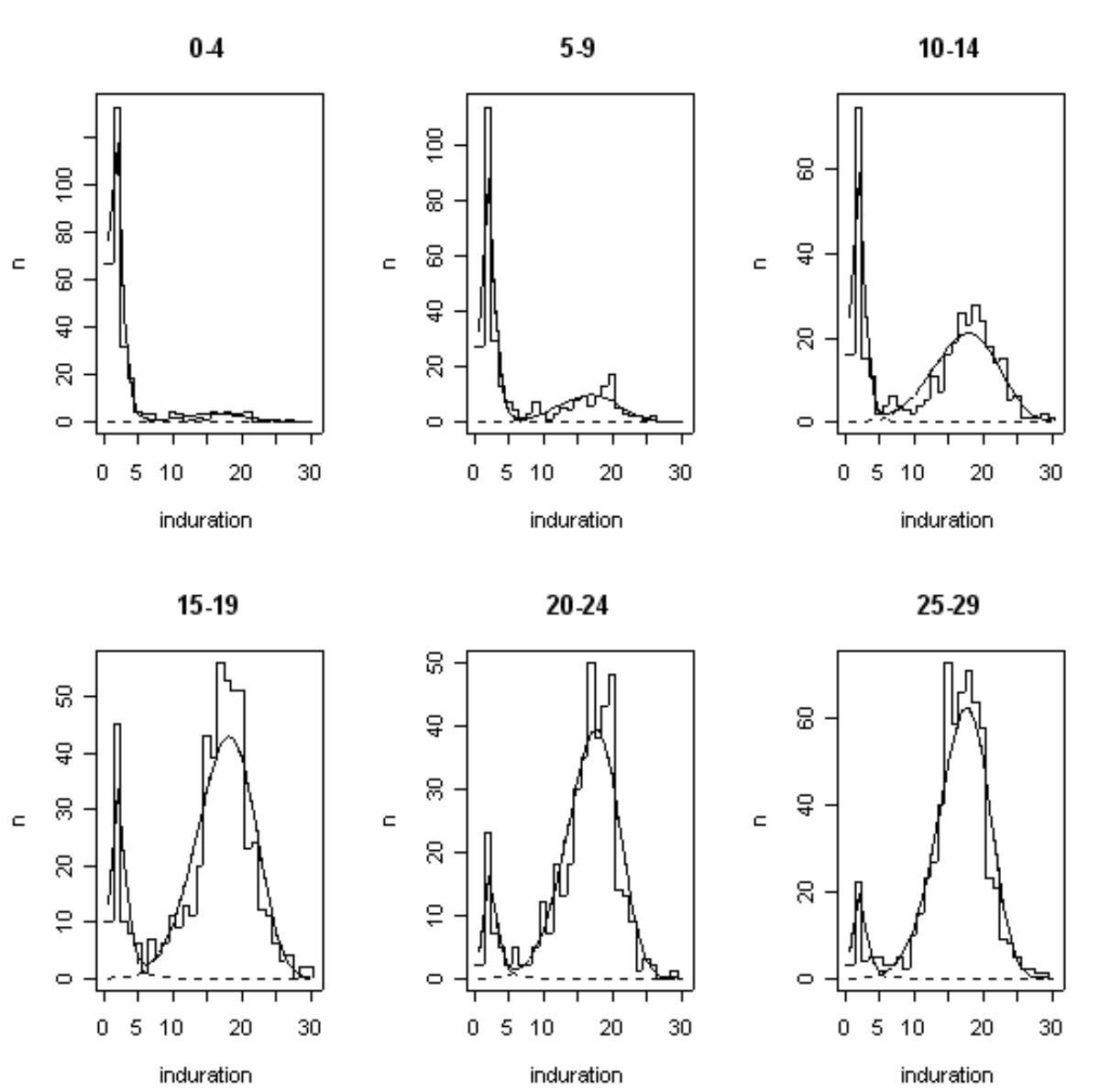


Figure 110-2 Application 3: Induration data and model fit



10.2 Application 4: Korea, males and females, all age groups

Application 4 provides an analysis of data from males and females from the Korean data set based on the extended model, i.e., data will be grouped by gender and age.

An alternative specification concerning the similarity of component distributions will be introduced here. Instead of specifying the maximum difference over all groups (this would include all age groups irrespective of gender), the following assumptions will be incorporated

1. males and females differ by at most 10% in their mixture component parameters, for both the TB and CR distributions.
2. For adjacent age groups, the ratio of TB parameters is between 0.9 and 1.1. For the ratio of CR parameters, the ratio corresponding ratio is between 0.85 and 1.15.

To allow for these specifications, the grouping structure needs to be set up as follows:

```
group.names <- list( c("SEX", "AGE"), c("M", "F"), c("0-4","5-9","10-14","15-
19","20-24","25-29") )
```

This means that the data are grouped by sex and age, i.e., the first six columns refer to males, the remaining columns to females.

Then the specifications for the ratios need to be specified as follows:

```
TB.qnt1.ratio <- list( 1.1, c(0.9,1.1))
TB.qnt2.ratio <- list( 1.1, c(0.9,1.1))
CR.qnt1.ratio <- list( 1.1, c(0.85,1.15))
CR.qnt2.ratio <- list( 1.1, c(0.85,1.15))
```

In each of specifications, the first element in the list refers to the gender constraints, the second one to the age constraints.

If the constraint is given by one number (e.g. 1.1), this means that over all groups (here for gender), the maximum ratio is 1.1.

If the constraint is 2-dimensional (e.g. c(0.9,1.1)), this refers to the constraints for adjacent age groups. For example, if the constraint for age groups would be c(1,1.2), this would mean that the parameters of mixture component distributions increase between 0 and 20% from one age group to the next.

10.2.1 Program input

```
infile <- "korea75mf.asc"
freq.column <- seq(1,12)

outfile <- "App4Out"

MS.run <- T
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247      # random seed (optional)
MS.size <- c(10000,2000,50)

distTB <- "Wb"
distCR <- "LN"
```

```
group.names  <-  list(  c("SEX",  "AGE"),  c("M",  "F"),  c("0-4","5-9","10-14","15-
19","20-24","25-29") )

TB.qnt1.ratio <- list( 1.1, c(0.9,1.1))
TB.qnt2.ratio <- list( 1.1, c(0.9,1.1))

CR.qnt1.ratio <- list( 1.1, c(0.85,1.15))
CR.qnt2.ratio <- list( 1.1, c(0.85,1.15))

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\NONBCG.r")
```

10.2.2 Results

Table 110-3 Parameter estimates for Application 4

```
Prevalence of TB infections (p.TB)
          mean  st.dev   2.5%    50%  97.5%
M 0-4   0.0424 0.00679 0.0304 0.0420 0.0562
M 5-9   0.1829 0.01684 0.1517 0.1825 0.2167
M 10-14 0.4951 0.02241 0.4539 0.4942 0.5410
M 15-19 0.7395 0.01734 0.7047 0.7400 0.7729
M 20-24 0.8559 0.01733 0.8209 0.8563 0.8881
M 25-29 0.9179 0.01113 0.8950 0.9185 0.9387
F 0-4   0.0620 0.00845 0.0465 0.0616 0.0793
F 5-9   0.1813 0.01614 0.1519 0.1805 0.2139
F 10-14 0.5454 0.02292 0.5007 0.5452 0.5889
F 15-19 0.7128 0.01808 0.6756 0.7129 0.7480
F 20-24 0.7738 0.01656 0.7397 0.7745 0.8061
F 25-29 0.7692 0.01558 0.7383 0.7698 0.7999
TB distribution (1st quantile, TB.qnt1)
        mean st.dev 2.5%  50% 97.5%
M 0-4   16.3  0.612 15.2 16.3  17.6
M 5-9   16.9  0.367 16.3 16.9  17.7
M 10-14 18.0  0.280 17.5 18.0  18.6
M 15-19 17.7  0.181 17.4 17.7  18.1
M 20-24 17.3  0.163 17.0 17.3  17.6
M 25-29 17.1  0.162 16.7 17.1  17.4
F 0-4   17.0  0.554 16.0 16.9  18.1
F 5-9   17.6  0.301 17.0 17.6  18.2
F 10-14 19.0  0.270 18.5 19.0  19.6
F 15-19 19.1  0.216 18.7 19.1  19.5
F 20-24 18.8  0.162 18.5 18.8  19.1
F 25-29 17.7  0.183 17.4 17.7  18.1
```

Table 110-4 Model checks for Application 4

```
Summary of predictive checks (+ predictions too large, - predictions too small)
    12.2 % predictive failures

1: M 0-4 2: M 5-9 3: M 10-14 4: M 15-19 5: M 20-24 6: M 25-29 7: F 0-4 8: F 5-9 9: F
10-14 10: F 15-19 11: F 20-24 12: F 25-29

      1  2  3  4  5  6  7  8  9 10 11 12 total
1mm   0  0  0  0  0  0  0  0  0  0  0  0     0
2mm   0 -1 -1  0  0  0 -1 -1  0  0  0  0     4
3mm   1  1  1  1  0  0  1  1  1  0  0  1     8
4mm   0  0  0  0  0  0  0  0  0  0  0  0     0
5mm   0  0  0  0  0  0  0  0  0  0  0  0     0
6mm   0  0  0  0  0  0  0  0 -1  0 -1  0     2
7mm   0  0 -1  0  0  0  0  0  0 -1  0 -1     3
8mm   0  0  0  0  0  0  0 -1  0  0  0  0     1
9mm   0 -1  0  0  0  1  0  0  0  0  0  0     2
10mm  0  0  0  0  0  0  0  0  0  0 -1  0     1
11mm  0  0  0  0  0  0  0  0  0  0  0  0     0
12mm  0  0  0  0  0  0  0  0  0  0  0  0     0
13mm  0  0  0  1  0  0  0  0  1  0  1  0     3
14mm  0  0  0  0  0  0  0  0  1  0  0  0     1
```

15mm	0	0	0	0	0	-1	0	0	1	0	0	0	2
16mm	0	0	0	0	0	0	0	0	0	0	0	1	1
17mm	0	0	0	0	0	0	0	0	0	0	0	-1	1
18mm	0	0	0	0	0	0	0	0	0	0	0	0	0
19mm	0	0	0	0	0	0	0	0	0	0	0	0	0
20mm	0	-1	0	0	-1	0	0	0	-1	-1	-1	-1	6
21mm	0	0	0	1	1	1	0	0	0	0	0	0	3
22mm	0	0	0	0	0	0	0	0	0	0	0	0	0
23mm	0	0	0	0	0	0	0	0	0	0	0	0	0
24mm	0	0	0	0	1	0	0	0	0	0	0	1	2
25mm	0	0	0	0	0	0	0	0	0	0	0	0	0
26mm	0	0	0	0	0	0	0	0	0	0	0	0	0
27mm	0	0	0	0	0	0	0	0	0	0	0	0	0
28mm	0	0	0	0	0	0	0	0	0	0	0	0	0
29mm	0	0	0	0	-1	-1	0	0	0	0	0	0	2
30mm	0	0	0	-1	0	0	0	0	0	-1	0	0	2
total	1	4	3	4	4	4	2	3	6	3	4	6	44

Figure 110-3 Application 4: Estimates of prevalence of *Mycobacterium Tuberculosis*, and prevalence of zero reaction (males and females, 6 age groups)

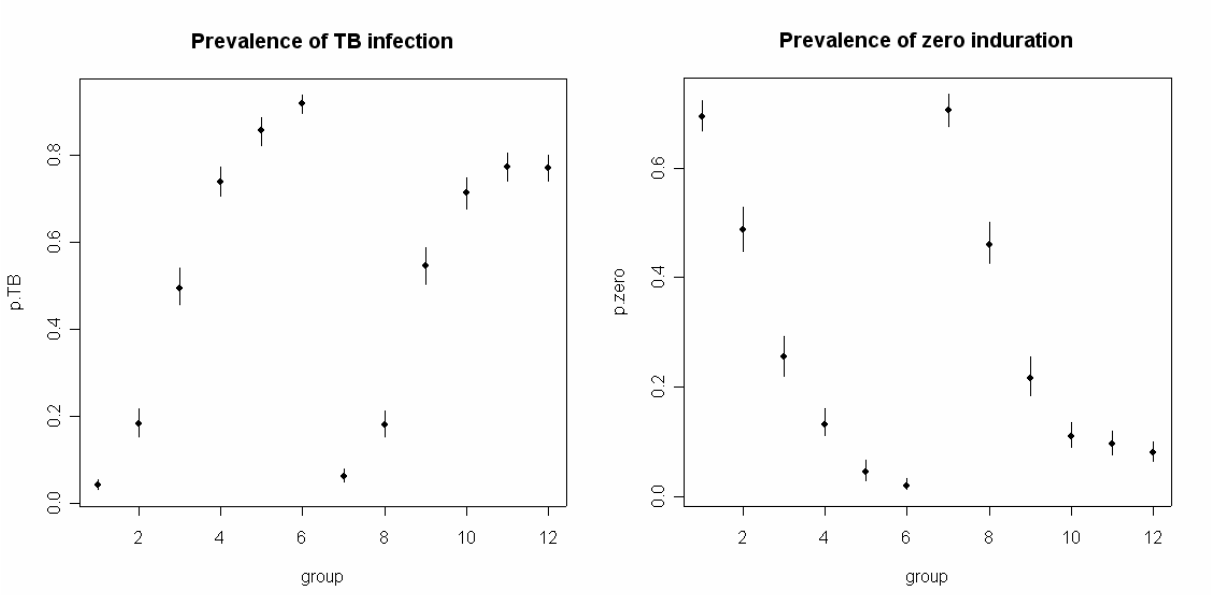
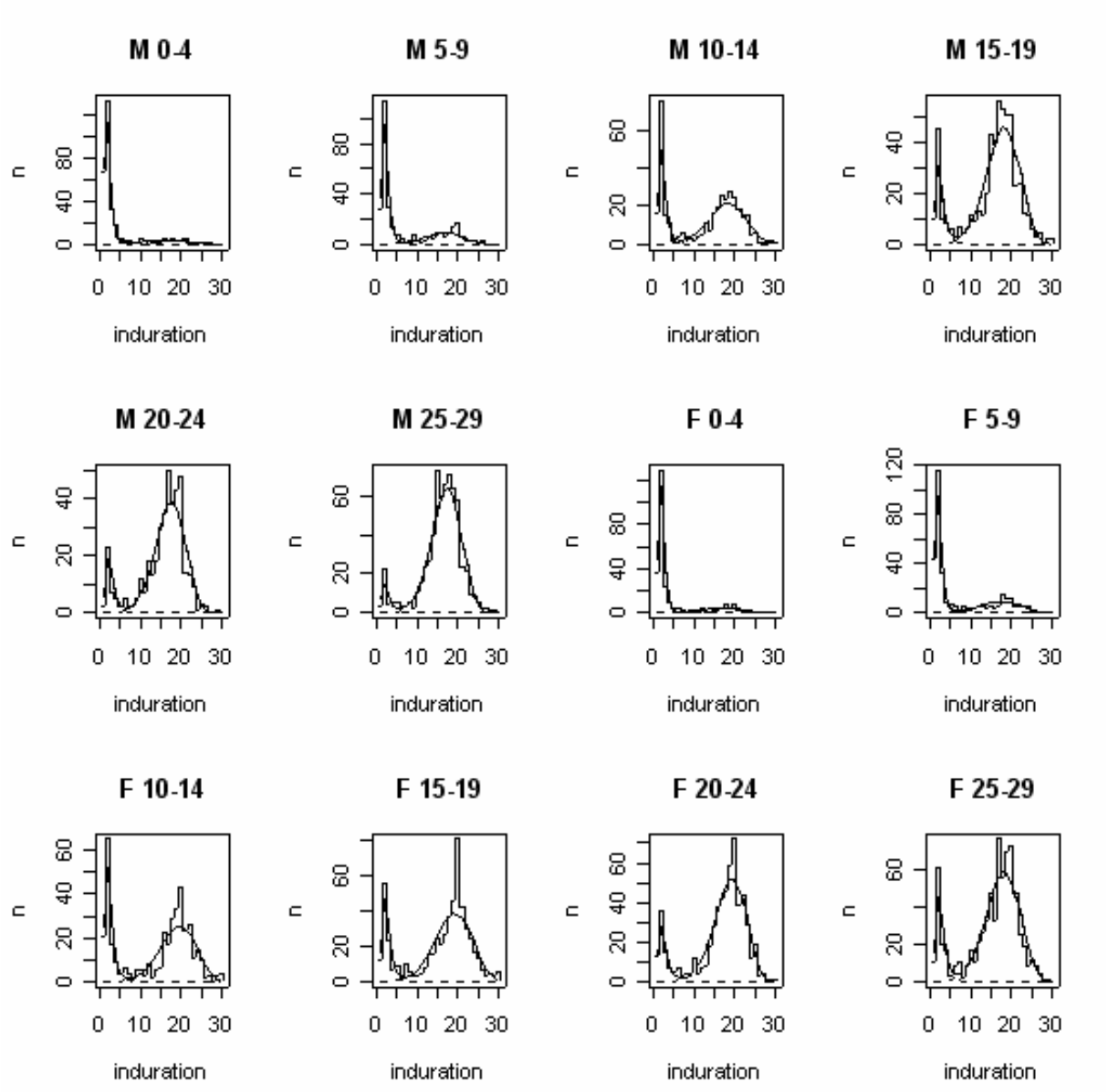


Figure 110-4 Application 4: Induration data and model fit



11 Grouped Induration Data: Application 5 (Navy Data)

So far it was assumed that induration data are available as frequencies for 0,1,…,30mm. This Section presents the analysis for grouped data. It should be noted that grouped induration data are more difficult to analyze, since information about the mixture component distributions is lost.

Application 5 is an extract from an old data set from the US Navy.

11.1 Data

```
alabama california florida georgia
7183      53755   12689   9823
120  696        262     197
144  754        375     253
125  479        284     198
127  380        261     137
92   414        196     118
96   419        129     108
70   427        101      74
63   385         75      57
68   385         63      55
27   309         27      31
11   220         17      26
6    133          6       7
2    57           4       2
2    29           1       2
3    15           0       1
```

The data set contains grouped induration data (frequencies) for 4 US states. The 1st row for zero-induration, the remaining 15 rows for the groups 0-2.5, 2.5-4.5,…, 26.5-28.5,>28.5. The 14 break-points defining the groups are defined in the variable `ind.brk` (see below).

11.2 Program input

```
infile <- "navy.asc"
header <- T

freq.column <- c(1,2,3,4)

group.names <- c( "Alabama", "California", "Florida", "Georgia" )

outfile <- "App5Out"

ind.brk <- c( 2.5, 4.5, 6.5, 8.5, 10.5, 12.5, 14.5, 16.5,
              18.5, 20.5, 22.5, 24.5, 26.5, 28.5 )

MS.size <- c(2000,5000,1)

RndSeed <- 7247

MS.run          <- T
MS.results      <- T
MS.check        <- T
```

```
MS.graph        <- T

distTB <- "N"
distCR <- "Wb"

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\NONBCG.r")
```

11.3 Results

Table 11-1 Parameter Estimates for Application 5

```
Prevalence of TB infections (p.TB)
             mean  st.dev    2.5%     50%  97.5%
Alabama    0.02470 0.00339 0.01817 0.02472 0.0312
California 0.03279 0.00177 0.02937 0.03288 0.0360
Florida    0.00746 0.00239 0.00295 0.00742 0.0123
Georgia    0.01404 0.00280 0.00882 0.01407 0.0199

 TB distribution (1st quantile, TB.qnt1)
           mean st.dev 2.5%  50% 97.5%
Alabama    16.6  0.233 16.2 16.6  17.1
California 16.6  0.233 16.2 16.6  17.1
Florida    16.6  0.233 16.2 16.6  17.1
Georgia    16.6  0.233 16.2 16.6  17.1
```

Table 11-2 Model checks for Application 5

```
Summary of predictive checks (+ predictions too large, - predictions too small)
    20 % predictive failures

1: Alabama 2: California 3: Florida 4: Georgia

        1  2  3  4 total
ind-1   1 -1  1  0     3
ind-2   0 -1 -1 -1     3
ind-3   0  1  0  0     1
ind-4   0  1  0  0     1
ind-5   0  0  0  0     0
ind-6  -1  0  0  0     1
ind-7   0  0  0  0     0
ind-8   0  0  0  0     0
ind-9  -1  0  0  0     1
ind-10  0  0  0  0     0
ind-11  1  0  0  0     1
ind-12  1  0  0  0     1
ind-13  0  0  0  0     0
ind-14  0  0  0  0     0
ind-15  0  0  0  0     0
total   5  4  2  1    12
```

Figure 11-1 Application 5: Estimates of prevalence of *Mycobacterium Tuberculosis*, and prevalence of zero reaction

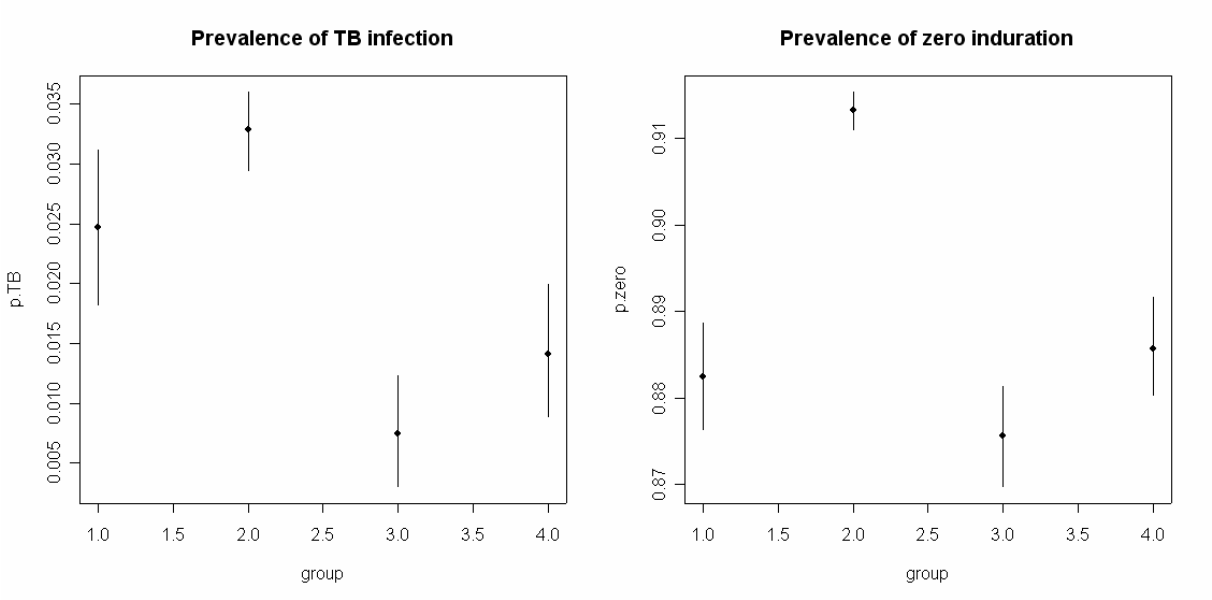
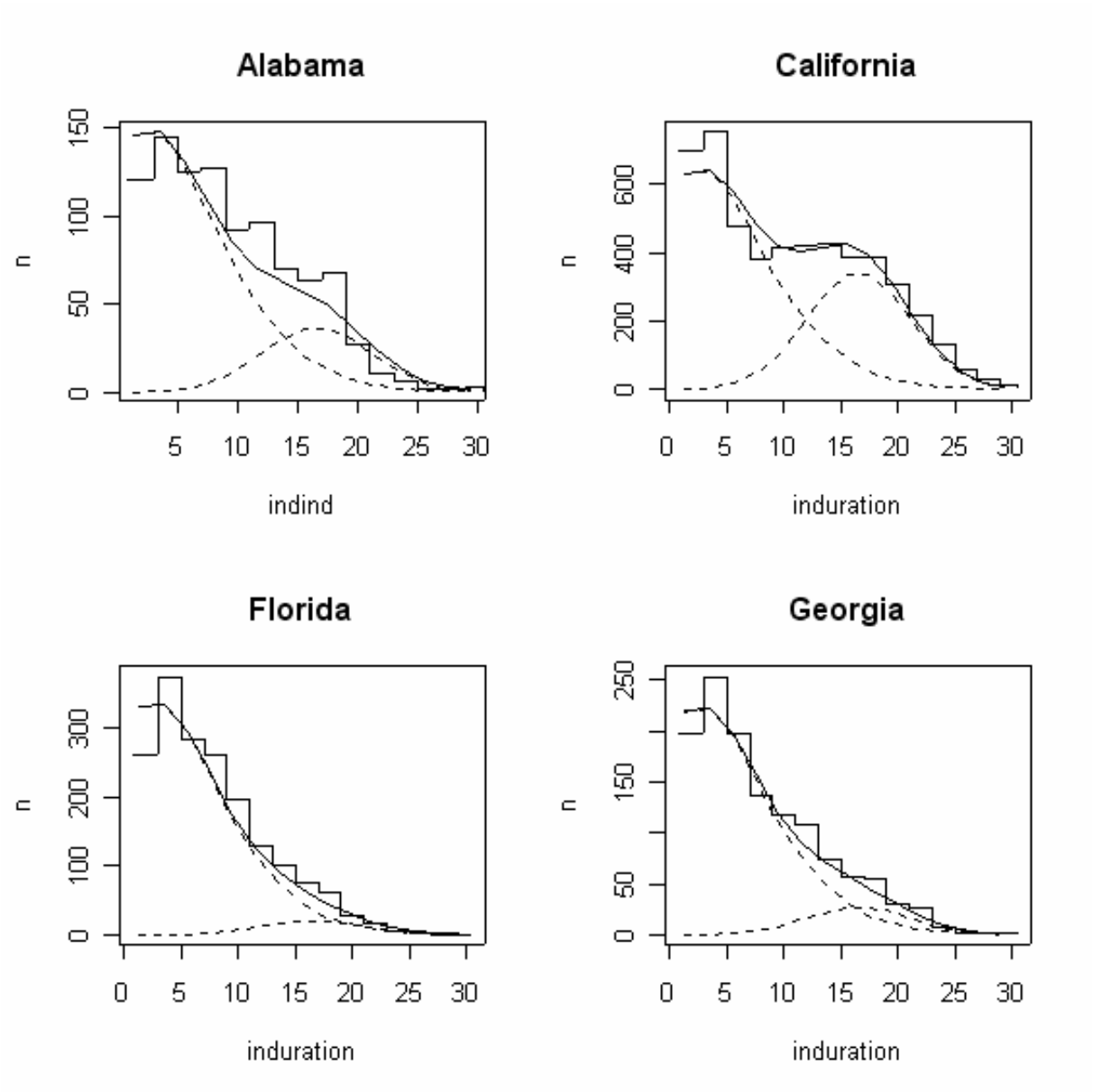


Figure 11-2 Application 5: Induration data and model fit



12 Analysis of Unvaccinated and Vaccinated Subjects: Application 6

12.1 Introduction

In the previous Sections we considered the analysis for data from unvaccinated subjects by mixture analysis involving two component distributions, one for infections with *Mycobaterium tuberculosis* and one for cross-reactions. Applications were shown for one group (Section 4), and for several groups using the basic (Section 7) and extended (Section 10) model.

In the remaining Sections we will consider both unvaccinated and vaccinated subjects, and extend the mixture framework in so far as we are going to allow for a third mixture component (BCG) in the group of vaccinated subjects. The basic model will assume that the TB and CR mixture components are the same in unvaccinated and vaccinated subjects.

The combined mixture analysis of unvaccinated and vaccinated subjects is a more complicated and ambitious endevour. In the sequel we will illustrate the analysis of unvaccinated and vaccinated subjects with examples of increasing complexity.

The following general remarks will apply:

1. Regarding notation, a **"0"** and **"1"** will be used to refer to the unvaccinated and vaccinated, respectively. For example

   ```
   infile0, freq0.column
   ```

   will be used to denote the input file and column specifications for unvaccinated subjects. In analogy,

   ```
   infile1, freq1.column
   ```

   are the corresponding specifications for the data from vaccinated subjects. The interpretation of all quantities involved remains the same as for the unvaccinated case.

2. *BCG* will be used for the third mixture component in the vaccinated group. Examples referring to this component are

   ```
   distBCG <- WB
   BCG.qnt1.init <- 10
   BCG.qnt1.ratio <- 1.2
   ```

   for specifying the type of distribution, the initial value for the first quantile, and constraints for the first quantile, respectively.

12.2 Data

Application 6 presents the basic analysis of the data for males of age 5-9 years, for both unvaccinated and vaccinated subjects (Appendix 16.1.3). Note that the basic model assumes that the TB and CR distribution are the same in unvaccinated and vaccinated subjects.

12.3 Program input

```
infile0 <- "korea75mf.asc"
freq0.column <- 2

infile1 <- "korea75mfBCG.asc"
freq1.column <- 2

outfile <- "App6Out"

MS.run <- T
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247      # random seed (optional)
MS.size <- c(2000,2000,10)

distTB <- "Wb"
distCR <- "LN"
distBCG <-"Wb"

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\BCG.r")
```

12.4 Results

```
Parameter estimates: non-BCG

 Prevalence of TB infections (p0.TB)
  mean st.dev  2.5%   50% 97.5%
 0.191 0.0171 0.158 0.191 0.226

 TB distribution (1st quantile, TB0.qnt1)
 mean st.dev 2.5%  50% 97.5%
 15.7  0.542 14.6 15.7  16.7

Parameter estimates: BCG

 Prevalence of TB infections (p1.TB)
  mean st.dev  2.5%   50% 97.5%
 0.249 0.0487 0.152 0.250 0.339

 TB distribution (1st quantile, TB1.qnt1)
 mean st.dev 2.5%  50% 97.5%
 15.7  0.542 14.6 15.7  16.7
```

Table 12-1 Model checks for Application 6

```
BCG: summary of predictive checks (+ predictions too large, - predictions too small)
    16.7 % predictive failures

1:

       1 total
1mm    0     0
2mm   -1     1
3mm    1     1
4mm    0     0
5mm    0     0
6mm    0     0
7mm    0     0
8mm    0     0
9mm    0     0
10mm   0     0
11mm   0     0
12mm  -1     1
13mm   0     0
14mm   0     0
15mm   0     0
16mm   1     1
```

```
17mm    0      0
18mm    0      0
19mm    0      0
20mm    0      0
21mm    0      0
22mm    0      0
23mm    1      1
24mm    0      0
25mm    0      0
26mm    0      0
27mm    0      0
28mm    0      0
29mm    0      0
30mm    0      0
total   5      5
```

13 The Basic Model for Several Groups: Application 7

The basic model assumes that all mixture component distributions are common across groups. In application 7 the basic model is used for data from males for all age groups, both unvaccinated and vaccinated.

13.1 Program input

```
infile0 <- "korea75mf.asc"
freq0.column <- seq(1,6)

infile1 <- "korea75mfBCG.asc"
freq1.column <- seq(1,6)

outfile <- "App7Out"

MS.run <- T
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247
MS.size <- c(5000,2000,10)

distTB <- "Wb"
distCR <- "LN"
distBCG <-"Wb"

group.names <- c("0-4","5-9","10-14","15-19","20-24","25-29")

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\BCG.r")
```

13.2 Results

```
Parameter estimates: non-BCG

 Prevalence of TB infections (p0.TB)
        mean  st.dev   2.5%    50% 97.5%
0-4   0.0405 0.00607 0.0295 0.0401 0.053
5-9   0.1845 0.01608 0.1540 0.1841 0.216
10-14 0.5002 0.02364 0.4554 0.4999 0.547
15-19 0.7489 0.01650 0.7164 0.7489 0.781
20-24 0.8656 0.01817 0.8282 0.8659 0.899
25-29 0.9247 0.01113 0.9011 0.9252 0.945

 TB distribution (1st quantile, TB0.qnt1)
      mean st.dev 2.5%  50% 97.5%
0-4   16.9  0.103 16.7 16.9  17.1
5-9   16.9  0.103 16.7 16.9  17.1
10-14 16.9  0.103 16.7 16.9  17.1
15-19 16.9  0.103 16.7 16.9  17.1
20-24 16.9  0.103 16.7 16.9  17.1
```

```
25-29 16.9  0.103 16.7 16.9  17.1


Parameter estimates: BCG

 Prevalence of TB infections (p1.TB)
        mean st.dev    2.5%    50%  97.5%
0-4   0.0466 0.0228 0.00565 0.0466 0.0924
5-9   0.1314 0.0237 0.08362 0.1310 0.1770
10-14 0.3732 0.0326 0.30594 0.3753 0.4331
15-19 0.4214 0.0503 0.32018 0.4279 0.5086
20-24 0.6669 0.0497 0.55927 0.6676 0.7614
25-29 0.8915 0.0401 0.81726 0.8922 0.9616

 TB distribution (1st quantile, TB1.qnt1)
      mean st.dev 2.5%  50% 97.5%
0-4   16.9  0.103 16.7 16.9  17.1
5-9   16.9  0.103 16.7 16.9  17.1
10-14 16.9  0.103 16.7 16.9  17.1
15-19 16.9  0.103 16.7 16.9  17.1
20-24 16.9  0.103 16.7 16.9  17.1
25-29 16.9  0.103 16.7 16.9  17.1
```

Table 13-1 Model checks for Application 7 (non-BCG group)

```
Non-BCG: summary of predictive checks (+ predictions too large, - predictions too small)
    15.6 % predictive failures

1: 0-4 2: 5-9 3: 10-14 4: 15-19 5: 20-24 6: 25-29

       1  2  3  4  5  6 total
1mm   -1  0  0  0  0  0     1
2mm    0  0  0  0  0  0     0
3mm    1  1  1  0  0  0     3
4mm    0  0  0  0  0  0     0
5mm    0  0  0 -1  0  0     1
6mm    0  0  0  0 -1  0     1
7mm   -1  0 -1  0  0  0     2
8mm    0  0  0  0  0  0     0
9mm    0 -1  0  0  0  1     2
10mm  -1  0  0  0  0  0     1
11mm   0  0  0  0  0  0     0
12mm   0  0  0  0  0  0     0
13mm   0  0  0  1  0  0     1
14mm   0  0  1  1  0  0     2
15mm   0  0  0  0  0 -1     1
16mm   0  0  0  0  0  0     0
17mm   0  0  0  0 -1  0     1
18mm   0  0  0  0  0  0     0
19mm   0  0  0  0  0  0     0
20mm   0 -1  0 -1 -1  0     3
21mm   0  0  0  0  0  1     1
22mm   0  0  0  0  0  0     0
23mm   0  0 -1  0  0  1     2
24mm   0  0  0  0  1  0     1
25mm   0  0  0  0  0  0     0
26mm   0  0  0  0  0  0     0
27mm   0  0  0 -1  0  0     1
28mm   0  0  0  0  0  0     0
29mm   0  0 -1 -1  0  0     2
30mm   0  0 -1 -1  0  0     2
total  4  3  6  7  4  4    28
```

Table 13-2 Model checks for Application 7 (BCG group)

```
BCG: summary of predictive checks (+ predictions too large, - predictions too small)
    13.3 % predictive failures

1: 0-4 2: 5-9 3: 10-14 4: 15-19 5: 20-24 6: 25-29

       1  2  3  4  5 6 total
1mm    0  1  0  1  0 0     2
2mm    0 -1 -1  0  0 0     2
3mm    0  1  1  0  0 0     2
4mm    0  0  0  0  0 0     0
5mm   -1 -1 -1  0  0 0     3
```

6mm	0	0	1	0	0	0	1
7mm	0	0	0	0	0	0	0
8mm	0	0	0	0	0	0	0
9mm	1	0	0	1	0	0	2
10mm	0	0	0	0	0	0	0
11mm	0	0	0	0	0	0	0
12mm	0	0	0	0	0	0	0
13mm	0	0	0	0	0	0	0
14mm	1	0	0	0	0	0	1
15mm	0	0	-1	-1	-1	0	3
16mm	0	0	0	0	0	0	0
17mm	0	0	0	0	0	0	0
18mm	0	0	0	0	0	0	0
19mm	0	0	0	0	0	0	0
20mm	0	0	0	0	0	0	0
21mm	0	0	0	1	0	0	1
22mm	0	0	1	1	1	0	3
23mm	0	1	0	0	1	0	2
24mm	0	0	0	1	0	0	1
25mm	0	0	-1	0	0	0	1
26mm	0	0	0	0	0	0	0
27mm	0	0	0	0	0	0	0
28mm	0	0	0	0	0	0	0
29mm	0	0	0	0	0	0	0
30mm	0	0	0	0	0	0	0
total	3	5	7	6	3	0	24

14 Extending the Basic Model for Several Groups*

The basic model of Section 13 (assuming that all mixture component distributions are common across groups) will now be extended. Note that this extension is only required if the results from the basic model show unsatisfactory results.

14.1 Application 8: Korea, males, all age groups

In application 8 the basic model will be extended to allow for more flexibility regarding the mixture component distributions.

- different mixture component distributions for TB reactions in the group of unvaccinated subjects:

  ```
  TB0.qnt1.ratio, TB0.qnt2.ratio
  ```

  In the example below, a ratio between 0.9 and 1.1 between adjacent age groups will be assumed.

- different mixture component distributions for CR reactions in the group of unvaccinated subjects

  ```
  CR0.qnt1.ratio, CR0.qnt2.ratio
  ```

  In the example below, a ratio between 0.85 and 1.15 between adjacent age groups will be assumed.

- different mixture component distributions for TB reactions in the group of vaccinated subjects

  ```
  TB1.qnt1.ratio, TB1.qnt2.ratio
  ```

  In the example below, a ratio between 0.9 and 1.1 between adjacent age groups will be assumed.

- different mixture component distributions for CR reactions in the group of vaccinated subjects

  `CR1.qnt1.ratio, CR1.qnt2.ratio`

  In the example below, a ratio between 0.85 and 1.15 between adjacent age groups will be assumed.

- different mixture component distributions for BCG reactions

  `BCG.qnt1.ratio, BCG.qnt2.ratio`

  In the example below, a ratio between 0.75 and 0.90 between adjacent age groups will be assumed. Note that this corresponds to a minimum (maximum) decrease of 10% (25%) for the parameters of the BCG distribution from one age group to the next.

  Note that the initial values for the parameters of the BCG distribution, i.e.,

  `BCG.qnt1.init, BCG.qnt2.init`

  must be chosen in such a way that they are consistent with the constraints.

- different mixture component distributions between the unvaccinated and vaccinated for TB infections

  `TB01.qnt1.ratio, TB01.qnt2.ratio`

  Note that here “01” refers to the specification between unvaccined and vaccinated.

  In the example below, a ratio of 1.1 was assumed, i.e., the parameters of the TB distribtution were allowed to be different by at most 10% between the unvaccinated and vaccinated.

- different mixture component distributions between the unvaccinated and vaccinated for CR infections

  `CR01.qnt1.ratio, CR01.qnt2.ratio`

  In the example below, a ratio of 1.1 was assumed, i.e., the parameters of the CR distribution were allowed to be different by at most 10% between the unvaccinated and vaccinated.

14.1.1 Program input

```
infile0 <- "korea75mf.asc"
freq0.column <- seq(1,6)

infile1 <- "korea75mfBCG.asc"
freq1.column <- seq(1,6)

outfile <- "App8Out"

MS.run <- T
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247      # random seed (optional)
MS.size <- c(10000,2000,50)

distTB <- "Wb"
distCR <- "LN"
distBCG <-"Wb"
```

```
group.names <- c("0-4","5-9","10-14","15-19","20-24","25-29")

TB0.qnt1.ratio <- c(0.9,1.1)
TB0.qnt2.ratio <- c(0.9,1.1)

CR0.qnt1.ratio <- c(0.85,1.15)
CR0.qnt2.ratio <- c(0.85,1.15)

TB1.qnt1.ratio <- c(0.90,1.1)
TB1.qnt2.ratio <- c(0.90,1.1)

CR1.qnt1.ratio <- c(0.85,1.15)
CR1.qnt2.ratio <- c(0.85,1.15)

BCG.qnt1.ratio <- c(0.75,0.90)
BCG.qnt2.ratio <- c(0.75,0.90)

TB01.qnt1.ratio <- 1.1
TB01.qnt2.ratio <- 1.1

CR01.qnt1.ratio <- 1.1
CR01.qnt2.ratio <- 1.1

BCG.qnt1.init <- seq(8,3.1,length=6)
BCG.qnt2.init <- seq(15,6,length=6)

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\BCG.r")
```

14.1.2 Results

```
Parameter estimates: non-BCG

 Prevalence of TB infections (p0.TB)
        mean  st.dev   2.5%    50%  97.5%
0-4   0.0431 0.00681 0.0310 0.0429 0.0582
5-9   0.1887 0.01652 0.1576 0.1882 0.2215
10-14 0.5040 0.02265 0.4598 0.5034 0.5473
15-19 0.7464 0.01795 0.7100 0.7466 0.7812
20-24 0.8642 0.01697 0.8299 0.8647 0.8961
25-29 0.9216 0.01060 0.8993 0.9219 0.9412

 TB distribution (1st quantile, TB0.qnt1)
      mean st.dev 2.5%  50% 97.5%
0-4   15.6  0.611 14.5 15.6  16.9
5-9   16.2  0.378 15.5 16.2  17.0
10-14 17.2  0.230 16.7 17.2  17.6
15-19 17.1  0.170 16.8 17.1  17.5
20-24 17.0  0.184 16.6 17.0  17.3
25-29 17.0  0.167 16.7 17.0  17.4

Parameter estimates: BCG

 Prevalence of TB infections (p1.TB)
        mean st.dev   2.5%    50% 97.5%
0-4   0.0712 0.0310 0.0109 0.0715 0.128
5-9   0.2049 0.0280 0.1496 0.2048 0.258
10-14 0.5435 0.0233 0.4988 0.5434 0.590
15-19 0.6644 0.0221 0.6217 0.6647 0.707
20-24 0.8352 0.0227 0.7865 0.8365 0.875
25-29 0.9423 0.0158 0.9070 0.9434 0.968

 TB distribution (1st quantile, TB1.qnt1)
      mean st.dev 2.5%  50% 97.5%
0-4   16.0  0.731 14.7 16.0  17.5
5-9   16.1  0.473 15.3 16.1  17.0
10-14 15.8  0.197 15.4 15.8  16.2
15-19 15.7  0.159 15.4 15.7  16.0
20-24 15.7  0.187 15.3 15.7  16.1
25-29 16.6  0.238 16.1 16.6  17.0
```

Table 14-1 Model checks for Application 8 (non-BCG group)

```
Non-BCG: summary of predictive checks (+ predictions too large, - predictions too small)
    10 % predictive failures

1: 0-4 2: 5-9 3: 10-14 4: 15-19 5: 20-24 6: 25-29

      1  2 3  4  5  6 total
1mm   0  0 0  0  0  0     0
2mm   0  0 0  0  0  0     0
3mm   1  0 1  0  0  0     2
4mm   0  0 0  0  0  0     0
5mm   0 -1 0 -1  0  0     2
6mm   0  0 0  0 -1  0     1
7mm   0  0 0  0  0  0     0
8mm   0  0 0  0  0  0     0
9mm   0  0 0  0  0  1     1
10mm  0  0 0  0  0  0     0
11mm  0  0 0  0  0  0     0
12mm  0  0 0  0  0  0     0
13mm  0  0 0  1  0  0     1
14mm  0  0 1  1  0  0     2
15mm  0  0 0  0  0 -1     1
16mm  0  0 0  0  0  0     0
17mm  0  0 0 -1  0  0     1
18mm  0  0 0  0  0  0     0
19mm  0  0 0  0  0  0     0
20mm  0 -1 0 -1 -1  0     3
21mm  0  0 0  0  1  1     2
22mm  0  0 0  0  0  0     0
23mm  0  0 0  0  0  0     0
24mm  0  0 0  0  0  0     0
25mm  0  0 0  0  0  0     0
26mm  0  0 0  0  0  0     0
27mm  0  0 0  0  0  0     0
28mm  0  0 0  0  0  0     0
29mm  0  0 0  0 -1  0     1
30mm  0  0 0 -1  0  0     1
total 1  2 2  6  4  3    18
```

Table 14-2 Model checks for Application 8 (BCG group)

```
BCG: summary of predictive checks (+ predictions too large, - predictions too small)
    7.2 % predictive failures

1: 0-4 2: 5-9 3: 10-14 4: 15-19 5: 20-24 6: 25-29

       1  2  3  4 5 6 total
1mm    0  0  0  0 0 0     0
2mm    0  0  0  0 0 0     0
3mm    0  1  0  0 0 0     1
4mm    0 -1  0  0 0 0     1
5mm   -1  0  0  0 0 0     1
6mm    0  0  1  0 0 0     1
7mm    0  0  0  0 0 0     0
8mm    0  0  0  0 0 0     0
9mm    0  0  0  0 0 0     0
10mm   0  0 -1 -1 0 0     2
11mm   0  0  0  0 0 0     0
12mm   0 -1  0  0 0 0     1
13mm   0  0  0  0 0 0     0
14mm   0  0  0  0 0 0     0
15mm   0  0 -1 -1 0 0     2
16mm   0  1  0  0 0 0     1
17mm   0  0  0  0 0 0     0
18mm   0  0  0  0 0 0     0
19mm   0  0  0  0 0 0     0
20mm   0  0  0  0 0 0     0
21mm   0  0  0  0 0 0     0
22mm   0  0  1  0 0 0     1
23mm   0  1  0  0 0 0     1
24mm   0  0  0  0 0 0     0
25mm   0  0 -1  0 0 0     1
26mm   0  0  0  0 0 0     0
27mm   0  0  0  0 0 0     0
28mm   0  0  0  0 0 0     0
29mm   0  0  0  0 0 0     0
30mm   0  0  0  0 0 0     0
total  1  5  5  2 0 0    13
```

14.2 Application 9: Korea, males, females, all age groups

Application 9 presents the analysis for males and females (6 age groups) for unvaccinated and vaccinated subjects. Age constraints are the same as for the analysis of males only (Application 8), and for gender it is assumed that males and females differ by at most 10% with regard to their mixture component parameters.

14.2.1 Program input

```
infile0 <- "korea75mf.asc"
freq0.column <- seq(1,12)

infile1 <- "korea75mfBCG.asc"
freq1.column <- seq(1,12)

outfile <- "App9Out"

MS.run <- T
MS.results <- T
MS.check <- T
MS.graph <- T

RndSeed <- 7247
MS.size <- c(10000,2000,50)

distTB <- "Wb"
distCR <- "LN"
distBCG <-"Wb"

group.names <- list( c("SEX", "AGE"), c("M", "F"), c("0-4","5-9","10-14","15-19","20-24","25-
29") )

TB0.qnt1.ratio <- list( 1.1, c(0.9,1.1))
TB0.qnt2.ratio <- list( 1.1, c(0.9,1.1))

CR0.qnt1.ratio <- list( 1.1, c(0.85,1.15))
CR0.qnt2.ratio <- list( 1.1, c(0.85,1.15))

TB1.qnt1.ratio <- list( 1.1, c(0.90,1.1))
TB1.qnt2.ratio <- list( 1.1, c(0.90,1.1))

CR1.qnt1.ratio <- list( 1.1, c(0.85,1.15))
CR1.qnt2.ratio <- list( 1.1, c(0.85,1.15))

BCG.qnt1.ratio <- list( 1.1, c(0.75,0.90) )
BCG.qnt2.ratio <- list( 1.1, c(0.75,0.90) )

TB01.qnt1.ratio <- 1.1
TB01.qnt2.ratio <- 1.1

CR01.qnt1.ratio <- 1.1
CR01.qnt2.ratio <- 1.1

BCG.qnt1.init <- rep( seq(8,3.1,length=6), 2)
BCG.qnt2.init <- rep( seq(15,6,length=6), 2)

source(file="h:\\Statistics\\TBmixtures\\MS.r")
source(file="h:\\Statistics\\TBmixtures\\BCG.r")
```

14.2.2 Results

```
Parameter estimates: non-BCG

 Prevalence of TB infections (p0.TB)
          mean  st.dev   2.5%    50% 97.5%
M 0-4   0.0435 0.00682 0.0307 0.0432 0.058
M 5-9   0.1892 0.01686 0.1580 0.1888 0.223
M 10-14 0.5034 0.02327 0.4575 0.5042 0.547
M 15-19 0.7444 0.01714 0.7105 0.7444 0.776
M 20-24 0.8617 0.01719 0.8270 0.8622 0.894
M 25-29 0.9206 0.01042 0.8992 0.9208 0.939
F 0-4   0.0639 0.00833 0.0488 0.0635 0.081
F 5-9   0.1882 0.01609 0.1570 0.1881 0.221
F 10-14 0.5550 0.02111 0.5129 0.5551 0.596
F 15-19 0.7234 0.01847 0.6872 0.7232 0.758
F 20-24 0.7797 0.01736 0.7440 0.7801 0.812
F 25-29 0.7735 0.01486 0.7440 0.7737 0.802

 TB distribution (1st quantile, TB0.qnt1)
        mean st.dev 2.5%  50% 97.5%
M 0-4   15.7  0.569 14.9 15.6  17.3
M 5-9   16.4  0.430 15.7 16.4  17.3
M 10-14 17.3  0.213 16.9 17.3  17.8
M 15-19 17.2  0.175 16.9 17.2  17.6
M 20-24 17.1  0.146 16.8 17.1  17.4
M 25-29 17.0  0.159 16.7 17.0  17.3
F 0-4   16.3  0.532 15.5 16.2  17.5
F 5-9   16.8  0.294 16.3 16.7  17.4
F 10-14 18.0  0.187 17.7 18.1  18.4
F 15-19 18.4  0.182 18.0 18.4  18.7
F 20-24 18.6  0.151 18.3 18.6  18.9
F 25-29 17.6  0.185 17.2 17.6  18.0

Parameter estimates: BCG

 Prevalence of TB infections (p1.TB)
          mean st.dev    2.5%    50%  97.5%
M 0-4   0.0404 0.0239 0.00303 0.0384 0.0903
M 5-9   0.1726 0.0232 0.12773 0.1727 0.2155
M 10-14 0.5258 0.0220 0.47937 0.5268 0.5665
M 15-19 0.6506 0.0229 0.60609 0.6513 0.6950
M 20-24 0.8286 0.0226 0.78295 0.8295 0.8699
M 25-29 0.9414 0.0155 0.90787 0.9424 0.9682
F 0-4   0.1023 0.0409 0.02623 0.1024 0.1821
F 5-9   0.2278 0.0294 0.16806 0.2282 0.2824
F 10-14 0.5766 0.0205 0.53627 0.5771 0.6146
F 15-19 0.6372 0.0214 0.59354 0.6377 0.6789
F 20-24 0.7090 0.0220 0.66593 0.7088 0.7530
F 25-29 0.8185 0.0243 0.77000 0.8193 0.8627

 TB distribution (1st quantile, TB1.qnt1)
        mean st.dev 2.5%  50% 97.5%
M 0-4   16.2  0.695 15.1 16.1  17.9
M 5-9   16.6  0.452 15.7 16.6  17.5
M 10-14 15.9  0.198 15.6 15.9  16.4
M 15-19 15.8  0.166 15.4 15.7  16.1
M 20-24 15.8  0.169 15.5 15.8  16.2
M 25-29 16.7  0.220 16.2 16.7  17.1
F 0-4   16.4  0.875 14.9 16.4  18.2
F 5-9   16.4  0.468 15.5 16.4  17.3
F 10-14 16.5  0.171 16.2 16.5  16.8
F 15-19 16.8  0.168 16.5 16.8  17.1
F 20-24 17.1  0.157 16.8 17.1  17.4
F 25-29 16.5  0.290 16.0 16.4  17.1
```

Table 14-3 Model checks for Application 9 (non-BCG group)

```
Non-BCG: summary of predictive checks (+ predictions too large, - predictions too small)
    11.7 % predictive failures

1: M 0-4 2: M 5-9 3: M 10-14 4: M 15-19 5: M 20-24 6: M 25-29 7: F 0-4 8: F 5-9 9: F 10-14 10:
F 15-19 11: F 20-24 12: F 25-29

      1  2  3  4  5  6  7  8  9 10 11 12 total
1mm   0  0  0  0  0  0  0  0  0  0  0  0     0
2mm   0  0  0  0  0  0  0  0  0  0  0  0     0
```

3mm	1	0	1	0	0	0	1	0	0	0	0	0	3
4mm	0	0	0	0	0	0	0	0	0	0	0	0	0
5mm	0	-1	0	0	0	0	0	0	0	-1	0	0	2
6mm	0	0	0	0	-1	0	-1	-1	-1	0	-1	-1	6
7mm	0	0	-1	0	0	0	0	0	0	-1	0	-1	3
8mm	0	0	0	0	0	0	0	0	0	0	0	0	0
9mm	0	0	0	0	0	1	0	0	0	0	0	0	1
10mm	0	0	0	0	0	0	0	0	0	0	0	0	0
11mm	0	0	0	0	0	0	0	0	0	0	0	0	0
12mm	0	0	0	0	0	0	0	0	0	0	0	0	0
13mm	0	0	0	1	0	0	1	0	1	0	1	0	4
14mm	0	0	1	1	0	0	0	0	1	0	0	0	3
15mm	0	0	0	0	0	-1	0	0	1	0	0	0	2
16mm	0	0	0	0	0	0	0	1	0	0	0	1	2
17mm	0	0	0	-1	0	0	0	0	0	0	0	-1	2
18mm	0	0	0	0	0	0	0	0	0	0	0	0	0
19mm	0	0	0	0	0	0	0	0	0	0	0	0	0
20mm	0	-1	0	-1	-1	0	0	0	-1	-1	-1	-1	7
21mm	0	0	0	0	1	1	0	0	0	0	0	0	2
22mm	0	0	0	0	0	0	0	0	0	0	0	0	0
23mm	0	0	0	0	0	0	0	0	0	0	0	0	0
24mm	0	0	0	0	1	0	0	0	0	0	0	1	2
25mm	0	0	0	0	0	0	0	0	0	0	0	0	0
26mm	0	0	0	0	0	0	0	0	0	0	0	0	0
27mm	0	0	0	0	0	0	0	0	0	0	0	0	0
28mm	0	0	0	0	0	0	0	0	0	0	0	0	0
29mm	0	0	0	0	0	0	0	0	0	0	0	0	0
30mm	0	0	0	-1	0	0	0	0	-1	-1	0	0	3
total	1	2	3	5	4	3	3	2	6	4	3	6	42

Table 14-4 Model checks for Application 9 (BCG group)

BCG: summary of predictive checks (+ predictions too large, - predictions too small)
8.6 % predictive failures

1: M 0-4 2: M 5-9 3: M 10-14 4: M 15-19 5: M 20-24 6: M 25-29 7: F 0-4 8: F 5-9 9: F 10-14 10: F 15-19 11: F 20-24 12: F 25-29

	1	2	3	4	5	6	7	8	9	10	11	12	total
1mm	0	0	0	0	0	0	0	0	0	0	0	0	0
2mm	0	0	0	0	0	0	0	0	0	0	0	0	0
3mm	0	1	0	0	0	0	0	1	1	0	0	0	3
4mm	0	-1	0	0	0	0	0	0	0	0	0	0	1
5mm	-1	0	0	0	0	0	0	-1	0	0	0	0	2
6mm	0	0	1	0	0	0	0	0	0	0	0	0	1
7mm	0	0	0	0	0	0	0	0	0	0	0	0	0
8mm	0	0	0	0	0	0	0	1	0	0	0	0	1
9mm	1	0	0	0	0	0	1	0	0	0	0	0	2
10mm	0	0	-1	-1	0	0	-1	-1	0	0	0	0	4
11mm	0	0	0	0	0	0	0	0	0	0	0	0	0
12mm	0	-1	0	0	0	0	0	0	0	-1	0	0	2
13mm	0	0	0	0	0	0	0	0	0	0	0	0	0
14mm	0	0	0	0	0	0	0	0	0	0	0	0	0
15mm	0	0	-1	-1	0	0	0	0	0	0	0	0	2
16mm	0	1	0	0	0	0	0	0	0	0	0	0	1
17mm	0	0	0	0	0	0	0	0	0	0	0	0	0
18mm	0	0	0	0	0	0	0	0	0	0	0	0	0
19mm	0	0	0	0	0	0	0	-1	0	0	0	0	1
20mm	0	0	0	0	0	0	0	-1	-1	-1	0	-1	4
21mm	0	0	0	0	0	0	0	1	1	0	0	0	2
22mm	0	0	1	0	0	0	0	0	0	0	0	0	1
23mm	0	0	0	0	0	0	0	0	0	0	1	0	1
24mm	0	0	0	0	0	0	0	0	0	0	0	1	1
25mm	0	0	-1	0	0	0	0	0	0	0	0	0	1
26mm	0	0	0	0	0	0	0	0	0	0	0	0	0
27mm	0	0	0	0	0	0	0	0	0	0	0	0	0
28mm	0	0	0	0	0	0	0	0	0	0	0	0	0
29mm	0	0	0	0	0	0	0	0	0	0	0	0	0
30mm	0	0	0	0	0	0	0	0	0	0	0	-1	1
total	2	4	5	2	0	0	2	7	3	2	1	3	31

15 Summary and Recommendations

Mixture analyses for induration data have been presented for

- unvaccinated subjects from one group (Section 4), from several groups using the basic model (Section 7), and from several groups using the extended model (Section 10).

- unvaccinated and vaccinated subjects from one group (Section 12), from several groups using the basic model (Section 13), and from several groups using the extended model (Section 14).

Mixture analysis for data from one group can be difficult if considerable overlap of mixture component distributions is present. If data from several groups is available, and prevalences across groups differ, mixture analysis using the basic model is more likely to be successful.

The statistical approach used is Bayesian and uses simulation techniques (Markov Chain Monte Carlo based on the Metropolis algorithm). Although this allows to fit the relatively complex mixture models, it comes at the price: convergence of the simulation algorithm needs to be checked, and stability of results can only be expected for sufficiently large simulation sample sizes.

- Look at the data first (histograms)
- Perfrom mixture analyses using the basic model (Section 7 or 13) and check whether the model fits the data reasonably well.
- A re-fined analysis may include different combinations of mixture component distributions (Wb,Ln,N). Pick the best of these models.
- If the model does not fit the data well enough, think about model extensions (Sections 10 or 14).
- If you have a model that fits the data reasonably well, perform a final analysis using a large number of simulations. This may require some time, but it makes sure that the results you get are stable.

If you encounter problems while running the program

- check the log-file to see where the program failed;
- check your input program accordingly;
- sometimes it helps to restart the program or to clear the memory (in the R-Menu list, select *Misc-Remove all objects*).

16 Appendix

16.1 Data sets

16.1.1 Data set 1: Korea75m.asc

korea75m.asc

0	678	282	125	82	18	12
1	67	27	16	10	2	3
2	133	114	75	45	23	22
3	32	29	15	10	7	4
4	18	13	11	8	5	5
5	4	7	2	6	2	5
6	3	4	4	1	5	3
7	3	1	6	7	2	3
8	1	3	4	4	2	5
9	1	7	3	6	5	2
10	4	3	2	11	12	10
11	3	1	4	9	7	15
12	0	3	5	13	18	23
13	1	5	11	11	13	27
14	2	4	7	20	18	40
15	1	8	16	43	30	73
16	4	9	19	39	35	59
17	3	6	26	56	50	66
18	4	9	23	53	38	71
19	3	13	28	51	43	64
20	3	17	24	51	48	58
21	4	5	18	23	14	23
22	0	3	14	24	13	21
23	1	2	15	12	9	9
24	0	2	5	11	1	8
25	1	1	6	6	3	5
26	0	2	1	3	2	2
27	1	0	1	4	0	2
28	0	0	1	0	0	1
29	0	0	2	2	1	1
30	0	0	1	2	0	0

16.1.2 Data set 2: Korea75mf.asc

korea75mf.asc

678	282	125	82	18	12	626	273	111	69	61	62
67	27	16	10	2	3	37	44	21	12	13	11
133	114	75	45	23	22	129	116	65	56	36	61
32	29	15	10	7	4	24	32	17	24	15	20
18	13	11	8	5	5	9	9	9	7	7	17
4	7	2	6	2	5	5	7	4	9	5	3
3	4	4	1	5	3	4	5	6	2	7	8
3	1	6	7	2	3	2	2	4	10	3	10
1	3	4	4	2	5	0	5	1	6	4	2
1	7	3	6	5	2	2	3	5	3	4	8
4	3	2	11	12	10	1	2	5	3	12	17
3	1	4	9	7	15	1	3	3	5	4	10
0	3	5	13	18	23	5	5	8	8	7	21
1	5	11	11	13	27	0	7	2	16	8	26
2	4	7	20	18	40	1	4	5	16	18	32
1	8	16	43	30	73	2	5	6	25	28	47
4	9	19	39	35	59	3	3	22	21	38	33
3	6	26	56	50	66	5	6	17	27	42	76
4	9	23	53	38	71	7	15	29	37	46	56
3	13	28	51	43	64	4	11	33	46	59	69

```
  3  17  24 51 48 58   7  11  43 81 72 72
  4   5  18 23 14 23   3   6  23 42 39 47
  0   3  14 24 13 21   3   6  26 30 44 25
  1   2  15 12  9  9   2   5  11 25 29 19
  0   2   5 11  1  8   1   5  14 17 12  8
  1   1   6  6  3  5   1   1  10 13 19 12
  0   2   1  3  2  2   1   0   2  7  3  8
  1   0   1  4  0  2   1   0   3  3  4  4
  0   0   1  0  0  1   0   1   0  5  1  1
  0   0   2  2  1  1   0   0   3  3  1  1
  0   0   1  2  0  0   0   0   4  6  1  0
```

16.1.3 Data set 3: Korea75mfBCG.asc

korea75mfBCG.asc

```
344 486 255 117  26   3 286 408 215 103  53  19
 30  67  41  14   1   2  27  63  49  20  17   3
142 316 209  91  17   3 156 257 195 109  51  18
 51  86  66  34   9   1  39  68  51  38  15   4
 32  59  40  21   6   1  21  41  28  11   8   2
 33  41  39  22   2   3  12  36  23  16   7   8
 23  25  16  20   5   1  12  23  21  16   7   2
 19  32  39  20   5   3  13  19  24  22   6   4
 15  34  35  35  10   1  26  17  35  18  13   6
 13  36  38  22   9   5  10  29  32  25   9   9
 24  37  70  52  19  11  31  53  51  33  13   8
 15  32  60  48  16  11  19  33  54  29  20  11
 28  58  86  51  23   8  22  43  59  57  19  17
 21  39  66  66  28  20  14  39  70  46  25  15
 13  45  83  53  27  19  13  40  71  42  24  13
 27  53 112  91  48  32  27  44 100  57  38  19
 16  30  87  76  39  24  20  42  96  57  31  23
 22  34  95  61  29  26  17  30  92  63  37  18
  9  48  67  56  33  27  16  32  86  57  55  21
 12  37  83  52  32  35  18  51  79  51  38  25
 10  34  66  53  25  25  13  43  88  63  45  26
  6  18  40  22  15  13   8   8  26  35  28  15
  4  12  17  13   4   6   3   8  29  23  23   9
  1   2  15  13   2   4   5   6  19  17   6   6
  2   3   8   3   3   4   2   4  14   9   5   1
  2   4  15   5   0   2   0   5   8   6   5   5
  0   1   2   1   0   1   1   2   8   2   3   1
  0   0   0   2   0   0   0   0   4   6   1   0
  1   1   0   1   0   0   0   1   1   1   0   1
  0   1   1   0   0   0   0   0   2   0   1   0
  0   0   0   0   0   0   0   0   2   1   0   3
```

16.1.4 Data set 4: Navy.asc

navy.asc

```
alabama california florida georgia
7183   53755   12689   9823
120    696       262    197
144    754       375    253
125    479       284    198
127    380       261    137
92     414       196    118
96     419       129    108
70     427       101     74
63     385        75     57
68     385        63     55
27     309        27     31
11     220        17     26
6      133         6      7
2      57          4      2
2      29          1      2
```

```
3       15          0       1
```

16.2 Applications

Table 16-1 An overview of the 9 applications

	Description/Description	Section
Application 1	Korean data Males age 5-9 yrs Analysis of a single frequency distribution	4
Application 2	Korean data Males: six age groups Analysis of six groups with basic model	7.4
Application 3	Korean data Males: six age groups Analysis of six groups with extended model	10.1
Application 4	Korean data Males and females: six age groups each Analysis with extended model	10.2
Application 5	Navy data Analysis with grouped induration data	11
Application 6	Korean data Males aged 5-9 yrs Unvaccinated and vaccinated subjects Analysis of a single frequency distribution	12
Application 7	Korean data Males: six age groups Unvaccinated and vaccinated subjects Analysis of six groups with basic model	13
Application 8	Korean data Males: six age groups Unvaccinated and vaccinated subjects Analysis of six groups with extended model	14.1
Application 9	Korean data Males and females: six age groups each Unvaccinated and vaccinated subjects Analysis with extended model	14.2

The following files contain the R-program for performing the mixture analysis, the main output file “Out1”, and the two additional output files (“Out2”, “Out3”) with the fitted indurations and model checks.

16.2.1 Application 1

16.2.2 Application 2

16.2.3 Application 3

16.2.4 Application 4

16.2.5 Application 5


App5.r


App5Out1.txt


App5Out2.txt


App5Out3.txt

16.2.6 Application 6


App6.r


App6Out1.txt


App6Out2.txt


App6Out3.txt

16.2.6.1 Application 7


App7.r


App7Out1.txt


App7Out2.txt


App7Out3.txt

16.2.7 Application 8


App8.r


App8Out1.txt


App8Out2.txt


App8Out3.txt

16.2.8 Application 9


App9.r


App9Out1.txt


App9Out2.txt


App9Out3.txt

16.3 Program details: files

Table 16-2 Program Files

File	Description
MS.r	The basic file responsible for the Metropolis sampling algorithm
NONBCG.r	Mixture analysis for induration data of non-vaccinated subjects. One input file is needed.
BCG.r	Mixture analysis for induration data of non-vaccinated and vaccinated subjects. Two input files of identical structure are needed.
IndurationPlot.r	Displays histograms of induration data

16.3.1 Input file

Table 16-3 Input for NONBCG and BCG analysis

Input	Description	Default
`infile`	A string pointing to the input file. If the file is not located in the working directory, the full path is needed. For BCG analysis, two input files are required (`infile0`, `infile1` for non-BCG and BCG data, respectively).	mandatory
`outfile`	A string denoting the output file. Four output files will be created by appending. For example, if `outfile <- "out"` the output files will be `out.log, out1.txt, out2.txt, out3.txt`.	
`freq.column`	Either a single number (if only one group is to be analyzed), or a vector pointing to the columns in the input file (`infile`). • `freq.column <- 3` • `freq.column <- c(1,2,3)` • `freq.column <- seq(1,4)` For BCG analysis, `freq.column0` and `freq.column1` are required.	
`group.names`	• a vector of group names (of length equal to the number of groups given in freq.column), or • a list if the groups arise from cross-classified data. Example:	

	group.names <- list(c("sex","age"), c("males", "females", c("0-4 yrs", "5-9 yrs", "10-14 yrs")) is an alternative way to specify group.names <- c("males 0-4 yrs", "males 5-9 yrs", "males 10-14 yrs", "females 0-4 yrs", "females 5-9 yrs", "females 10-14 yrs")	
distTB	The type of distribution for TB infections. Wb (for Weibull), LN (for log-normal), or N (normal)	Mandatory
distCR	The type of distribution for cross-reactions: Wb (for Weibull), LN (for log-normal), or N (normal)	Mandatory
distBCG	The type of distribution for cross-reactions: Wb (for Weibull), LN (for log-normal), or N (normal)	Mandatory for BCG program
MS.run MS.results MS.check MS.graph MS.graph.page	MS.run, MS.results, MS.check are either T (TRUE) or F (FALSE). They will be executed if set to T. MS.run: execution of the sampling algorithm. MS.results: processing of results. MS.check: executes posterior predictive model checks. MS.graph: produces graphical output. 5 different types of graphs are supported: [1] MCMC plots (used to visually assess the convergence of the Metropolis sampler. [2] Histograms of posterior distributions for model parameters. [3] Probability of infection (as a function of induration) [4] Confidence interval plots (only if more than one group is analyzed) [5] Data histograms with model fits (component distributions). By default: only [1] and [5] are presented, i.e., MS.graph <- c(T,F,F,F,T) Note that if MS.graph is not specified as a vector of size 5, the input for MS.graph is replicated 5 times: MS.graph <- T for all , MS.graph <- F for no graphs.	MS.run TRUE MS.results TRUE MS.check FALSE MS.graph (T,F,F,F,T)
MS.size	A vector of 3 elements denoting the length of the sampling algorithm MS.size[1]: the length of the initial phase (burn-in) of the algorithm. Values from this phase will not be saved. MS.size[2]:	

	the size of the sample to be used for the analysis. `MS.size[3]`: the thinning of the sample. Example: `MS.size <- c( 1000, 2000, 5)` The algorithm will run for 11000 iterations. An initial phase of 1000 iterations, then 10000 iterations, only every 5th iteration will be saved, i.e., the sample size used for the analysis will be 2000.	
`RndSeed`	A random seed for the random number generators.	

`source()`
`MS.r`
`NONBCG.r`
`BCG.r`
`IndurationPlot.r`
	Two programs (`MS` and `NONBCG`, or `MS` and `BCG`) need to be sourced at the end of the input file:
	`source(file="h:\\Statistics\\TBmixtures\\MS.r")`
	`source(file="h:\\Statistics\\TBmixtures\\NONBCG.r")`
	Note that the full path must be given if the programs do not reside in the current working directory.	

`qnt1TB qnt2TB qnt1CR qnt2CR` `qnt1BCG, qnt2BCG`	The quantiles of the TB and CR mixture component distributions. BCG analysis only	0.50 0.95 0.50 0.95

`TB.qnt1.init`
`TB.qnt1.int`
`TB.qnt1.diff`
`TB.qnt1.ratio`
`same for CR and BCG and qnt2`
	Initial values (.init), interval constraints (.int), constraints on differences (.diff), constraints on ratios (.ratio). (See examples).
	For BCG analysis use TB0, TB1, CR0 and CR1 for the non-BCG and BCG groups, respectively	

16.3.2 Output files

Table 16-4 Program Output

File name	Description
Outfile.log	Output log-file with text output about current state of program
Outfile1.txt	Main output with results from mixture analysis
Outfile2.txt	Output file with fitted values
Outfile3.txt	Output file with model checks

16.3.3 IndurationPlot function

Table 16-5 IndurationPlot Function

Syntax	`IndurationPlot(infile, freq.column, group.labels, header = F, page.layout, zeromm = "no", perc = F, xlab = "Induration (mm)", ylab = "", plot.type = "histogram", pch = 16)`	
Input argument	Description	Default
`Infile`	A string pointing to the input file (an ASCI file)	mandatory input
`freq.column`	A vector of column indices	Data from all columns will be displayed
`group.labels`	A vector of group labels of the same length as `freq.column`	"Group 1", "Group 2"… For only one group, no label
`Header`	`T` if a header (first line in infile) is present, `F` if not	`F`
`page.layout`	A 2-vector denoting the number of rows and columns used for graphical display	Depends on the number of distributions (specified in `freq.column`) to be displayed
`Zeromm`	Options for zero induration frequencies • `"no"` for not displaying zero induration frequencies • `"yes"` for displaying zero indruation frequencies • `"text"` for not displaying zero induration frequencies (a text message is added in figure)	`"no"`
`Perc`	`T` for displaying frequencies as percentages, `F` for absolute frequencies	`F`

`xlab`	Label for x-axis	`"Induration (mm)"`
`ylab`	Label for y-axis	No label
`plot.type`	• `"histogram"` standard histogram display • `"polygon"` histogram displayed as a polygon plot • `"polygon.points"` histogram displayed as a polygon with points	`"histogram"`
`pch`	Symbol type in `"polygon.points"`	`pch=16` (a solid circle)

16.4 Program details: parameters

16.4.1 Model parameters

Table 16-6 Parameter naming conventions

	Program		
Parameter	NONBCG	BCG	
		unvaccinated	vaccinated
TB infection prevalence	p.TB	p0.TB	p1.TB
Zero induration prevalence	p.zero	p0.zero	p1.zero
BCG prevalence			p1.BCG
Distribution of TB infection (1^{st} and 2^{nd} quantile)*	TB.qnt1 TB.qnt2	TB0.qnt1 TB0.qnt2	TB1.qnt1 TB1.qnt2
Distribution of cross-reactions (1^{st} and 2^{nd} quantile)*	CR.qnt1 CR.qnt2	CR0.qnt1 CR0.qnt2	CR1.qnt1 CR1.qnt2
Distribution of BCG reactions (1^{st} and 2^{nd} quantile)*		BCG.qnt1 BCG.qnt2	

* Default: 50% and 95% quantiles

16.4.2 Initial values for parameters

For all model parameters in Table 16-6, initial values can be specified. They are specified using the .init extension. For example

```
TB.qnt1.init <- 17
```

assigns the initial value 17 to the first (50%) quantile of the TB distribution.

16.4.3 Parameter constraints

Parameter constraints can be of the following type:

- **Interval constraints**, forcing the corresponding parameter to be within the boundaries given by the interval (applicable for all models)
- **Difference constraints**, specifying the maximum difference for parameter values across groups (for extended models of Section 10 and 14)
- **Ratio constraints**, specifying the maximum ratio for parameter values across groups (for extended models of Section 10 and 14)

For details regarding the constraints see Table 16-7 and 16.8 for the NONBCG and BCG program, respectively.

Table 16-7 Parameter constraints (NONBCG program)

TB.qnt1.int	Interval constraint for parameter. Example: `TB.qnt1.int <- c(16,18).` The median of the distribution of TB infections is between 16 and 18mm.
TB.qnt1.diff	One-number: Maximum difference between parameter values for different groups. Example: `TB.qnt1.diff <- 2.` The difference in medians for the distribution of TB infections over the different groups is at most 2mm. Two-numbers: Difference of parameter values between adjacent groups (current group minus previous group). Example: `TB.qnt1.diff <- c(-1,0)` Medians for the distribution of TB infections between adjacent groups are decreasing by at most 1mm.
TB.qnt1.ratio	One-number: Maximum ratio between parameter values for different groups Example: `TB.qnt1.ratio <- 1.1` Medians for the distribution of TB infections differ by at most 10%. Two-numbers: Ratio of parameter values for adjacent groups (current group

	divided by previous group) Example: `TB.qnt1.diff <- c(0.90,0.95)` Medians for the distribution of TB infections between adjacent groups are decreasing by 5% to 10%.

Table 16-8 Parameter constraints (BCG program)

TB0.qnt1.int	Interval constraint for parameter. Example: `TB0.qnt1.int <- c(16,18).` The median of the distribution of TB infections is between 16 and 18mm.
TB0.qnt1.diff	One-number: Maximum difference between parameter values for different groups. Example: `TB0.qnt1.diff <- 2.` The difference in medians for the distribution of TB infections over the different groups is at most 2mm. Two-numbers: Difference of parameter values between adjacent groups (current group minus previous group). Example: `TB0.qnt1.diff <- c(-1,0)` Medians for the distribution of TB infections between adjacent groups are decreasing by at most 1mm.
TB0.qnt1.ratio	One-number: Maximum ratio between parameter values for different groups Example: `TB0.qnt1.ratio <- 1.1` Medians for the distribution of TB infections differ by at most 10%. Two-numbers: Ratio of parameter values for adjacent groups (current group divided by previous group) Example: `TB0.qnt1.diff <- c(0.90,0.95)` Medians for the distribution of TB infections between adjacent groups are decreasing by 5% to 10%.
TB01.qnt1.diff	Difference (for 1st quantile of TB distributions) between unvaccinated (non-BCG) and vaccinated (BCG).
TB01.qnt1.ratio	Ratio (for 1st quantile of TB distributions) between unvaccinated (non-BCG) and vaccinated (BCG).
CR01.qnt1.diff	Difference (for 1st quantile of CR distributions) between unvaccinated (non-BCG) and vaccinated (BCG).
CR01.qnt1.ratio	Ratio (for 1st quantile of CR distributions) between unvaccinated (non-BCG) and vaccinated (BCG).

16.5 Program details: output to R-object TBmix

Table 16-9 R-object TBmix from NONBCG.r and BCG.r program

Components of object	Description
TBmix$Program	Program name and version
TBmix$fit	Information from Metropolis Sampling algorithm • **TBmixfitSample** sampled values, a list • **TBmixfitLoglikLogpostSample** sampled values for log-likelihood and log-posterior • **TBmixfitEstimates** approximate maximum-likelihood and maximum posterior estimates • **TBmixfitDeviance** mean deviance at deviance at maximum-likelihood and maximum posterior estimates • **TBmixfitInput** main input specification for MS.r; see TBmix$MSinfo for more detailed information • **TBmixfitOutput** last sampled value and scaling information; see TBmix$MSinfo for more detailed information.
TBmix$pred.table	Table of posterior predictive model checks as shown in output file 1 and 3
TBmix$MSinfo	Detailed parameters of the Metropolis Sampler MS.r. The ones in boldface can be changed in the Input Program: **MS.size, MS.block,** MS.low, MS.high, **MS.adj, MS.block, MS.screen, MS.screen.dig,** MS.names, **MS.InitialValues, MS.run, MS.results, MS.check, MS.graph, MS.graph.page**
TBmix$constraints	Parameter constraints: from .int, .diff, .ratio specifications
TBmix$files	Name of input and output files
TBmix$dist	Distributions for TB, CR and BCG mixture components
TBmix$Data	Input data sets
TBmix$p.TB.summary, TBmix$p.zero.summary …	Posterior summaries for model parameters
TBmix$mpe	Maximum posterior estimate
TBmix$LastValues	Last sampled parameter values If the algorithm needs to be restarted from these values, use **MS.InitialValues <- TBmix$LastValues** in Input Program
TBmix$RndSeedAtStart	Random seed at start of algorithm

References

Berry, D. (2006). Bayesian Clincial Trials, *Nature Reviews*, 5:27-36.

Couzin, J. (2004). The new math of clinical trials. *Science*, 6 Feb 2004, 303: 784-786.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004). *Bayesian Data Analysis*, Chapman & Hall.

Goodman, S.N. (1999). Towards evidence-based medical statistics 1: the p-value fallacy. *The Annals of Internal Medicine*, 130:995-1004.

Goodman, S.N. (1999). Towards evidence-based medical statistics 2: the Bayes factor. *The Annals of Internal Medicine*, 130:1005-1013

Goodman, S.N. (2005). Introduction to Bayesian methods I: measuring the strength of evidence. *Clinical Trials*, 2:282-290.

Malakoff, D. (1999). Bayes offers a 'new way' to make sense of numbers. *Science*, Nov 19, 1999.

Neuenschwander, B., Zwahlen, M., Kim, S.J., Engel, R.R., and Rieder, H.L. (2000). Trends in the prevalence of infection with *Mycobacterium tuberculosis* in Korea from 1965 to 1995: an analysis of seven surveys by mixture models. *INT J TUBERC DIS* 4(8): 719-729.

Neuenschwander, B., Zwahlen, M., Kim, S.J., Lee, E.G., and Rieder, H.L. (2002). Determination of the prevalence of infection with *Mycobacterium tuberculosis* among persons vaccinated against Bacillus Calmette-Guérin in South Korea. *American Journal of Epidemiology*, 155(7): 654-663.

O'Hagan, A., Luce, B.R. (2003). A Primer on Bayesian Statistics in Health Economics and Outcomes Research.

Spiegelhalter, D.J., Abrams, K.R., Myles, J.P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley.