Modellwahl (Bayes-Faktor, DIC)

Volker Schmid

3. Juli 2017

Bayes-Faktor

Devianz-Informationskriterium (DIC)

Bayes-Faktor

Bayes-Faktor

Für zwei alternative Hypothesen $H_0 : \theta \in \Theta_0$ und $H_1 : \theta \in \Theta_1$ ist der **Bayes-Faktor**:

$$B(x) = \frac{P(x|H_0)}{P(x|H_1)} = \frac{\frac{P(\theta\in\Theta_0|x)}{P(\theta\in\Theta_1|x)}}{\frac{p(\theta\in\Theta_0)}{p(\theta\in\Theta_1)}}$$

Modellwahl über Bayes-Faktor

- Gegeben: Daten x, K verschiedene Modelle
- Wahrscheinlichkeit, x unter Modell k zu beobachten: $p(x|M_k)$
- Priori auf Modelle: $p(M_k)$
- Posteriori-Odds:

$$\frac{p(M_k|x)}{p(M_l|x)} = \frac{p(x|M_k)p(M_k)}{p(x|M_l)p(M_l)} \frac{p(x)}{p(x)}$$

Anmerkungen

- $\frac{p(x|M_k)}{p(x|M_l)}$ ist der Bayesfaktor zugunsten von M_k
- Bayesfaktor unabhängig von der Priori auf die Modelle
- Zähler/Nenner ist die **marginale Likelihood**

$$p(x|M_k) = \int p(x|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \tag{1}$$

- Für einfache Hypothesen $(H_0 : \theta = \theta_0, H_1 : \theta = \theta_1)$ ist der Bayesfaktor gleich dem Likelihood ratio.

Skala des Bayes-Faktors

Nach Jeffreys (1961) kann der Bayes-Faktor wie folgt interpretiert werden:

$B < 1$	H_0 wird gestützt
$B \in [1, 10^{1/2}]$	Anzeichen gegen H_0, aber kaum erwähnenswert
$B \in [10^{1/2}, 10]$	beachtliche Anzeichen gegen H_0
$B \in [10, 10^{3/2}]$	starke Anzeichen gegen H_0
$B \in [10^{3/2}, 100]$	sehr starke Anzeichen gegen H_0
$B > 100$	ausschlaggebende Anzeichen gegen H_0

Beispiel Brandmeldungen in Frankville, NC

Quelle: Jim Albert, LearnBayes Vignette

Daten: Anzahl von Brandmeldungen in aufeinanderfolgenden Monaten: $y_1, ..., y_N$. Modell

- $y_1, ..., y_N \sim f(y|\theta)$
- $\theta \sim g(\theta)$

- Priorannahmen über den Erwartungswert von y: Erwartungswert ist 70, Standardabweichung ist 10.
- Unklar: welche Verteilung haben die Daten?

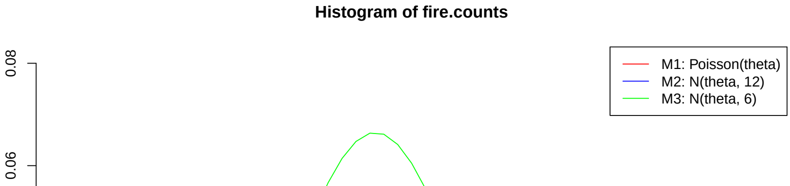
Verschiedene Modellannahmen

1. $y \sim Po(\theta)$
2. $y \sim N(\theta, 12^2)$
3. $y \sim N(\theta, 6^2)$

In allen Fällen: $\theta \sim Ga(280, 4)$. $(E(\theta) = 70, sd(\theta) = 4.2)$

Visueller Vergleich

```
fire.counts <- c(75, 88, 84, 99, 79, 68, 86, 109, 73, 85, 1
                 75, 81, 64, 77, 83, 83, 88, 83, 78, 83, 78
                 82, 90, 74, 72, 69, 72, 76, 76, 104, 86, 9
hist(fire.counts, probability=TRUE, ylim=c(0, .08))
x <- 60:110
lines(x, dpois(x, lambda=mean(fire.counts)), col="red")
lines(x, dnorm(x, mean=mean(fire.counts), sd=12), col="blue
lines(x, dnorm(x, mean=mean(fire.counts), sd=6), col="green
legend("topright", legend=c("M1: Poisson(theta)",
                            "M2: N(theta, 12)",
                            "M3: N(theta, 6)"),
       col=c("red", "blue", "green"), lty=1)
```



Prädiktive/Marginale Dichte von y

$$f(y) = \int \prod_{j=1}^{N} f(y_j|\theta)g(\theta)d\theta.$$

Laplace-Approximation {.allowframebreaks}

- Berechne Posteriori-Modus
- Taylor-Approximation der Log-Likelihood

```
laplace <- function (logpost, mode, ...)
{
  fit = optim(mode, logpost, gr = NULL,
    ..., hessian = TRUE, control = list(fnscale = -1))
  mode = fit$par
  h = -solve(fit$hessian)
  p = length(mode)
  int = p/2 * log(2 * pi) + 0.5 * log(det(h)) +
  logpost(mode, ...)
  stuff = list(mode = mode, var = h, int = int,
               converge = fit$convergence == 0)
```

Erstes Modell

```r
model.1 <- function(theta, y){
  sum(log(dpois(y, theta))) +
    dgamma(theta, shape=280, rate=4)
}
log.pred.1 <- laplace(model.1, 80, fire.counts)$int
log.pred.1
```

```
## [1] -131.6253
```

Zweites und drittes Modell

```r
model.2 <- function(theta, y){
  sum(log(dnorm(y, theta, 6))) +
    dgamma(theta, shape=280, rate=4)
}
model.3 <- function(theta, y){
  sum(log(dnorm(y, theta, 12))) +
    dgamma(theta, shape=280, rate=4)
}
log.pred.2 <- laplace(model.2, 80, fire.counts)$int
log.pred.3 <- laplace(model.3, 80, fire.counts)$int
```

Modellvergleich

```r
data.frame(Model=1:3, log.pred=c(log.pred.1, log.pred.2, lo
```

```
##   Model  log.pred
## 1     1 -131.6253
## 2     2 -144.7554
## 3     3 -132.9464
```

```r
exp(log.pred.1 - log.pred.3)
```

```
## [1] 3.74751
```

Nach Jeffreys *beachtliche Anzeichen* für Modell 1 gegenüber Modell 3.

Devianz-Informationskriterium (DIC)

Informationskriterien

Übliche Informationskriterien (z.B. AIC und BIC) messen die Anpassung plus Anzahl der Parameter. Bei hierarchischen Modellen ist die Anzahl der Parameter jedoch nicht aussagekräftig, insbesondere wenn die Parameter abhängig sind. Deshalb ist die Idee hier, die *effektive Anzahl* der Parameter zu schätzen.

Devianz

Der Ausdruck

$$D(y) := -2\Big(\log\left(p(y|\hat{\theta})\right) - \log\left(p(y|\hat{\theta}_s)\right)\Big)$$

heißt Devianz

Dabei sind $\hat{\theta}$ die geschätzen Parameter eines Modells M und $\hat{\theta}_s$ die geschätzten Parameter in einem saturierten Modell (Daten komplett angepasst). Dies entspricht minus zweimal dem Logarithmus des Likelihoodratios. Manchmal wird auch nur $-2\log\left(p(y|\hat{\theta})\right)$ als Devianz bezeichnet.

Maß der Reduktion der Überraschung

Sei $\hat{\theta}$ ein Schätzer für θ^{TRUE}. Betrachte nun

$$d(y, \theta^{TRUE}, \hat{\theta}) = -2 \log \left(p(y|\theta^{TRUE}) \right) + 2 \log \left(p(y|\hat{\theta}) \right)$$

Dies kann interpretiert werden als *Maß der Reduktion der Überraschung* oder Unsicherheit aufgrund der Schätzung.

Vorschlag von Spiegelhalter et al.(2002): Benutze Posteriori-Erwartungswert von $d(y, \theta^{TRUE}, \hat{\theta})$ als *Effektive Anzahl der Parameter* im Bayes-Modell.

Effektive Anzahl von Parametern

$$E(d(y, \theta^{TRUE}, \hat{\theta}) = E(-2 \log \left(p(y|\theta^{TRUE})\right)) + E(2 \log \left(p(y|\hat{\theta})\right))$$
$$=: \widehat{D(\theta)} - D(\hat{\theta}) =: p_D$$

p_D ist abhängig von

- den Daten y
- dem Datenmodell
- der Priori $p(\theta)$
- der Wahl des Punktschätzers $\hat{\theta}$

Definition DIC

Das **Deviance Information Criterion (DIC)** ist definiert als

$$DIC = \widehat{D(\theta)} + p_D$$

Dies entspricht dem Bayesianischen Maß für die Anpassung des Fits plus den Komplexitätsterm p_D.

Es gilt: $\widehat{D(\theta)} = E(D(\theta)) = E(-2\log(p(y|\theta)))$ wird kleiner für bessere Anpassung.

Anmerkungen DIC

- Bei substantiellem Konflikt zwischen Priori und Daten sowie bei nicht unimodalen Posterioris kann p_D negativ werden.
- Falls kein hierarchische Modell vorliegt und alle Prioris komplett spezifiziert sind, gilt

$$AIC = D(\hat{\theta}_{ML}) + 2 \cdot p$$

Berechnung des DIC bei MCMC I

- $\hat{D(\theta)}$: In jeder MCMC-Iteration k berechne $D(\theta^{(k)}$ und schätze $\hat{D(\theta)}$ über Mittelwert (Median)
- $D(\hat{\theta})$: Benutze Punktschätzer (Mittelwert, Median) und setze diesen in D ein (plug-in Schätzer)

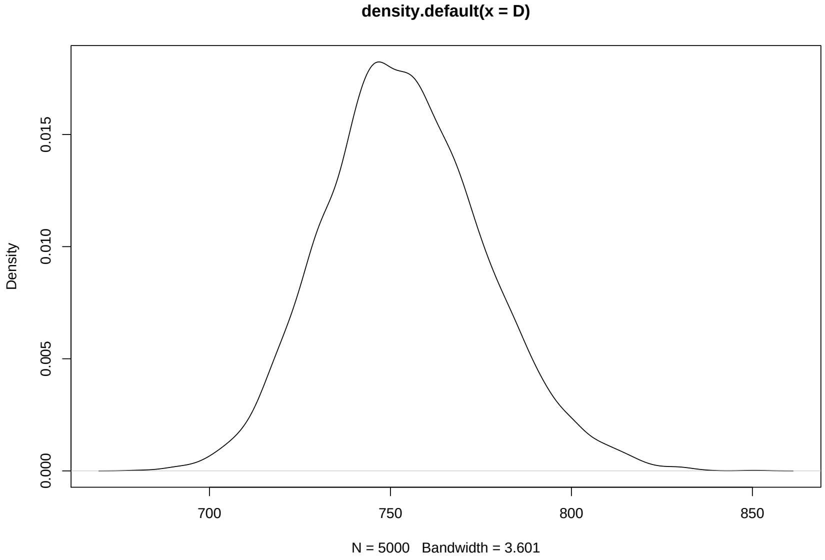
Poisson-Log-Normal-Modell (Hausübung 1):

```
rw=1
source("drivers_example_mcmc.R")
```

Berechnung des DIC bei MCMC II

```
mu<-array(0,c(T,nr.it))
D <- rep(NA,nr.it)
sigma<-sigma2.save
for (i in 1:nr.it)
     {
        mu[,i]=alpha.save[i]+belt*beta.save[i]+
          gamma.save[,i]+delta.save[,i]
        D[i] <- -2*sum(dnorm(y, mean=mu[,i],
                    sd=sqrt(sigma[i]),log=TRUE))
}
plot(density(D))
```

Berechnung des DIC bei MCMC III



Berechnung des DIC bei MCMC IV

```
D.theta.hat <- -2*sum(dnorm(y,apply(mu,1,median),
                     sqrt(median(sigma)),log=TRUE))
D.hat <- median(D)
pD <- D.hat - D.theta.hat
(DIC1<-data.frame("rw"=1,"D"=D.hat,"pD"=pD,
                                "DIC"=D.hat+pD))
```

```
##   rw        D       pD      DIC
## 1  1 753.0946 29.02179 782.1164
```

Als Vergleichsmodell Zeittrend mit Random Walk zweiter Ordnung als Priori (RW2)

```
rw=2
source("drivers_example_mcmc.R")
```

Berechnung des DIC bei MCMC V

```
## One iteration takes roughly 0.01 seconds. Estimated tota
```

```
mu2<-array(0,c(T,nr.it))
D2 <- rep(NA,nr.it)
for (i in 1:nr.it)
     {
        mu2[,i]=alpha.save[i]+belt*beta.save[i]+
                    gamma.save[,i]+delta.save[,i]
        D2[i] <- -2*sum(dnorm(y, mean=mu2[,i],
              sd=sqrt(sigma2.save[i]),log=TRUE))
}
D.theta.hat2 <- -2*sum(dnorm(y,apply(mu2,1,median),
              sqrt(median(sigma2.save)),log=TRUE))
D.hat2 <- median(D2)
pD2 <- D.hat2 - D.theta.hat2
DIC2<-data.frame("rw"=2,"D"=D.hat2,"pD"=pD2,
                                   "DIC"=D.hat2+pD2)
```

Berechnung des DIC bei MCMC VI

und ohne zeitlichen Trend (0):

```
print(rbind(DIC3,DIC1,DIC2))
```

```
##   rw        D       pD      DIC
## 1  0 880.2732 15.90610 896.1793
## 2  1 753.0946 29.02179 782.1164
## 3  2 753.9989 29.80447 783.8033
```