Empirical Bayes methods for missing data analysis

John Brandel

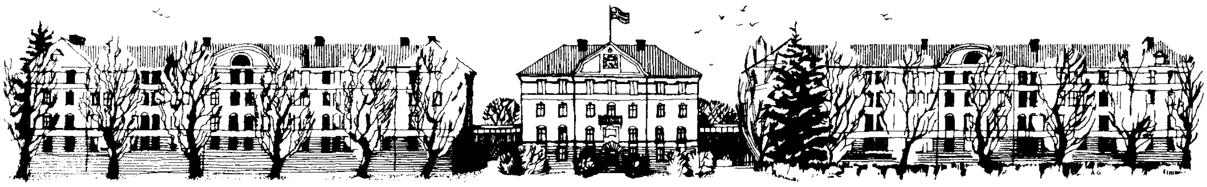
U.U.D.M. Project Report 2004:11

Examensarbete i matematisk statistik, 20 poäng

Handledare: Silvelyn Zwanzig, Uppsala universitet och Johan Bring, Statisticon

Examinator: Silvelyn Zwanzig

Juni 2004



Department of Mathematics

Uppsala University

Empirical Bayes methods for missing data analysis

John Brandel

June 6, 2004

Abstract

Missing data in clinical trials can ruin the significance of test results and violates the Intention To Treat principle. Therefore imputation methods are an important tool that is used to deal with this problem. In this essay two different Bayesian imputation methods will be compared to four different general imputation methods. The general methods are the LOCF, worst case, best case and a mean value method. The Bayesian methods are Empirical Bayes method and a method using the Expectation Maximization algorithm. From the results of analysis made on the data sets used here it can be concluded that the Bayesian methods give better estimates than general imputation methods. This conclusion is valid for different parameter values such as variance, percentage missing values and number of observations per patient.

Acknowledgment

I would like to thank my supervisor and examinator Silvelyn Zwanzig at the Department of Mathematics for answering questions and giving a lot of support. I also want to thank Ass. professor Johan Bring at Statisticon for introducing me to the interesting subject and for answering questions. Finally I want to thank my fellow student Martin Elfsberg for support and interesting discussions.

Contents

1 Introduction 5

2 Method for the clinical trial 6

3 General imputation methods 7

3.1 Excluding non-complete data 7

3.1.1 The LOCF method 8

3.1.2 Best or worst case imputation 9

3.1.3 Mean value methods 9

3.1.4 Regression methods 10

4 Empirical Bayes methods 10

4.1 Empirical Bayes theory 10

4.2 Assumptions 12

4.3 The beta/binomial model 12

4.3.1 Moment Estimates 14

4.3.2 Maximum likelihood estimation 15

4.4 The Gaussian/Gaussian model 17

4.4.1 Estimates 18

4.5 The EM-algorithm 19

5 Data analysis 20

5.1 Analysis of trial data set 20

5.2 Analysis of simulations 21

5.2.1 A small simulation 22

5.2.2 Large data set simulation 24

6 Discussion 26

7 Conclusions 27

8 References 29

A Appendix 30

A.1 The beta distribution 30

A.2 Small simulation 31

1 Introduction

In clinical trials there often exists missing data. There can be different reasons to why these occur. A few examples can be that some patients get to ill to attend to a meeting, patients might refuse to continue the study (dropouts) or there might be treatment failures. For a small number of unrecorded data there is no problem but if you have a substantial amount of them problems will occur. Another complication is non-compliant patients (i.e. patients who in some way does not follow the test as they should, which can be that they don't take the drug twice per day as is the instructions for the test or they simply don't take the drug at all).

If the clinic trial contains missing data, dropouts or non-compliant patients and the patients that have these incorrect protocol values are excluded the trial is not following the intention to treat (ITT) principle. There are demands that all clinical trials should be tested with an ITT analysis[1]. The intention to treat analysis requires that patients must be analyzed in the groups they were randomized into, regardless of whether they complied with the treatment they were given or not. It is also required that the outcome data is complete. Imputation methods is therefore needed to make most clinical trials approved by the intention to treat principle.

There are two different purposes with a method for imputation of missing data. One is that the company analyzing a drug wants the missing data to be replaced by such a correct value as possible, so that the result of the analysis become accurate. This is of course very important for the company because it is likely that there is some missing values which can not be allowed to ruin the test significance.

Another purpose for imputation methods is for them to be conservative. These methods are made to make it harder for the company or the trial to show significant differences between the test groups. If the company wants to show that group A has a better development than group B a conservative imputation method makes it harder to show significant differences, but if it does it is more reliable than if we would have used a non-conservative imputation method.

One thing to have in mind is that test individuals with missing values might be more likely to have bad protocol values compared to the others. A method which then replace the missing value with a value estimated as a mean value from the individuals test group can gain the company. A method that is very

[1] emea (2001), p. 1

conservative is the worst case method which is discussed in subsection 3.1.2. So either we can use imputation to get a good estimated value or to get a conservative.

The main purpose with this essay is to evaluate if Bayesian imputation methods can be used to get better estimated values than general methods. I am going to analyze a method using Empirical Bayes (EB) analysis and a method using the Expectation Maximization (EM) algorithm. These will be analyzed in a certain case involving Parkinson patients and when data sets are generated randomly. Presumably we will find that the EB and EM methods give good estimated values for the missing values while it might not be very conservative. These will be analyzed in comparison with other imputation methods. Two different models of Bayesian analysis will be analyzed and compared to each other. These are the Gaussian/Gaussian model and the beta/binomial model and these come from making different assumptions regarding the prior and likelihood distribution.

2 Method for the clinical trial

The clinical trial we are going to look at is one involving Parkinson patients. Video recordings of each patient were made 17 occasions per day at two different occasions. This procedure was done twice, once with an 'old' drug and once with the 'drug' that hopefully turns out to be better. These video clips were then studied by two doctors that observe the clips independently of each other. For every observation (video clip) the patients were given a number between -3 and +3 on a degree of how ill they were. A -3 states that they can hardly move at all and a +3 states the shaking is very bad, i.e. the closer to zero the better the state of the patient. When the patient has a number between -1 and 1 they are considered to be quite functional. So the idea with the test is to find out for how many percent of the treatment time each patient is in this interval. This proportion of time is denoted $\hat{\pi}_i$ for the i:th patient, $i = 1, \ldots, N$ where N is the number of patients. x_i is the number of observations that the individual i is in the 'good interval' and n_i is the number of observations for $i = 1, \ldots, N$. Clearly

$$\hat{\pi}_i = \frac{x_i}{n_i}. \tag{1}$$

A weighted estimate for the overall mean of all patients is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} n_i \hat{\pi}_i}{\sum_{i=1}^{N} n_i} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} n_i}. \tag{2}$$

When the number of observations n_i for a patient is high this estimate is quite good, but if we have a low number of observations it can be a poor estimate of the 'true' π_i. Then we wish to use an estimator $\tilde{\pi}_i$ that uses the overall mean estimate $\hat{\mu}$. This is what Empirical Bayes methods does. A formula for the so called *shrinkage estimator* $\tilde{\pi}_i$ will then be

$$\tilde{\pi}_i = \widehat{B}_i\hat{\mu} + (1 - \widehat{B}_i)\hat{\pi}_i. \tag{3}$$

$\widehat{B}_i$ is the factor deciding how big the shrinkage will be and will therefore be referred to as the *shrinkage factor*.

3 General imputation methods

There are many different imputation methods and not always easy to know which one of them to use or even if you can use imputation at all. One thing to be aware of when using imputation methods is that it can lead to bias or decrease of standard deviation as we shall see later.

In this section five different methods of imputation are going to be discussed. The first four of these are also going to be analyzed together with the Empirical Bayes and the EM method that is described in section 4.

There are two different ways of handling missing data. One way is to simply exclude the patients with missing outcomes from the analysis and another way is to impute data where they are missing.

3.1 Excluding non-complete data

When excluding patients from a clinical trial the sample size get smaller and that affect both the power and the variability of the test. A smaller sample size gives a greater possibility of a non-significant result i.e. the larger the sample size the greater the statistical power of the test. There is also possible that the patients with non-complete treatments have more extreme values than the others and therefore excluding them lessens the variability and the confidence interval.

Another effect to be aware of when dealing with missing values in a data material is bias and this is presumably also the most important concern [2]. Missing values will lead to bias if the unmeasured values are related to the real value (i.e. if a higher or lower value is more likely to be missing), but not if they are only related to the treatment (i.e. if one treatment arm is more likely to have

[2] emea (2001), p. 2

missing values than another).

The method used when you simply ignore missing data and go through with the statistical analysis with only complete data is called ”complete case analysis”. Often this is not a good way to deal with the data material because of the aspects pointed out above. It violates the intention to treat principle. Instead you probably should use some method of imputation.

3.1.1 The LOCF method

One method is last observation carried forward (LOCF) which is widely used [3]. Here the last measured observation before the missing one is imputed. This method works best if the observations is expected to remain at some level or if there is only a few missing values. If the observations in a test is expected to increase or decrease over time this method does not work very well. One example (figure 1) is a test where we look at how a medicine affects the condition of MS (Muscular Sclerosis) patients. Let us say that the observations of the patients muscle strength X_ij are graded at a scale from 1 to 10 where i is patient nr i and j is the test occasion. The natural development is that the muscle strength is slowly decreasing over time and the intention of the treatment would be to slow this development. If one patient drops out after say j = 10 and we use the LOCF method X_{ij} for j $>$ 10 will be put equal to $X_{i,10}$ and we do overestimate the missing observations. If there were only a few missing observations the LOCF method could be used, but some kind of regression model would presumably fit better (see chapter 3.1.4).

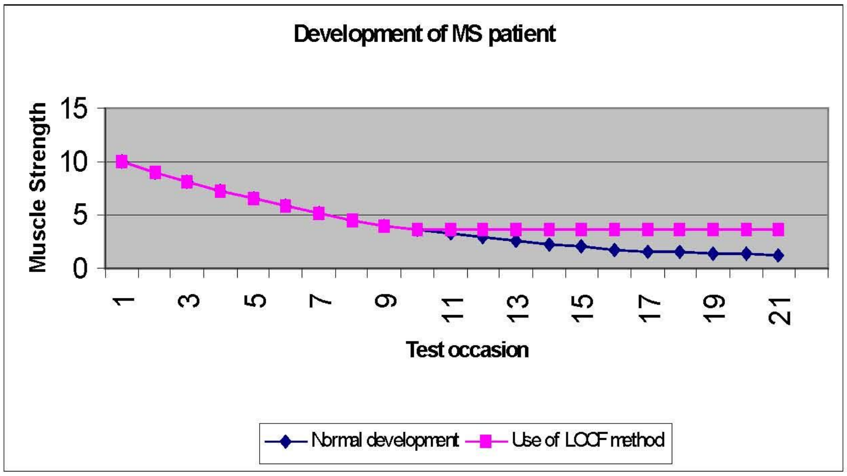


Figure 1: *We get an overestimation with the use of LOCF in the case with MS patients.*

[3] emea (2001), p. 3

In this case where we have a downward sloping trend the LOCF method is not conservative nor does it give a good estimate. If we have a upward sloping trend the LOCF method can be considered conservative. If there wouldn't have been a trend in the observations the method gives a fairly good estimated value depending on the last observed value. If the last observed value is an outlier the LOCF method does not give a good estimated value (though maybe conservative) but if it is somewhat 'normal' for the individual it is probably a better estimate.

3.1.2 Best or worst case imputation

Two other methods when dealing with missing data are best case and worst case imputation. Here, as the name tells, you impute the best or worst data. This leads to either an under or over evaluation of the data and can be used "to assess a lower bound of efficacy as a demonstration of robustness" [4]. There are different ways of using worst case imputation. I will give some examples of worst case imputation, where best case imputation works similar but but of course the other way around: If a patient, let's call him Adam, in a trial where there is no trend only has a few values recorded you can take the worst of these and impute it instead of the missing ones. But let's say that all of the observations actually are quite good compared to the other patients. Then the worst case would be if Adam gets the worst value of all of the patients observations. This method is of course conservative.

3.1.3 Mean value methods

A natural method of imputation is to use the mean value of the recorded observations. This method leads to lower variance and a concern here is that the dropouts might be more likely to be patients with more extreme values (i.e. a very ill patient might not show up). Another aspect of using the mean value is that it is not always clear on which data you should calculate the mean value. In the MS case you can not use the mean value for the whole period but if for example $X_{5,12}$ is missing you can take the mean value of all $X_{i,12}$ as an approximation, where i is patient number. Another idea would be to use the mean value of $X_{5,11}$ and $X_{5,13}$. This method can also be used for several missing values in a row, i.e. if j = 11, 12, 13 is missing you simply estimate the mean of $X_{5,10}$ and $X_{5,14}$. If we don't have a trend in the data we can randomly impute values from the knowledge of the patients mean value. We first calculate a mean value from the non-missing values and then use this mean value to produce new observations that we replace the missing ones with.

[4] Ibid

One problem with these methods is that they may lower the standard error because they estimate a central value and ignore its uncertainty. One way to deal with this is to use multiple imputation methods where you generate multiple copies of the data set and replace missing values by randomly generated values. Another way to avoid lowering of the standard error is to use maximum-likelihood methods where you fit a model by an iterative process [5].

Mean-value methods is made to give good estimates and is not conservative.

3.1.4 Regression methods

When trials are somehow linear as in the MS example linear regression methods can be used for imputation. One way is to use a simple regression model: $y_i = b_0 + b_1 x_i + e$ where x_i can be the time since the trial started.

If we take the example with MS patients again y_i would be the patients muscle strength and x_i would be the treatment time after i test occasions. If only the first 3 test occasions has been recorded we can predict the other by first decide b0, b1 and then calculate y_4, y_5 or y_6 by: $y_4 = b_0 + b_1 x_4 + e$.

Another way is to use a multiple regression model and thus the data from all patients. The model is $y_i = a + b_1 x_i + c * age + d * gender + e$ where age and gender is the age respectively the gender of the test person. Of course other variables as for example time since the outbreak of the disease is possible to add. With this model the gain is that we include other variables and we are then able to adjust the imputations better to the specific patient than if we use a simple regression model.

4 Empirical Bayes methods

4.1 Empirical Bayes theory

This subsection is partly based on theory from *Encyclopedia of Biostatistics [1998, p 1314]* and from the book by Bradley P. Carlin and Thomas A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis [2000]*. Generally Bayesian analysis uses past experience, guesses or convenient assumptions in the form of a prior distribution. Let us assume that we have a variable X with likelihood function p(x|θ) where θ is an unknown parameter that we wish to estimate. In classical procedures no care will be taken to the past data of θ. Here we will instead use the past data to get a better estimate of θ. We suppose that θ has a prior distribution $g(\theta|\eta)$ where η is a vector of so called *hyperparameters*. X is observed

[5] emea (2001), p. 3

why the estimate $\hat{\theta}$ are found in the posterior distribution. We wish to estimate θ with a small mean square error, thus

$$E[(\hat{\theta}-\theta)^2|x] = [\hat{\theta}(x) - E(\theta|x)]^2 + var(\theta|x)$$

and because var($\theta|x$) is not a function of $\hat{\theta}$, E($\theta|x$) is the estimator to be minimized. E($\theta|x$) is called the *Bayes* estimator. To evaluate it we first use Bayes formula to determine the posterior distribution[6]

$$p(\theta|x,\eta) = \frac{p(\mathbf{x}|\theta)g(\theta|\eta)}{m(\mathbf{x}|\eta)} \tag{4}$$

where m($\mathbf{x}|\eta$) is the marginal distribution of **x**,

$$m(\mathbf{x}|\eta) = \int p(\mathbf{x}|\theta)g(\theta|\eta)d\theta \tag{5}$$

whereby the posterior expectation is given by

$$E(\theta|x) = \frac{\int \theta p(\mathbf{x}|\theta)g(\theta|\eta)d\theta}{\int p(\mathbf{x}|\theta)g(\theta|\eta)d\theta}. \tag{6}$$

In the Bayesian approach equation (4) is used if η is known. If η is unknown (as it is in our Parkinson trial) we would in the Bayesian approach have to use a *hyperprior* distribution $h(\eta)$ and now we obtain the posterior distribution of θ by also marginalizing over η:

$$p(\theta|x) = \frac{\int p(\mathbf{x}|\theta)g(\theta|\eta)h(\eta)d\eta}{\int\int p(\mathbf{x}|\theta)g(\theta|\eta)h(\eta)d\theta d\eta}$$

This is often not trivial why we instead use the marginal distribution of X (5) to estimate the hyperparameter η. This can be done with marginal maximum likelihood estimation (MMLE) of $\hat{\eta}$. Inferences are then made on $p(\theta|\mathbf{x},\hat{\eta})$ by inserting $\hat{\eta}$ into equation (4). This kind of procedure is called *empirical Bayes* (EB) analysis.

When we obtain the posterior distribution it is convenient to use the shorthand

$$p(\theta|x,\eta) \propto p(\mathbf{x}|\theta)g(\theta|\eta) \tag{7}$$

which states that the posterior is proportional to the likelihood times the prior. This can be done because any constant or function of **y** can be multiplied with the likelihood without altering the posterior. When we choose a prior to our likelihood it can be convenient to choose a prior that is *conjugate* to the likelihood $p(\mathbf{x}|\theta)$. A conjugate is a prior that leads to a posterior distribution that belongs to the same distribution family as the prior.

[6] Gut (1995), p. 6

4.2 Assumptions

The data set we study in our trial is graded in an ordinal scale. We have $i = 1 \ldots N$ patients that are assumed to be independent of each other. Each patient have $t = 1 \ldots n_i$ observations that are also assumed to be independent. We assign $Z_{it} \in \{-3, \ldots, 3\}$ as variable for one observation. Z_{it} is either good $Y_{it} = 1$ if $Z_{it} \in \{-1, 0, 1\}$ or bad $Y_{it} = 0$ if $Z_{it} \in \{-3, -2, 2, 3\}$. The stochastic variable for one observation to be 'good' thus is Bernoulli-distributed, $\mathbf{Y}_{it} \overset{iid}{\sim} Be(\pi_i)$. For n_i observations we assign the stochastic variable $\mathbf{X}_i = \sum_{t=1}^{n_i} \mathbf{Y}_{it}$ which is binomial distributed, $\mathbf{X}_i|\pi_i \overset{ind}{\sim} Bin(n_i, \pi_i)$. The patients are independent from each other but because of different n_i and π_i, $\mathbf{X}_i|\pi_i$ is not identically distributed.

4.3 The beta/binomial model

A natural way to choose prior and likelihood function is to choose the $beta(r, s)$ distribution as prior (see Appendix A for more information on the beta distribution) and the binomial as likelihood. The reasons to why we choose the beta as prior are:

- because it stays in the interval [0,1].
- because it is a conjugate to the binomial distribution.
- because it can take different kinds of shape depending on our specific data set.

The reason to why we choose the binomial as likelihood is simply because the way the observations are drawn follows the definition of the binomial distribution. We start by giving the formula for the likelihood function:

$$P(\mathbf{X}_i = x_i|\pi_i) = \binom{n_i}{x_i} \pi_i^{x_i}(1-\pi_i)^{n_i - x_i}. \tag{8}$$

In the following formulas the subscript i will be left out for convenience. The expectation and variance for $\frac{\mathbf{X}}{n}|\pi$ is:

$$E\left(\frac{X}{n}|\pi\right) = \frac{E(X|\pi)}{n} = \frac{n\pi}{n} = \pi$$

,

$$Var\left(\frac{X}{n}|\pi\right) = \frac{Var(X|\pi)}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \pi(1-\pi)\frac{1}{n}.$$

To make the beta distribution more easy to work with we reparametrize $beta(r, s)$ to $beta_{rep}(\mu, M)$. We put $\mu = r/(r+s)$ which is the overall mean of all patients and $M = r+s$ is a factor affecting the variance. Increasing M decreases the variance. We assume that $\pi|\mu \overset{iid}{\sim} beta_{rep}(\mu, M)$-distributed and the density function is:

$$g(\mathbf{\Pi} = \pi|\mu, M) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\pi^{M\mu-1}(1-\pi)^{M(1-\mu)-1} \tag{9}$$

where μ and M are the hyperparameters. The expectation and variance for $beta_{rep}(\mu, M)$ is:

$$E(\pi|\mu, M) = \mu, \;\; Var(\pi|\mu, M) = \frac{\mu(1-\mu)}{M+1}.$$

We get a beta/binomial distribution for marginal density function of $\mathbf{X}$ [7]:

$$\begin{aligned} m(\mathbf{X} = x|\mu, M) = \int_0^1 p(\mathbf{X}|\pi)g(\mathbf{\Pi}|\mu, M)d\pi = \\ \binom{n}{x}\frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\int_0^1 \pi^{x+M\mu-1}(1-\pi)^{n-x+M(1-\mu)-1}d\pi = \\ \binom{n}{x}\frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\frac{\Gamma(x+M\mu)\Gamma(n-x+M(1-\mu))}{\Gamma(n+M)}. \end{aligned} \tag{10}$$

The expectation and variance for $\frac{X}{n}$ is:

$$E\left(\frac{X}{n}\right) = E\left[E\left(\frac{X}{n}|\pi\right)\right] = E(\pi) = \mu,$$

$$Var\left(\frac{X}{n}\right) = E\left[Var\left(\frac{X}{n}|\pi\right)\right] + Var\left[E\left(\frac{X}{n}|\pi\right)\right] = E[\pi(1-\pi)\frac{1}{n}] + Var(\pi) =$$

$$\frac{1}{n}(\mu(1-\mu) - Var(\pi)) + Var(\pi) = \frac{\mu(1-\mu)}{n}\left(1 + \frac{n-1}{M+1}\right). \tag{11}$$

The posterior distribution is

$$\begin{aligned} P(\mathbf{\Pi}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{\Pi})P(\mathbf{\Pi}) = \\ \binom{n}{x}\frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\pi^{x+M\mu-1}(1-\pi)^{n-x+M(1-\mu)-1} \end{aligned} \tag{12}$$

which is $beta(r_{EB}, s_{EB})$-distributed, with $r_{EB} = x + M\mu$ and $s_{EB} = n - x + M(1-\mu)$. Notice that this is not the reparametrized beta distribution. We see that both the prior and the posterior is beta-distributed and we can conclude that the beta distribution is the conjugate for the binomial likelihood.

[7] Gut (1995), p. 19

4.3.1 Moment Estimates

For each patient i, π_i is estimated by

$$\hat{\pi}_i = \frac{x_i}{n_i}. \tag{13}$$

A weighted moment estimate for μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} n_i \hat{\pi}_i}{\sum_{i=1}^{N} n_i} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} n_i}. \tag{14}$$

To find a moment estimate for M we use equation (11):

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} Var(\frac{x_i}{n_i}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\hat{\mu}(1-\hat{\mu})}{n_i}\left[1 + \frac{n_i - 1}{\widehat{M}+1}\right] \tag{15}$$

where a weighted estimate for s^2 is

$$s^2 = \frac{N\sum n_i(\hat{\pi}_i - \hat{\mu})^2}{(N-1)\sum n_i}. \tag{16}$$

We solve the equation for M and get:

$$\widehat{M} = \frac{\hat{\mu}(1-\hat{\mu}) - s^2}{s^2 - \frac{\hat{\mu}(1-\hat{\mu})}{N}\sum_{i=1}^{N}\frac{1}{n_i}}. \tag{17}$$

The estimated posterior expectation

$$\tilde{\pi}_i = E(\mathbf{\Pi}|\hat{\mu}, \widehat{M}) = \frac{r_{EB}}{r_{EB} + s_{EB}} = \frac{\frac{x_i}{n_i} + \widehat{M}\hat{\mu}}{n_i + \widehat{M}} = \frac{\widehat{M}}{n_i + \widehat{M}}\hat{\mu} + \frac{n_i}{n_i + \widehat{M}}\frac{x_i}{n_i} \tag{18}$$

and the estimated posterior variance

$$\begin{aligned}\tilde{\sigma}_i^2 = Var(\mathbf{\Pi}|\hat{\mu}, \widehat{M}) = \frac{r_{EB}s_{EB}}{(r_{EB} + s_{EB})^2(r_{EB} + s_{EB} + 1)} = \\ \frac{(x_i + \widehat{M}\hat{\mu})(n_i - x_i + \widehat{M}(1-\hat{\mu}))}{(n_i + \widehat{M})^2(n_i + \widehat{M} + 1)} = \\ \frac{\tilde{\pi}_i}{n_i + \widehat{M}}\frac{n_i - x_i + \widehat{M}(1-\hat{\mu})}{n_i + \widehat{M} + 1} = \frac{\tilde{\pi}_i(1-\tilde{\pi}_i)}{n_i + \widehat{M} + 1}.\end{aligned} \tag{19}$$

We see here that $\tilde{\pi}_i$ is a weighted average between the prior mean $\hat{\mu}$ and $\tilde{\pi}_i$, depending on the weight of $\widehat{M}$ (information in the prior) relative to n_i (information in the data). To visualize the weighting more and to make it easier to compare with other methods we can write the posterior expectation as

$$\tilde{\pi}_i = \widehat{B}_i\hat{\mu} + (1 - \widehat{B}_i)\hat{\pi}_i. \tag{20}$$

where $\widehat{B}_i = \frac{\widehat{M}}{\widehat{M}+n_i}$ is the so called *shrinkage factor*. If we recall formula (3) from chapter 2 it is observed that this formula is similar. Though $\widehat{B}_i$ is a function of $\hat{\mu}$ (see equation (17) and (18)) why equation (20) is not a linear function as equation (3) appears to be. Clearly $\widehat{B}_i$ depends on the relationship between M and n_i. The higher M is and the smaller n_i is the bigger is the shrinkage or vice versa. If n_i = 0, B_i = 1 and if n_i is large B_i is small.

As $M \rightarrow \infty$ the prior variance $Var(\pi_i|\mu, M) \rightarrow 0$ and we have exact knowledge of π_i (all π_i are the same). Also $B_i \rightarrow 1$ and $\tilde{\pi}_i \rightarrow \hat{\mu}$ for all i. If equation (17) is observed it can be concluded that for M to turn to infinity, s^2 has to turn to the value $\frac{\hat{\mu}(1-\hat{\mu})}{N}\sum_{i=1}^{N}\frac{1}{n_i}$.

As $M \rightarrow 0$ the prior variance $Var(\pi_i|\mu, M)$ turns to its maximum value $\mu(1-\mu)$, $B_i \rightarrow 0$ and $\tilde{\pi}_i \rightarrow \hat{\pi}_i$. From equation (17) we conclude that s^2 has to turn to its maximum value $\hat{\mu}(1-\hat{\mu})$ for M to turn to zero.

M is a parameter that is the same for the whole dataset, while n_i can be different for each patient, so differences between $\widehat{B}_i$ within the same dataset depends solely on n_i. If a patient has zero observations, $n_i = 0$, $B_i = 1$ and $\hat{\pi}_i$ is equal to the population mean $\hat{\mu}$.

4.3.2 Maximum likelihood estimation

A better way to find $\hat{\mu}$ and $\hat{M}$ is with maximum likelihood estimates. The derivation of the ML-estimates are as follows:

The likelihood function is

$$L(\mu, M) = \prod_{i=1}^{N}\left(\frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))}\binom{n_i}{x_i}\frac{\Gamma(x_i+M\mu)\Gamma(n_i-x_i+M(1-\mu))}{\Gamma(n_i+M)}\right).$$

The loglikelihood is

$$l(\mu, M) = log(n_i) + log\Gamma(M) - log[\Gamma(M\mu)\Gamma(M(1-\mu))] +$$
$$+\sum[log\binom{n_i}{x_i} + log\Gamma(x_i+M\mu) + log\Gamma(n_i-x_i+M(1-\mu)) - log\Gamma(n_i+M)].$$

The derivative of the likelihood function with respect to μ and M are

$$\frac{dl(\mu, M)}{d\mu} = -\frac{\Psi(M\mu)M}{\Gamma(M\mu)} + \frac{M\Psi(M(1-\mu))}{\Gamma(M(1-\mu))} +$$
$$+\sum_{i=1}^{N}\left[\frac{M\Psi(x_i+M\mu)}{\Gamma(x_i+M\mu)} - \frac{M\Psi(n_i-x_i+M(1-\mu))}{\Gamma(n_i-x_i+M(1-\mu))}\right] = 0, \qquad (21)$$

$$\frac{dl(\mu, M)}{dM} = \frac{\Psi(M)}{\Gamma(M)} - \frac{\mu\Psi(M\mu)}{\Gamma(M\mu)} - \frac{(1-\mu)\Psi(M(1-\mu))}{\Gamma(M(1-\mu))} +$$

$$-\sum_{i=1}^{N}\left[\frac{\Psi(n_i + M)}{\Gamma(n_i + M)} + \frac{\mu\Psi(x_i + M\mu)}{\Gamma(x_i + M\mu)} + \frac{(1-\mu)\Psi(n_i - x_i + M(1-\mu))}{\Gamma(n_i - x_i + M(1-\mu))}\right] = 0. \quad (22)$$

The greek letter Ψ is used as the derivative of the Γ-function. A problem in equation (21) and (22) is that it is not possible to express one of the variables as a function of the other. I have used a so called *hybrid* method to solve these nonlinear equations. The moment estimate for μ has been inserted into equation (21) and (22) to find the MLE of M and vice versa. Because of the usage of both the moment estimate and the MLE the method is called a hybrid method. The result is shown in figure 2, 3, 4 and 5 below. In figure 2 we see that $l(\mu, M)$ has a maximum for $M \approx 8$ and a minimum when $M \to \infty$. $M \approx 8$ is not a

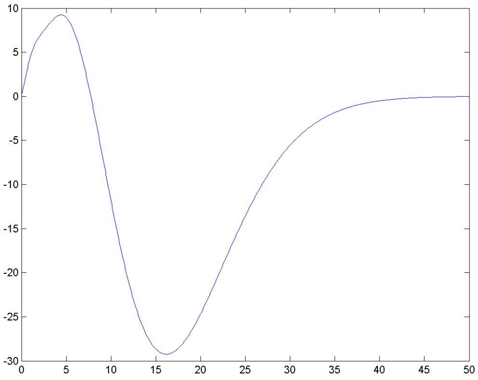


Figure 2: *The derivative of $l(\mu, M)$ using formula 21 with $\hat{\mu}$ from equation 14.*

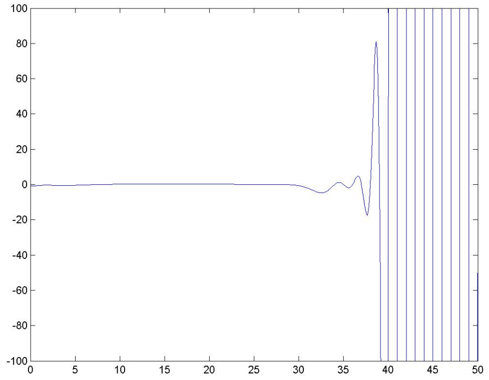


Figure 3: *The derivative of $l(\mu, M)$ using formula 22 with $\hat{\mu}$ from equation 14.*

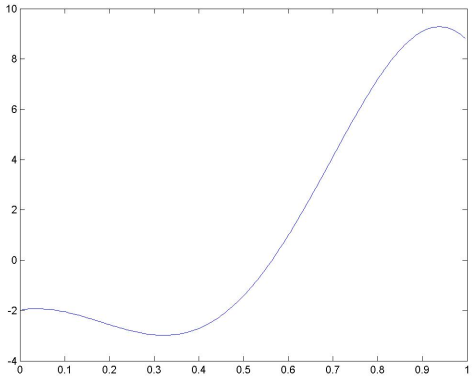


Figure 4: *The derivative of $l(\mu, M)$ using formula 21 with $\widehat{M}$ from equation 17.*

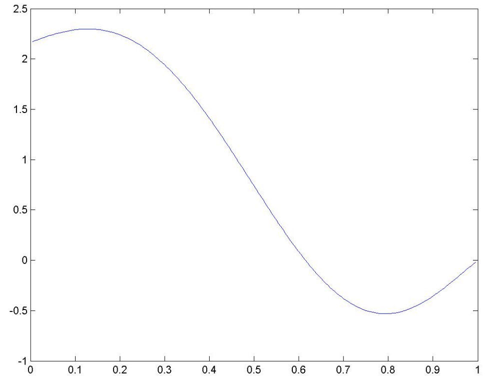


Figure 5: *The derivative of $l(\mu, M)$ using formula 22 with $\widehat{M}$ from equation 17.*

good estimate of M because if we construct samples with M equal to 1 or 50 the ML estimate is still somewhere between 6 and 15. In figure 3 we can not come to any single estimate for M because $l(\mu, M)$ has many maximums and minimums throughout the x-axis. The conclusion is that we cannot come to any single estimate for M from any of these figures or forth worth with the ML method. In figure 4 $l(\mu, M)$ has a maximum value at $\mu \approx 0.2$ and a minimum at $\mu \approx 0.8$. In figure 5 $l(\mu, M)$ has a maximum at $\mu \approx 0.8$ which is consistent with the moment estimate of μ that is 0.83. Unfortunately this is the only estimate that is somewhat consistent with the moment estimates why the moment method will be used instead of the hybrid method.

4.4 The Gaussian/Gaussian model

Another way to analyze the data set is to use normal approximation and then get a Gaussian/Gaussian (normal/normal) model instead of the beta/binomial model we used above. The likelihood function is binomial distributed why the expectation and variance for $\frac{X_i}{n_i}|\pi_i$ is

$$E\left(\frac{X_i}{n}|\pi_i\right) = \frac{E(X_i|\pi_i)}{n_i} = \frac{n_i\pi_i}{n_i} = \pi_i,$$

$$Var\left(\frac{X_i}{n}|\pi_i\right) = \frac{Var(X_i|\pi_i)}{n_i^2} = \frac{n_i\pi_i(1-\pi_i)}{n_i^2} = \frac{\pi_i(1-\pi_i)}{n_i}.$$

We approximate the binomial distribution with a normal (or Gaussian) distribution, thus $\frac{X_i}{n_i}|\pi_i \overset{ind}{\sim} N(\pi_i, \frac{\pi_i(1-\pi_i)}{n_i})$.

We use a normal distribution as a prior instead of the beta distribution why $\pi_i|\mu$ now is assumed to be i.i.d. $N(\mu, \tau^2)$-distributed, with hyperparameters μ and τ^2. It can be questioned if a normal distribution is reasonable to use here because π_i is restricted to the interval [0,1]. We will get the answer on that question when we do the tests.

We get a Gaussian/Gaussian distribution for marginal density function of $\mathbf{X}$ [8]:

$$m(x|\mu,\tau^2) = \int p(x|\theta,\sigma^2)g(\theta|\mu,\tau^2)d\theta \sim N(\mu, \sigma^2+\tau^2) \qquad (23)$$

According to Bayes formula (4) the posterior distribution is[9]:

$$P(\pi_i|x_i) \propto \frac{1}{\sqrt{2\pi\sigma_i^2}}\,exp\left\{-\frac{1}{2}(x_i-\pi_i)^2/\sigma_i^2\right\}\frac{1}{\sqrt{2\pi\tau^2}}\,exp\left\{-\frac{1}{2}(\pi_i-\mu)^2/\tau^2\right\} =$$

[8] Carlin, Louis, (2000) p. 11, 62

[9] Ibid p. 63

$$\frac{1}{2\pi\sigma_i\tau}\,exp\left\{-\frac{1}{2}\frac{\tau^2(x_i-\pi_i)^2+\sigma_i^2(\pi_i-\mu)^2}{\sigma_i^2\tau^2}\right\}=$$
$$\frac{1}{2\pi\sigma_i\tau}\,exp\left\{-\frac{1}{2}\frac{\pi_i^2(\sigma_i^2+\tau^2)-2\pi_i(\tau^2x+\sigma_i^2\mu)+\tau^2x_i^2+\sigma_i^2\mu^2}{\sigma_i^2\tau^2}\right\}$$

which is $N(B_i\mu+(1-B_i)x_i,(1-B_i)\sigma_i^2)$-distributed, where $B_i=\frac{\sigma_i^2}{\sigma_i^2+\tau^2}$.

4.4.1 Estimates

Before we continue with the EB estimate we have to find estimates for the expectation and variance of $\frac{X_i}{n_i}|\pi_i$ and $\pi_i|\mu$. A moment estimate for μ is as before equation (14) and a weighted variance estimate for the whole populace variance τ^2 is

$$\hat{\tau}^2=\frac{N\sum n_i(\hat{\pi}_i-\hat{\mu})^2}{(N-1)\sum n_i}. \tag{24}$$

For $\frac{X_i}{n_i}|\pi_i$ a moment estimate for $\hat{\pi}_i$ is equation (1) and a moment estimate for the variance, $\hat{\sigma}_i$, is

$$\hat{\sigma}_i^2=\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}$$

because $\frac{X_i}{n_i}|\pi_i$ is binomial distributed. The estimated posterior expectation of π_i then is,

$$\tilde{\pi}_i=\widehat{B}_i\hat{\mu}+(1-\widehat{B}_i)\frac{x_i}{n_i} \tag{25}$$

and the estimated posterior variance

$$Var(\pi_i|\hat{\mu})=(1-\widehat{B}_i)\hat{\sigma}_i^2. \tag{26}$$

The estimate $\widehat{B}_i$ is simply found by

$$\widehat{B}_i=\frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2+\hat{\tau}^2}. \tag{27}$$

Equation (25) follows similar to equation (20) the same principle as the formula mentioned in chapter 2 equation (3). Though $\widehat{B}_i$ is a function of $\hat{\pi}_i$ ($\hat{\sigma}_i^2=\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}$) why equation (27) is not a linear function as equation (3) appears to be. Here $\tilde{\pi}_i$ is a weighted average between the prior mean $\hat{\mu}$ and $\hat{\pi}_i$, depending on the weight of $\hat{\tau}^2$ (information in the prior) relative to $\hat{\sigma}_i^2$ (information in the data). If any of σ^2 or τ^2 would have been known we would have found a better estimate by using the marginal likelihood function to get a marginal MLE as we did in subsection (4.3.2). Now that both are unknown it is unfortunately no idea to do this which is understood after some calculations which I consider not necessary to bring up here. We will simply have to be satisfied with the estimate of $\hat{B}_i$ as it is above in equation (27).

As $\hat{\tau}^2 \to 0$ we have exact knowledge of π_i, $B_i \to 1$ and $\tilde{\pi}_i \to \hat{\mu}$ for all i(same as for the beta/binomial case when $M \to \infty$).

As $\hat{\tau}^2 \to \infty$ the prior gives no information, so $B_i \to 0$ and $\tilde{\pi}_i \to \hat{\pi}_i$ for all i. This can be compared with the beta/binomial case when $M \to 0$ with the difference that then the prior variance turns to $\hat{\mu}(1-\hat{\mu})$ instead of infinity.

A large $\hat{\sigma}_i^2$ comes from a low number of observations n_i (little information from the data set) and results in bigger shrinkage. $\hat{\tau}^2$ is the same for the whole data set while $\hat{\sigma}_i^2$ is different for each patient. $\hat{\sigma}_i^2$ depends on both $\hat{\pi}_i$ and n_i why these parameters affect the difference between single patients $\widehat{B}_i$. We can compare this with the beta/binomial model where it was only n_i that affected the difference in shrinkage factor between patients. If a patient has zero observations, $n_i = 0$, it is easy to see that $B_i = 1$ and that $\tilde{\pi}_i$ is equal to the population mean $\hat{\mu}$.

In both the EB models we do not really impute values instead of the missing ones. What we did was to estimate a new $\tilde{\pi}_i$ for each patient even if the patient in question did not have any missing values. We did come to the conclusion that patients with a smaller number of observations get a higher shrinkage, which is good indeed, but if we want a patient with no missing values to have approximately no shrinkage at all the number of observations would have to be very large (possibly over 1000), i.e. the individual variance would have to tend to zero. If we don't have that many observations it is not possible to get zero variance or no shrinkage. The estimate $\tilde{\pi}_i$ is an estimate of the true π_i, with missing values or not, but of course if the variance tends to zero we do get a very close estimate of π_i and the shrinkage is very small. It is mostly not possible to accomplish very large number of observations per patient (n_i) when the trial is made. Hopefully Empirical Bayes method, as described in the two subsections above, will give us good estimates of the true vector π.

4.5 The EM-algorithm

If we want to impute values directly and not change the $\hat{\pi}_i$ for the patients that do not have any missing values a method that follows the *Expectation Maximization* algorithm (EM-algorithm) can be used. A short version on how the

EM-algorithm works is given.

Consider a model where a vector $Y = T(X)$ is observed, with prior $g(Y|\theta)$, of the complete model $X = (X_1, \ldots, X_n)$, with distribution function $f(X|\theta)$ where the MLE of θ is to be found. Now a score function can be assigned:

$$S(Y|\theta) = \frac{\delta}{\delta\theta} ln(g(Y|\theta)). \tag{28}$$

Then we go on with the 'E'- or expectation-step which is to calculate

$$\overline{S}(\theta|\theta^{(k)}) = E(S(Y|\theta)|\mathbf{y}, \theta^{(k)}). \tag{29}$$

$\theta^{(k)}$ is of course the value of θ at iteration j where $j = 0, 1, 2, \ldots$. To manage this the Bayesian techniques described in the previous sections are used. With the 'M'- or maximization-step a new estimate of the parameter θ is then found:

$$\theta^{(k+1)} = argmax\overline{S}(\theta|\theta^{(k)}). \tag{30}$$

We then start from the beginning again with the 'E'-step and repeat the recursion until $\theta^{(k)}$ converges to some value $\hat{\theta}^{EM}$. This version of the EM-algorithm has its background in the works of Carlin, B.P. and Louis, T.A. (2000) and Zwanzig, S. (2003).

The way the 'E'- and 'M'-step are used in this essay is as follows. For the 'E'-step the analysis of Empirical Bayes described in section 4.3 and 4.4 will be used to find $\overline{S}$ which is our $\tilde{\pi}_i$ for $i = 1, \ldots, N$. Then for the 'M'-step new values will be generated from the Be($\tilde{\pi}_i$)-distribution which will be imputed where the missing values was. Then a new estimate for $\hat{\pi}_i$ is calculated which is called $\hat{\pi}_i^{EM}$, i.e. the recursion is made once.

5 Data analysis

5.1 Analysis of trial data set

How the trial data is built was described in section 2 and 4.2. The trial consisted of 22 patients with 136 video observations each if no missing values are recorded. Half of of the observations were done when the patients had received drug A and half when they received drug B. Comparing differences between drugs is not the topic of this essay why the analysis will be made on all of the observations (i.e. there will be 136 observations per patient if the patient have no missing values). The analysis was made in *Matlab*.

If the histogram of the data set in figure (6) is compared with the plots of the normal distribution in figure (7) it is seen that the normal/normal model

method	simple est	EB (p/p)	EM (p/p)	LOCF	best c	worst c	mean
$\hat{\mu}$ b/b	0.831	0.832	0.803	0.786	0.847	0.755	0.801
$\hat{\mu}$ n/n		0.834	0.797				
$\widehat{M}$ b/b	2.83	2.94/117	2.94/2.47	2.46	4.88	1.33	2.38
$\hat{\tau}^2$ b/b	0.037	0.036/0.0012	0.036/0.047	0.049	0.022	0.080	0.047
$\hat{\tau}^2$ n/n		0.037/0.034	0.037/0.050				

Table 1: *Results from analysis. p/p stands for prior/posterior values, b/b for beta/binomial method and G/G for Gaussian/Gaussian method. The blanks are of course used instead of writing the value above again.*

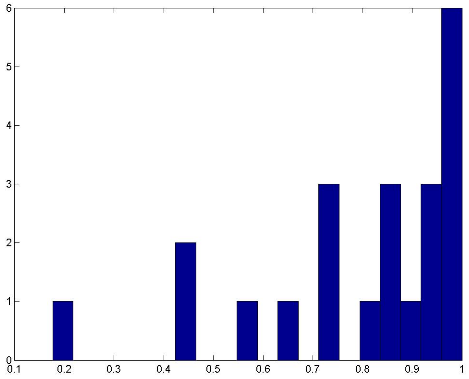


Figure 6: *Histogram of the trial data set.*

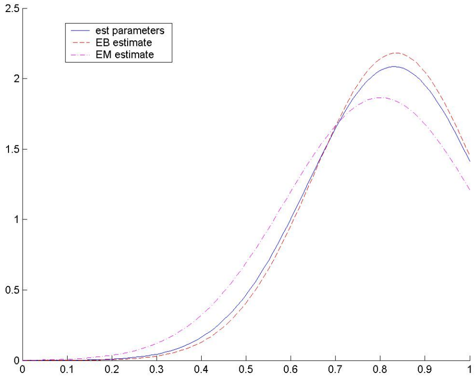


Figure 7: *Graph of the $N(\hat{\mu}, \hat{\tau}^2)$-distribution for the simple-, EB- and EM-estimate in the normal/normal model.*

gives a very poor fit to the data. This is somewhat as expected because it is not really appropriate to use the normal distribution in this case where it is quite obvious that the beta is the superior distribution to use.

The trial data had few missing values, it is less than five percent for most patients why it is necessary to generate data sets so that the amount of missing values and other variables can be chosen. This is done in the next section.

5.2 Analysis of simulations

To evaluate the EB-model and EM-model new data sets with different parameter values for the underlying 'true' values of π and M will be generated. The way it is done are as follows:

- Generate N observations $\pi_1, \ldots, \pi_N$ from the $beta(\mu, M)$-distribution. μ and M are the chosen 'true' parameters and π_i are the 'true' values that later is to be estimated with our different imputation methods.

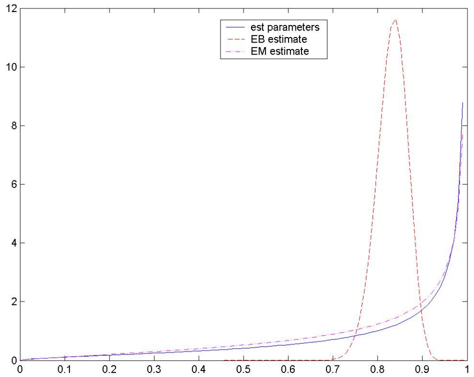


Figure 8: *Graph of the beta$(\hat{\mu}, \widehat{M})$-distribution for the simple-, EB- and EM-estimate in the beta/binomial model.*

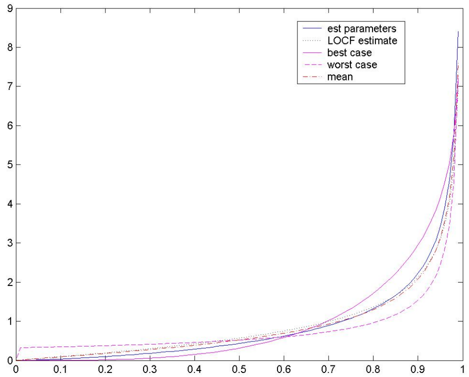


Figure 9: *Graph of the beta$(\hat{\mu}, \widehat{M})$-distribution for the estimates of LOCF-, bestcase-, worst case- and mean method-models.*

- Generate n observations N times from the Bernoulli distribution. That is for every π_i n observations are drawn from $Be(\pi_i)$. Thus an $N \times n$ matrix has been created.

- Now randomly remove observations for each row of the matrix in the fashion that each row get a different expected number of missing values. This is done by generating values p_i from the beta(r,s)-distribution where $\frac{r}{r+s}$ is the expected percentage of missing values for the whole data set. For each row, $i = 1, \ldots, N$, p_i is the probability that each observation is a missing value.

5.2.1 A small simulation

To get a visualization of how the Empirical Bayes- and the EM-method works a simulation with a small N is made and showed in table 2. The 'true' parameters of the beta distribution is $\mu = 0.65$ and $M = 10$, and the expected amount of missing values is 25%. For the Empirical Bayes method it can be seen from this table that the higher the difference is between $\hat{\pi}_i$ and the mean value the larger is the shrinkage, while the number of missing values does not affect the shrinkage as much. For the EM-method the number of missing values affect the shrinkage most, while the difference between $\hat{\pi}_i$ and mean does not have that big influence. But of course for the shrinkage to be big, as in row 13, both missing values and difference $(\hat{\pi}_i - \hat{\mu})$ has to be quite large. Generally the shrinkage is bigger for the EB-estimate than the EM-estimate, unless the amount of missing values are large. As an example of this row number 4 and 11 can be studied, where the amount of missing values are about 60% and the

i	n_i	B_i	true π_i $beta_{rep}(0.65, 10)$	$\hat{\pi}_i$	$\hat{\pi}^{EB}$	(shr EB)	$\hat{\pi}^{EM}$	(shr EM)
$\hat{\mu}$			0.670	0.639	0.637		0.645	
$\widehat{M}\ prior$			7.014	5.854	9.33		9.33	
$\widehat{M}\ post.$			7.014	5.854	31.11		12.33	
1	28	0.25	0.5667	0.6429	0.6418	-0.0011	0.6333	-0.0095
2	16	0.37	0.6767	0.5625	0.5905	0.0280	0.6333	0.0708
3	26	0.26	0.4607	0.4231	0.4800	0.0569	0.4000	-0.0231
4	9	0.51	0.4747	0.6667	0.6523	-0.0143	0.6000	-0.0667
5	16	0.37	0.6496	0.7500	0.7089	-0.0411	0.7667	0.0167
6	26	0.26	0.8916	0.8846	0.8196	-0.0650	0.8333	-0.0513
7	29	0.24	0.5318	0.3103	0.3902	0.0799	0.3333	0.0230
8	21	0.31	0.7974	0.7143	0.6910	-0.0233	0.7333	0.0190
9	28	0.25	0.5451	0.6429	0.6418	-0.0011	0.6667	0.0238
10	29	0.24	0.6470	0.5517	0.5728	0.0211	0.5667	0.0149
11	13	0.42	0.8021	0.7692	0.7146	-0.0546	0.6333	-0.1359
12	27	0.26	0.6247	0.6296	0.6319	0.0023	0.6667	0.0370
13	9	0.51	0.5030	0.2222	0.4341	0.2119	0.4333	0.2111
14	30	0.24	0.6747	0.7000	0.6854	-0.0146	0.7000	0
15	15	0.38	0.7963	0.8667	0.7792	-0.0875	0.7667	-0.1000
16	27	0.26	0.8929	0.8519	0.7971	-0.0548	0.8667	0.0148
17	20	0.32	0.7963	0.8500	0.7827	-0.0673	0.8000	-0.0500
18	14	0.40	0.4999	0.5000	0.5554	0.0554	0.5667	0.0667
19	15	0.38	0.9075	0.8667	0.7792	-0.0875	0.8667	0
20	28	0.25	0.5702	0.4286	0.4810	0.0525	0.4333	0.0048

Table 2: *Shrinkage (shr), estimate (EB or EM) - $\hat{\pi}_i$, is within parenthesis, with the usage of the EB- and the EM-method.*

EM-shrinkage higher than the EB-shinkage.

To see how the Gaussian/Gaussian model works in comparison with the beta/binomial another simulation was made and the table to that is shown in appendix A.2 table 3. From this table where the true $M = 20$ we can see that the estimated prior variance is much higher in the Gaussian/Gaussian model and that results in much lesser shrinkage than in the beta/binomial model. The reason for the high variance is probably the poor fit of the model as stated in section 5.1. For other M within the interval [1,100] the shrinkage still is much lesser. The only case when the estimates are quite similar is when the number of observations per patient is high (at least above 100). In figure 19 and 20 the difference between the 'true' estimate of pi_i and beta/binomial respectively Gaussian/Gaussian estimates are shown in box-plots. There we see that the beta/binomial model gives better results than the Gaussian/Gaussian, especially when the variance is high (figure 19, M=40). The following tests will be made only with the beta/binomial model.

5.2.2 Large data set simulation

The observations made in the subsection above are consistent with the expectations in section 4.3 and 4.4. In that small sample it is not possible to conclude if the EB- or EM-method is efficient or better than the general methods of imputation, why bigger samples with different parameter values and different amounts of missing values has to be made. That is done in this subsection.

Box-plots of differences between the 'true' model and the different general methods and Bayesian methods can be used to show the efficiency of the imputation methods. $\mu = 0.3$ and $N = 1000$ for all figures. What the box-plots different boxes shows are the difference between the 'true' π_i and:

1. Non-Bayesian estimate.
2. Empirical Bayes estimate, that is with $\hat{\mu}$ and $\widehat{M}$.
3. Bayes estimate, that is with the 'true' μ and M.
4. The EM method with $\hat{\mu}$ and $\widehat{M}$.
5. The EM method with 'true' μ and M.
6. LOCF method.
7. Best case method.
8. Worst case method.
9. Mean method.

Figure 10 shows that the EB-method gives slightly better estimates of π_i than the other methods and when the variance is lowered as in figure 11 (M=40 instead of 5) it can be seen that it is clearly better. For this low amount of missing values there is almost no difference between the EM-method and the non-Bayesian estimate.

For lower number of observations the EB-method gives significantly better estimates than the other methods as we see in figure 12 where n=10. When the number of missing values are increased to 30% as in figure 13 it can be noted that the EM-method begin to show better estimates than the general imputation methods.

The amount of missing values are raised to 50% in figure 14 and 15. We

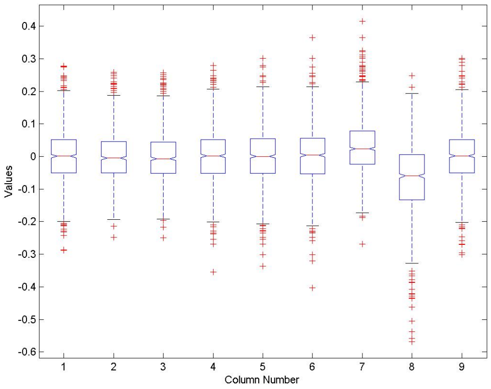


Figure 10: *Missing values 10%, M=5 and n=30.*

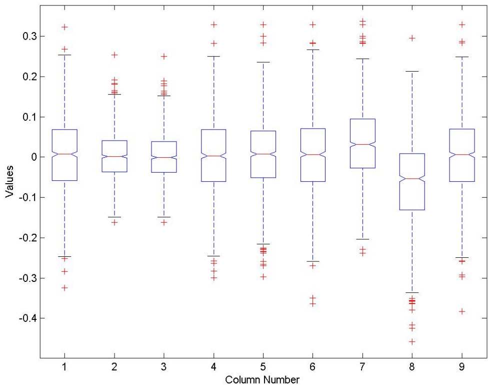


Figure 11: *Missing values 10%, M=40 and n=30.*

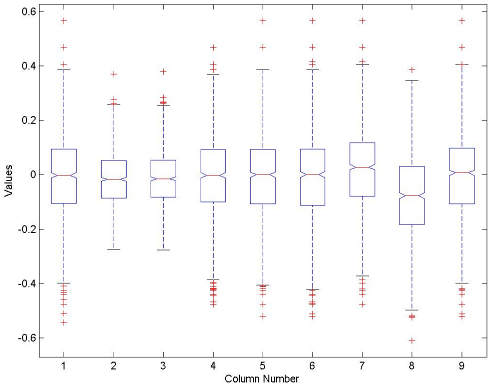


Figure 12: *Missing values 10%, M=10 and n=10.*

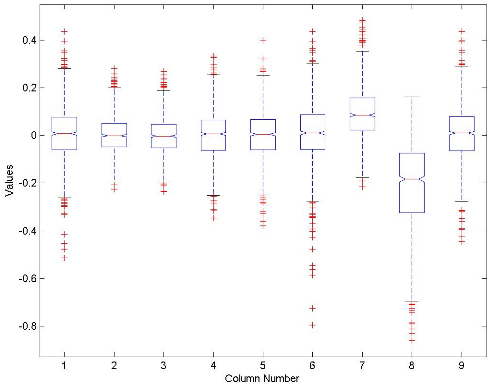


Figure 13: *Missing values 30%, M=15 and n=30.*

can spot a slight difference between the EB-estimate and the Bayes estimate when the variance is high (figure 15, M=5), but for lower variance (figure 14, M=40) the difference is small. It can also be stated that the EM-method now has significantly better estimates than the regular methods and the difference between the EB- and EM-method is smaller.

For very large numbers of observations as in figure 16 where n=500 it is hard to spot any differences between methods other than that EB and EM gives a lower amount of outliers. For very large number of missing values (80%) and high variance (M=3) as in figure 17 the EB and EM method has very similar estimates. From the figures above we can also draw the conclusions that the estimates get better as M increases (the prior variance decreases).

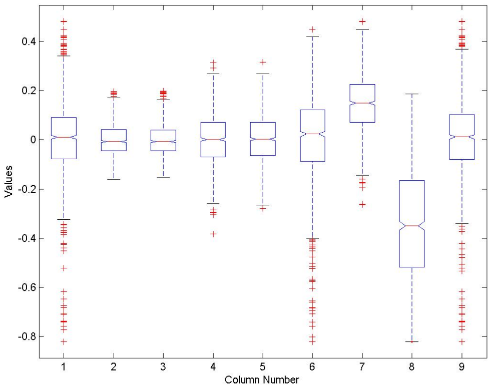


Figure 14: *Missing values 50%, M=40 and n=30.*

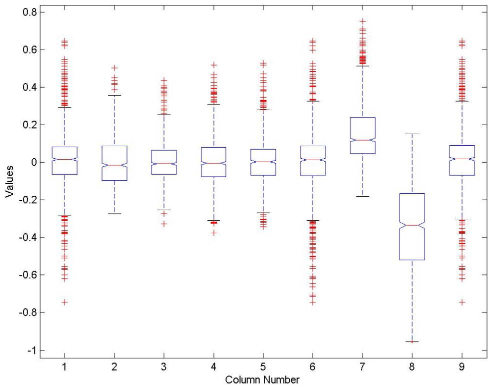


Figure 15: *Missing values 50%, M=5 and n=30.*

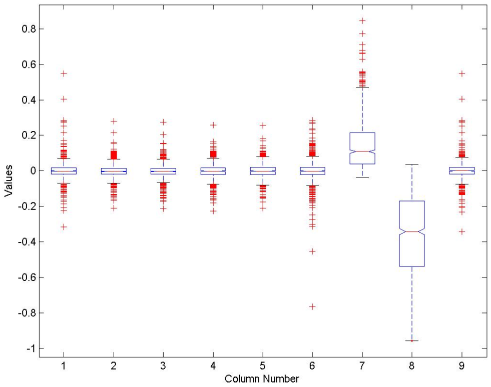


Figure 16: *Missing values 50%, M=5 and n=500.*

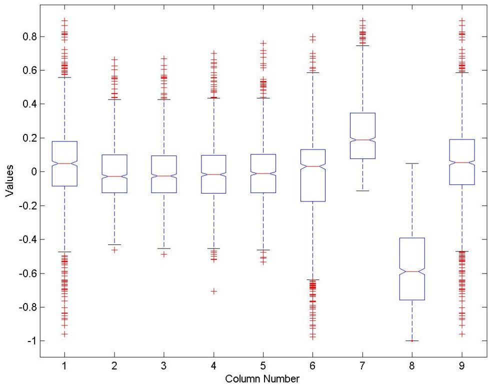


Figure 17: *Missing values 80%, M=3 and n=20.*

6 Discussion

With the Empirical Bayes method we get new estimates of π_i for each patient $i = 1, \ldots, N$, where π_i is the probability that the patient i is in a healthy state. The Empirical Bayes method may not really be seen as an imputation method where new observations is created instead of the missing ones, but the difference in *shrinkage* between patients are dependent on the number of missing values (number of observations n_i). If it is completely necessary to impute new observations the Expectation Maximization method can be used instead of the EB method but the shrinkage factor will be smaller depending on the amount of missing values. When the amount is about 5-10% the shrinkage factor is small relative to the EB method, but when it increases to about 50% the difference is less and the new estimates of pi_i is approximately as good compared to the general methods as the EB method is.

The Expectation Maximization method as it is implemented in this essay is used in such a way that the recursion is made once. It can possibly be made until convergence is reached but I did not think it a necessity to do so because the values assessed by doing the recursion once was satisfying.

The data over Parkinson patients is on the scale [-3,3] and were divided into two health states. These were the 'good' interval [-1,0,1] and the 'bad' interval [-3,-2,2,3]. As described in section 2 a value below zero states that the patient suffer from stiffness and above zero that the patient suffer from shaking. It could then be interesting to divide the scale into three health states and these would be [-3,-2], [-1,0,1] and [2,3]. The reason for doing this would be that the tested drug may be of more help against stiffness than shaking or vice versa and therefore it is interesting to use three classifications. If we want to do this we would use a multinomial distribution as likelihood together with a Dirichlet distribution for prior. The Dirichlet is conjugate to the multinomial why the calculations are fairly easy.

7 Conclusions

From the results in chapter 5.2.2 we can conclude that the Empirical Bayes method gives better estimates of the true value of π_i compared to the other general imputation methods. This conclusion is valid for different amount of missing values and different values of the data parameters, such as number of observations per patient, number of patients, mean and variance.

For the Expectation Maximization method the same conclusion can be drawn as for the Empirical Bayes method with the restriction that the amount of missing values at least should be more than 10% if we are to get significantly better estimates.

When the beta/binomial model was compared to the Gaussian/Gaussian model it was concluded that the latter gave a poor fit to the data as we expected. Because of the poor fit the estimated prior variance is much higher than in the beta/binomial model and as a result of that the shrinkage is smaller. When examining how efficient the models were to estimate the 'true' estimates the Gaussian/Gaussian could be considered the inferior. Observe that these conclusions were made when the data sets used in this essay was analyzed and that the Gaussian/Gaussian model surely can be a well functioning model for other data sets.

The EB and EM method are not conservative imputation methods because they do not make the estimates worse if missing values exist. They simply pull the estimate toward the overall mean value with weight depending on the number of missing values and that is not a procedure that makes the method conservative.

8 References

1. Carlin, B.P. and Louis, T.A. (2000) *Bayes an Empirical Bayes Methods for Data Analysis*, New York: Chapman & Hall.

2. Commitee for Proprietary Medicinal Products (2001) *Points to consider on Missing Data*, London: The European Agency for the Evaluation of Medicinal Products.

3. Encyclopedia of Biostatistics (1998), Wiley.

4. Gut A. (1995) *An Intermediate Course in Probability*, New York: Springer-Verlag.

5. Louis, T.A and Bailey, J.K. (1990). Controlling error rates using prior information and marginal totals to select tumor sites. *Journal of Statistical Planning and Inference*, 24, 297-316.

6. Zwanzig, S. (2003) *Computer Intensive Statistical Methods*, Department of Mathematics: Uppsala University.

A Appendix

A.1 The beta distribution

The beta distribution is a continuous distribution with density function

$$P(x|r,s) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1-x)^{s-1}$$

$$E(X) = \frac{r}{r+s}$$

$$Var(X) = \frac{r(r+s)}{(r+s)(r+s+1)},$$

where $x \in [0,1]$, r $>$ and s > 0. The beta distribution is defined on the unit interval and depending on r and s it can take several different types of shapes. If both r = s = 1 it is the U(0,1)-distribution, see figure (18). If r < 1 the distribution reaches infinity as x $\rightarrow 0$ and if s < 1 it goes to infinity as x $\rightarrow 1$. If r and s < 1 the beta is concave up and if r and s > 1 it is concave down. Note that the higher r+s is the lower is the variance. Some illustrations of the beta distribution with different parameter values are shown in figure (18).

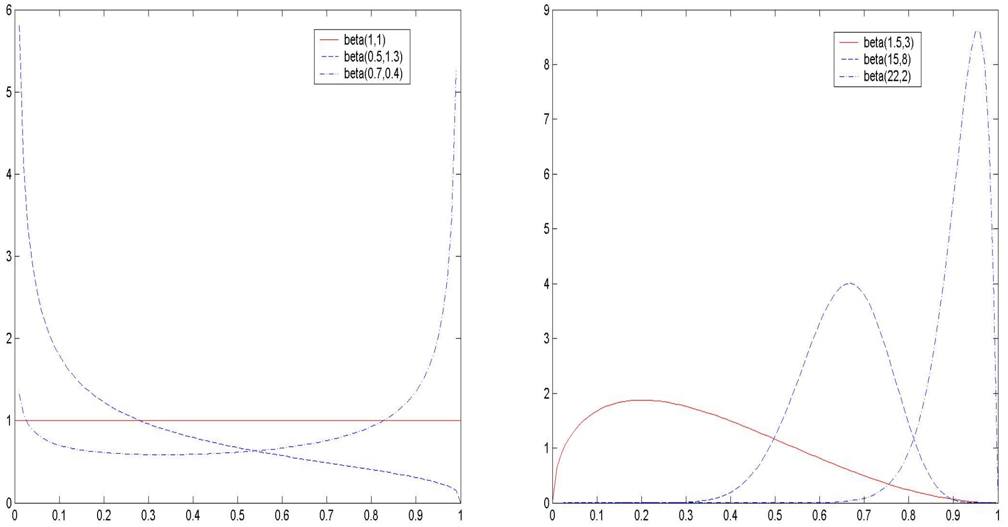


Figure 18: *The beta distribution with six sets of different parameter values.*

A.2 Small simulation

i	n_i	true π_i	$\hat{\pi}_i$	shr EB	shr EM	shr EB	shr EM
		$beta_{rep}(0.6, 20)$		b/b	b/b	G/G	G/G
$\hat{\mu}$		0.6625	0.671	0.675	0.637	0.684	0.696
$\widehat{M}/\hat{\tau}^2$ *prior*		6.396	2.971	5.53/0.034	5.53/0.034	-/0.056	-/0.056
$\widehat{M}/\hat{\tau}^2$ *post.*		6.396	2.971	12.38/0.016	8.38/0.025	-/0.033	-/0.061
1	1	0.609	0	0.568	0.414	0.579	0.426
2	5	0.719	0.800	-0.068	-0.047	-0.011	-0.116
3	13	0.457	0.385	0.085	0.070	0.036	0.036
4	10	0.713	0.700	-0.011	-0.008	-0.068	0.037
5	3	0.657	0	0.435	0.109	0.368	0.268
6	3	0.653	1.000	-0.214	-0.053	-0.158	0
7	0.5	0.807	0	0.615	0.571	0.420	0.689
8	8	0.502	0.750	-0.032	-0.023	-0.013	0.039
9	6	0.691	0.333	0.162	0.135	0.088	-0.070
10	2	0.704	1.000	-0.242	-0.098	-0.210	0
11	11	0.679	0.909	-0.080	-0.028	-0.014	-0.014
12	3	0.640	1.000	-0.213	-0.053	-0.210	0
13	15	0.774	0.933	-0.071	-0.018	-0.038	0.014
14	15	0.518	0.667	0.001	0.001	0.017	0.017
15	17	0.634	0.882	-0.052	-0.020	-0.040	0.012
16	19	0.608	0.526	0.033	0.027	0	0
17	2	0.619	1.000	-0.242	-0.098	-0.421	0
18	5	0.497	0.200	0.247	0.171	0.063	0.221
19	14	0.476	0.571	0.028	0.024	0.008	0.007
20	19	0.627	0.684	-0.003	-0.002	0	0

Table 3: *Shrinkage (shr), estimate (EB or EM) - $\hat{\pi}_i$, with the usage of the EB- and the EM-method with the beta/binomial and Gaussian/Gaussian models. Expected number of missing values are 50%*

The box-plots in figure 19 and 20 shows:

1. difference between Non-Bayesian estimate of π_i and 'true' π_i.
2. difference between EB estimate of π_i and 'true' π_i, beta/binomial model.
3. difference between EB estimate of π_i and 'true' π_i, Gaussian/Gaussian model.
4. difference between EB estimate of π_i and EM estimate of π_i, beta/binomial model.

5. Difference between EB estimate of π_i and EM estimate of π_i, Gaussian/Gaussian model.

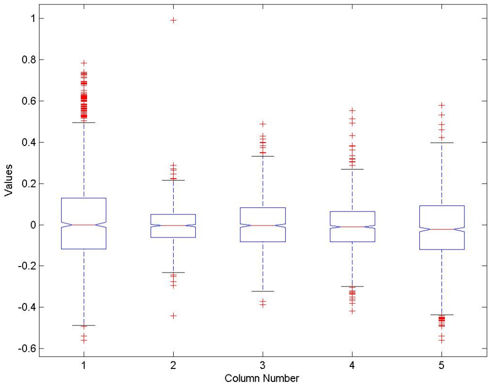


Figure 19: *Missing values 50%, M=40, μ=0.6 and n=20. Repeated 50 times for 20 patients.*

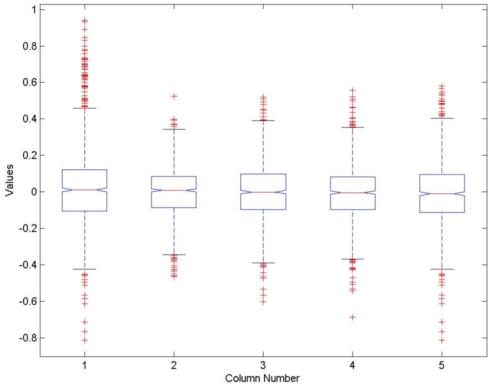


Figure 20: *Missing values 50%, M=5, μ=0.6, N=20 and n=20. The simulation is repeated 50 times.*