UiO : Department of Mathematics
University of Oslo

STK4021

Course notes and examples

Ida Scheel

October 27, 2016

Table of contents

The empirical Bayes approach- based on Ch 5 in [1]

- Consider the model

$$y \mid \theta \sim p(y \mid \theta)$$
$$\theta \mid \varphi \sim p(\theta \mid \varphi)$$

- For a full Bayesian approach, we either fix φ based on prior knowledge (two-level model), or we give φ a prior distribution $p(\varphi)$ (more than two levels), which again can depend on parameters. At some level however, we have to stop adding parameters, and at the top level, some quantities must be fixed
- For the empirical Bayes approach, we use point estimates of φ, estimated from data
- Can in principle be used for any number of levels of the hierarchy, for example if we give φ a prior distribution $p(\varphi \mid \rho)$, then we could use empirical Bayes estimates of ρ

The empirical Bayes (EB) estimates and the estimated posterior distribution

- Remember the marginal distribution of y given φ

$$p(y \mid \varphi) = \int p(y \mid \theta) p(\theta \mid \varphi) d\theta$$

- This is used to find the EB estimates $\hat{\varphi} \equiv \hat{\varphi}(y)$, e.g. by maximising $p(y \mid \varphi)$ w.r.t φ
- The estimated posterior distribution is then $p(\theta \mid y, \hat{\varphi})$
- This is a parametric EB approach, non-parametric approaches also exist
- NB: Posterior probability intervals for θ must be constructed with care (see e.g. [1]), to incorporate uncertainty about φ

EB for a Normal model

- Consider the two-layer Normal model

$$y_i \mid \theta_i \sim N(\theta_i, \sigma^2), i = 1, \ldots, n$$
$$\theta_i \mid \mu \sim N(\mu, \tau^2), i = 1, \ldots, n$$

  where σ^2 and τ^2 are assumed known constants, hence μ is the only random hyperparameter, for which we wish to find a EB estimate
- Now it is quite straightforward to show that

$$p(y_i \mid \mu) = \int \left[\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(y_i - \theta_i)^2}{2\sigma^2} \right) \frac{1}{(2\pi\tau^2)^{1/2}} \exp\left(-\frac{(\theta_i - \mu)^2}{2\tau^2} \right) \right] d\theta_i$$
$$= \frac{1}{(2\pi(\sigma^2 + \tau^2))^{1/2}} \exp\left(-\frac{1}{2(\sigma^2 + \tau^2)} (y_i - \mu)^2 \right), \; i = 1, \ldots, n$$

EB for a Normal model

- Hence

$$
\begin{aligned}
p(y \mid \mu) &= \prod_{i=1}^{n} p(y_i \mid \mu) \\
&= \frac{1}{\left(2\pi(\sigma^2 + \tau^2)\right)^{k/2}} \exp\left(-\frac{1}{2(\sigma^2 + \tau^2)} \sum_{i=1}^{n} (y_i - \mu)^2 \right)
\end{aligned}
$$

which is obviously maximised for $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- The EB estimated posterior distribution is hence given by

$$
\begin{aligned}
p(\theta_i \mid y_i, \hat{\mu}) &\propto p(y_i \mid \theta_i) \cdot p(\theta_i \mid \hat{\mu}) \\
&= N\left(\frac{\sigma^2 \hat{\mu} + \tau^2 y_i}{\sigma^2 + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \right), \ i = 1, \ldots, n
\end{aligned}
$$

Single-parameter model for epidemiology

- Concerns estimating a rate from Poisson data (idealized example from the textbook [2], pp 45-46)
- Consider a survey of the causes of death in a single year for a city in the US
- Population 200.000, $y = 3$ persons died of asthma
  - Crude estimate of $3/200.0000 = 1.5$ per 100.000 persons per year
- For epidemiological data like this, a Poisson sampling distribution is commonly used, assuming exchangeability given exposure and rate parameter
  - Let θ be the true, underlying long-term asthma mortality rate per 100.000 persons per year in the city
  - The exposure is $x = 2.0$ (since θ) is defined per 100.000 persons per year)
  - Hence, the sampling distribution is $y \sim \text{Poisson}(2.0\theta)$

Prior distribution

- Asthma mortality rates around the world typically are around 0.6 per 100.000, and rarely above 1.5 per 100.000 in Western countries
- Assume exchangeability between this city and other Western cities, and this year and other years
- Know that Gamma(a, b) is the conjugate prior prior distribution, use that for convenience, must find suitable values of a and b that match the prior information
- Book: $\theta \sim$ Gamma$(3.0, 5.0)$ (mean=0.6, 97.5% of the mass lies below 1.44, prior probability of $\theta < 1.5$ 98.0%)
- Slightly different (with more uncertainty) suggestion: $\theta \sim$ Gamma$(1.2, 2.0)$ (mean=0.6, 97.5% of the mass lies below 2.05, prior probability of $\theta < 1.5$ 92.9%)
- “rarely above 1.5 per 100.000” is open for interpretation

Posterior distribution

- We know that the posterior distribution for θ will be $Gamma(a + y, b + x)$
- Book prior: Posterior is Gamma$(6.0, 7.0)$
- Alternative prior: Posterior is Gamma$(4.2, 4.0)$
- The two different priors yields somewhat different posterior distributions and conclusions (see R-script)
- Little data!!! Prior is influential

Posterior distribution with additional data

- Additional data: Suppose we now have 10 years of data, with 30 deaths caused by asthma over the 10 years. Assume the population size is constant at 200.000 over the period
- Now $y = 30$ and the exposure is $x = \frac{200.000 \times 10}{100.000} = 20$ (since θ is defined per 100.000 persons per year)
- Book prior: Posterior is Gamma$(33.0, 25.0)$
- Alternative prior: Posterior is Gamma$(31.2, 22.0)$
- The posterior results with the two different priors are more similar now with more data, but still slightly different (see R-script)
- More data, prior is less influential

Alternative specification with n independent outcomes

- Could alternatively say that $y_i, i = 1, \ldots, n$ is the number of deaths caused by asthma per 100.000 persons per year
- Let θ still be the true, underlying long-term asthma mortality rate per 100.000 persons per year in the city
- Then

$$y_i \sim \text{Pois}(x_i\theta), i = 1, \ldots, n$$

  where the exposure is $x_i, i = 1, \ldots, n$
- Then we know that the likelihood is

$$p(y \mid \theta) \propto \theta^{\sum_{i=1}^n y_i} e^{-\theta \sum_{i=1}^n x_i}$$

where in the example we have $y_i = 3, x_i = 2, i = 1, \ldots, n$ for (i) $n = 1$ and (ii) $n = 10$

The normal approximation

- The normal approximation to the posterior distribution for θ based on $\log p(y \mid \theta) = C + \sum_{i=1}^n y_i \log\theta - \theta\sum_{i=1}^n x_i$ can easily be found. First find the mode

$$\frac{d\log p(y \mid \theta)}{d\theta} = \frac{\sum_{i=1}^n y_i}{\theta} - \sum_{i=1}^n x_i$$

which is =0 for $\theta = \hat\theta = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$

- Then the Fisher information (since we allow for different exposure-values x_i, the y_i's are not iid):

$$n \cdot J(\theta) = E\left[-\frac{d^2\log p(y \mid \theta)}{d\theta^2} \mid \theta\right] = E\left[\frac{\sum_{i=1}^n y_i}{\theta^2}\right]$$

$$\underset{(\text{using } E[y_i]=x_i\theta)}{=} \frac{\sum_{i=1}^n x_i\theta}{\theta^2} = \frac{\sum_{i=1}^n x_i}{\theta}$$

The normal approximation

- Hence

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$
$$n \cdot J(\hat{\theta}) = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{\sum_{i=1}^{n} y_i}$$

and for large n

$$p(\theta \mid y) \approx N\left(\theta \mid \hat{\theta}, \left(n \cdot J(\hat{\theta})\right)^{-1}\right)$$

- For $n = 1$ the Normal approximation is $N\left(\frac{3}{2}, \frac{3}{2^2}\right) = N(1.5, 0.75)$ and for $n = 10$ it is $N\left(\frac{30}{20}, \frac{30}{20^2}\right) = N(1.5, 0.075)$

Multinomial sampling distribution with a Dirichlet prior

- Application: 2016 US presidential election poll (Sept-16)
  - $n = 911$ representative, likely voters were asked which candidate they prefer in the 2016 US presidential election
  - $y_1 = 392$ preferred Clinton, $y_2 = 364$ preferred Trump, and $y_3 = 155$ preferred other candidates or had no opinion
- Multinomial sampling distribution with
  - probability θ_1 of preferring Clinton
  - probability θ_2 of preferring Trump
  - probability θ_3 of preferring other candidates or having no opinion
  - $\sum_{i=1}^{3} \theta_i = 1$ (hence there are in fact only two parameters)
- A non-informative uniform Dirichlet(1,1,1) prior for $(\theta_1, \theta_2, \theta_3)$
- Hence, the posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is Dirichlet$(1 + y_1, 1 + y_2, 1 + y_3)$ = Dirichlet$(393, 365, 156)$

The posterior distribution of an estimand of interest

- Suppose we are interested in the posterior distribution of $\frac{\theta_1}{\theta_2}$
- This can easily be approximated by for $i = 1, \ldots, S$ doing
  - Sample $\theta^{(i)}$ from the posterior distribution of $(\theta_1, \theta_2, \theta_3)$
  - Compute $\frac{\theta_1^{(i)}}{\theta_2^{(i)}}$
- The S values of $\frac{\theta_1^{(i)}}{\theta_2^{(i)}}$ for $i = 1, \ldots, S$ are a then samples from the posterior distribution of $\frac{\theta_1}{\theta_2}$

The Normal approximation

- Our interest primarily lies in θ_1 and θ_2, therefore we focus on these two parameters and replace θ_3 by $1 - \theta_1 - \theta_2$
- The likelihood for $\theta = (\theta_1, \theta_2)$ is

$$p(y \mid \theta) \propto \prod_{i=1}^{3} \theta_i^{y_i} = \theta_1^{y_1} \cdot \theta_2^{y_2} \cdot (1 - \theta_1 - \theta_2)^{y_3}$$

- The Normal approximation to the posterior distribution for θ based on
$\log p(y \mid \theta) = C + y_1 \log \theta_1 + y_2 \log \theta_2 + y_3 \log (1 - \theta_1 - \theta_2)$ can easily be found. First find the mode

$$\frac{d \log p(y \mid \theta)}{d\theta_i} = \frac{y_i}{\theta_i} - \frac{y_3}{1 - \theta_1 - \theta_2}, i = 1, 2$$

which is =0 for $\theta_i = \widehat{\theta}_i = \frac{y_i}{\sum_{j=1}^{3} y_j}$

The Normal approximation

- Then the Fisher information:

$$\frac{d^2 \log p(y \mid \theta)}{d\theta_i^2} = -\frac{y_i}{\theta_i^2} - \frac{y_3}{(1-\theta_1-\theta_2)^2}, i = 1, 2$$

$$\frac{d^2 \log p(y \mid \theta)}{d\theta_1 d\theta_2} = -\frac{y_3}{(1-\theta_1-\theta_2)^2}$$

$$\Downarrow$$

$$n \cdot J(\theta) = E\left[-\begin{pmatrix} -\frac{y_1}{\theta_1^2} - \frac{y_3}{(1-\theta_1-\theta_2)^2} & -\frac{y_3}{(1-\theta_1-\theta_2)^2} \\ -\frac{y_3}{(1-\theta_1-\theta_2)^2} & -\frac{y_2}{\theta_2^2} - \frac{y_3}{(1-\theta_1-\theta_2)^2} \end{pmatrix}\right]$$

$$\underset{(\text{using } E[y_i]=n\theta_i)}{=} \begin{pmatrix} -\frac{n}{\theta_1} - \frac{n}{1-\theta_1-\theta_2} & -\frac{n}{1-\theta_1-\theta_2} \\ -\frac{n}{1-\theta_1-\theta_2} & -\frac{n}{\theta_2} - \frac{n}{1-\theta_1-\theta_2} \end{pmatrix}$$

The Normal approximation

- Hence

$$\widehat{\theta}_i = \frac{y_i}{\sum_{j=1}^{3} y_j}$$

$$n \cdot J(\widehat{\theta}) = n \cdot \begin{pmatrix} -\frac{1}{\widehat{\theta}_1} - \frac{1}{1-\widehat{\theta}_1-\widehat{\theta}_2} & -\frac{1}{1-\widehat{\theta}_1-\widehat{\theta}_2} \\ -\frac{1}{1-\widehat{\theta}_1-\widehat{\theta}_2} & -\frac{1}{\widehat{\theta}_2} - \frac{1}{1-\widehat{\theta}_1-\widehat{\theta}_2} \end{pmatrix}$$

  and for large n

$$p(\theta \mid y) \approx N\left(\theta \mid \hat{\theta}, \left(n \cdot J(\hat{\theta})\right)^{-1}\right)$$

- Here we know the exact posterior distribution, can compare it to the Normal approximation by for example contour-plots (see R-script)

The application and sampling distribution

- Example from the textbook [2], section 3.7
- A bioassay experiment typically concerns giving various dose levels of a drug/chemical compound to a batch of animals and measure a binary response (alive/dead or tumor/no tumor)
- The data for k dose levels are of the form

$$(x_i, n_i, y_i),\ i = 1, \dots, k$$

  where x_i is the i'th dose level given to n_i animals of which y_i animals responded with "success" (e.g. death)
- Reasonable to model the response of the animals within the i'th group (given dose x_i) as exchangeable, by modelling them as independent with equal probabilities of success θ_i, i.e. a binomial model

$$y_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$$

Logistic regression model for the probabilities

- The parameters $\theta_1, \ldots, \theta_k$ should be not be modelled as exchangeable, since we have the dose levels $x_1, \ldots, x_k$
- Rather model the pairs $\theta_i \mid x_i,\ i = 1, \ldots, k$ by a logistic regression model

$$\text{logit}(\theta_i) = \alpha + \beta x_i,\ i = 1, \ldots, k$$

  where $\text{logit}(\theta_i) = \log \frac{\theta_i}{1-\theta_i}$ is the logistic transformation
- Hence $\theta_i = \text{logit}^{-1}(\alpha + \beta x_i) = \frac{e^{\{\alpha+\beta x_i\}}}{1+e^{\{\alpha+\beta x_i\}}}$ and the likelihood contribution from group i for the parameters α and β is

$$\begin{aligned} p(y_i \mid \alpha, \beta) &\propto \theta_i^{y_i}(1-\theta_i)^{n_i - y_i} \\ &= \left(\frac{e^{\{\alpha+\beta x_i\}}}{1 + e^{\{\alpha+\beta x_i\}}}\right)^{y_i} \left(\frac{1}{1 + e^{\{\alpha+\beta x_i\}}}\right)^{n_i - y_i} \end{aligned}$$

Prior and posterior distributions

- We assume an improper prior distribution for the parameters α and β: $p(\alpha, \beta) \propto 1$
- Hence, α and β are independent apriori and marginally uniformly distributed
- Hence the joint posterior distribution for α and β can be expressed as

$$\begin{aligned} p(\alpha, \beta \mid y) &\propto p(\alpha, \beta) \prod_{i=1}^{k} p(y_i \mid \alpha, \beta, n_i, x_i) \\ &= \prod_{i=1}^{k} \left(\frac{e^{\{\alpha+\beta x_i\}}}{1 + e^{\{\alpha+\beta x_i\}}} \right)^{y_i} \left(\frac{1}{1 + e^{\{\alpha+\beta x_i\}}} \right)^{n_i - y_i} \end{aligned}$$

Data and graph of the model

Bioassay data from an experiment (table 3.1 from the textbook [2], see the textbook for reference)

Dose, x_i (log g/ml)	Number of animals n_i	Number of deaths y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

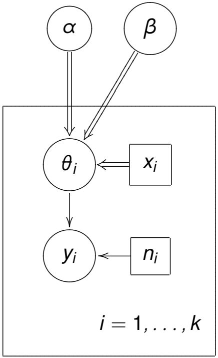

Figure: Graph representation of the model

Posterior analysis

- The normalised posterior distribution is not available analytically
- Hence, some numerical approximation must be performed, e.g. sampling
- Book, ch 3.7, compute the posterior density on a grid of points, then normalise by setting the total probability over the grid of points equal to1
- Later we can do e.g. MCMC
- Now: Normal approximation (Exercise 2 in Chapter 4)

The normal approximation

- The ML-estimates of (α, β) can be found by using standard software for logistic regression, the results are (from the textbook) $(\hat{\alpha}, \hat{\beta}) = (0.8, 7.7)$
- Log-likelihood for one datapoint:

$$
\begin{aligned}
l_i &= \log p(y_i \mid \alpha, \beta) \\
&= C + y_i \log\left(\frac{e^{\{\alpha+\beta x_i\}}}{1 + e^{\{\alpha+\beta x_i\}}}\right) + (n_i - y_i) \log\left(\frac{1}{1 + e^{\{\alpha+\beta x_i\}}}\right) \\
&= C + y_i(\alpha + \beta x_i) - n_i \log\left(1 + e^{\{\alpha+\beta x_i\}}\right)
\end{aligned}
$$

- Hence

$$
\begin{aligned}
\frac{dl_i}{d\alpha} &= y_i - \frac{n_i e^{\{\alpha+\beta x_i\}}}{1 + e^{\{\alpha+\beta x_i\}}} \\
\frac{dl_i}{d\beta} &= y_i x_i - \frac{n_i x_i e^{\{\alpha+\beta x_i\}}}{1 + e^{\{\alpha+\beta x_i\}}}
\end{aligned}
$$

The normal approximation

- The second partial derivatives:

$$\frac{d^2 l_i}{d\alpha^2} = -\frac{n_i e^{\{\alpha+\beta x_i\}}\left(1 + e^{\{\alpha+\beta x_i\}}\right) - n_i e^{\{\alpha+\beta x_i\}} e^{\{\alpha+\beta x_i\}}}{(1 + e^{\{\alpha+\beta x_i\}})^2}$$

$$= -\frac{n_i e^{\{\alpha+\beta x_i\}}}{(1 + e^{\{\alpha+\beta x_i\}})^2}$$

$$\frac{d^2 l_i}{d\beta^2} = -\frac{n_i x_i^2 e^{\{\alpha+\beta x_i\}}\left(1 + e^{\{\alpha+\beta x_i\}}\right) - n_i x_i e^{\{\alpha+\beta x_i\}} x_i e^{\{\alpha+\beta x_i\}}}{(1 + e^{\{\alpha+\beta x_i\}})^2}$$

$$= -\frac{n_i x_i^2 e^{\{\alpha+\beta x_i\}}}{(1 + e^{\{\alpha+\beta x_i\}})^2}$$

$$\frac{d^2 l_i}{d\alpha d\beta} = -\frac{n_i x_i e^{\{\alpha+\beta x_i\}}\left(1 + e^{\{\alpha+\beta x_i\}}\right) - n_i e^{\{\alpha+\beta x_i\}} x_i e^{\{\alpha+\beta x_i\}}}{(1 + e^{\{\alpha+\beta x_i\}})^2}$$

$$= -\frac{n_i x_i e^{\{\alpha+\beta x_i\}}}{(1 + e^{\{\alpha+\beta x_i\}})^2}$$

The normal approximation

- The normal approximation for (α, β) has mean $(\hat{\alpha}, \hat{\beta})$ and covariance matrix $\left(n \cdot J((\hat{\alpha}, \hat{\beta}))\right)^{-1}$, where (remember that $y_1, \ldots, y_k$ are not identically distributed)

$$n \cdot J((\hat{\alpha}, \hat{\beta})) = \begin{pmatrix} \sum_{i=1}^{k} \frac{n_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2} & \sum_{i=1}^{k} \frac{n_i x_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2} \\ \sum_{i=1}^{k} \frac{n_i x_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2} & \sum_{i=1}^{k} \frac{n_i x_i^2 e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2} \end{pmatrix}$$

The normal approximation

- The normal approximation variances are the diagonal elements of $\left(n \cdot J((\hat{\alpha}, \hat{\beta}))\right)^{-1}$, hence

$$\widehat{\mathrm{Var}(\alpha)} = \frac{\sum_{i=1}^{k} \frac{n_i x_i^2 e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}}{\left(\sum_{i=1}^{k} \frac{n_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}\right)\left(\sum_{i=1}^{k} \frac{n_i x_i^2 e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}\right) - \left(\sum_{i=1}^{k} \frac{n_i x_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}\right)^2}$$

$$\widehat{\mathrm{Var}(\beta)} = \frac{\sum_{i=1}^{k} \frac{n_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}}{\left(\sum_{i=1}^{k} \frac{n_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}\right)\left(\sum_{i=1}^{k} \frac{n_i x_i^2 e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}\right) - \left(\sum_{i=1}^{k} \frac{n_i x_i e^{\{\hat{\alpha}+\hat{\beta}x_i\}}}{\left(1+e^{\{\hat{\alpha}+\hat{\beta}x_i\}}\right)^2}\right)^2}$$

References I

B. P. Carlin and T. A. Louis
Bayesian Methods for Data Analysis, Third edition.
Chapman&Hall/CRC Texts in statistical science, 2009.

A. Gelman, J. B. Carlin, H. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin
Bayesian Data Analysis, Third edition.
Chapman&Hall/CRC Texts in statistical science, 2014.