Bayes'scher Ansatz

Peter von Rohr

2016-05-22

Abkürzungen

Abk	Bedeutung
REML	Restricted oder Residual Maximum Likelihood
BLUP	Best Linear Unbiased Prediction
ML	Maximum Likelihood
MCMC	Markov Chain Monte Carlo

Contents

Erklärung

Dieses Dokument enthält eine Einführung in Bayes'sche Statistik im Hinblick auf deren Verwendung in der genomischen Selektion. Der präsentierte Inhalt ist als Lerntext organisiert. Das heisst nach jedem Abschnitt wird eine Kontrollfrage gestellt. Kann der/die LeserIn die Frage beantworten, kann der nachfolgende Abschnitt in Angriff genommen werden. Andernfalls empfielt es sich den vorangegangenen Abschnitt noch einmal zu repetieren.

Einführung

In der Statistik gibt es zwei verschiedene Lehrmeinungen. Es sind dies

1. die **Frequentisten** und
2. die **Bayesianer**.

Alle bisher [1] vorgestellten statistischen Konzepte, so zum Beispiel `Least Squares`, `Maximum Likelihood`, `REML` und `BLUP` stammen aus dem Lager der Frequentisten.

Die Unterschiede zwischen Frequentisten und Bayesianern bestehen hauptsächlich in

- deren Verständnis von Wahrscheinlichkeiten
- deren Unterteilung von Modell- und Datenkomponenten
- deren Techniken zur Schätzung von Parametern

Die folgende Tabelle gibt eine Übersicht über die Unterschiede.

Was	Frequentisten	Bayesianer
Wahrscheinlichkeiten	Eigenschaften von Zufallsvariablen, welche bei grossen Stichproben eintreten	Mass für Informationsgehalt unabhängig von Stichprobengrösse
Modell- und Datenkomponenten	Unterscheidung zwischen Modellparametern und Daten. Parameter sind unbekannt, Daten sind bekannt. Fehlende Daten werden ignoriert	Unterscheidung zwischen unbekannten und bekannten Grössen, unabhängig ob Parameter oder Daten. Fehlende Daten können simuliert werden
Schätztungen von Parametern	ML oder REML werden für Parameterschätzung verwendet	MCMC Zufallszahlen zur Approximation der gewünschten a posteriori Verteilungen

---

Kontrollfrage 1: Unabhängig davon ob Parameter oder Daten unterteilen Bayesianer die Grössen in Kategorien von ...

---

[1] Hier ist nicht nur diese Vorlesung sondern auch die Züchtungslehre und die angewandet Zuchtwertschätzung gemeint

Das Lineare Modell

Die Bayes'sche Art der Parameterschätzung soll an einem einfachen linearen Modell gezeigt werden. Angenommen, wir betrachten das Modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{1}$$

wobei y_i die i-te Beobachtung einer Zielgrösse ist, β_0 für den Achsenabschnitt steht, x_1 eine erklärende Variable ist und ϵ_i für den Restterm steht. Für den Restterm nehmen wir an, dass deren Varianz konstant gleich σ^2 ist.

Bekannte und Unbekannte

Unter der Annahme, dass wir für die Zielgrösse y_i und die erklärende Variable x_1 keine fehlenden Daten haben, dann machen wir als Bayesianer folgende Einteilung in bekannte und unbekannte Grössen.

und als **bekannte** Grössen

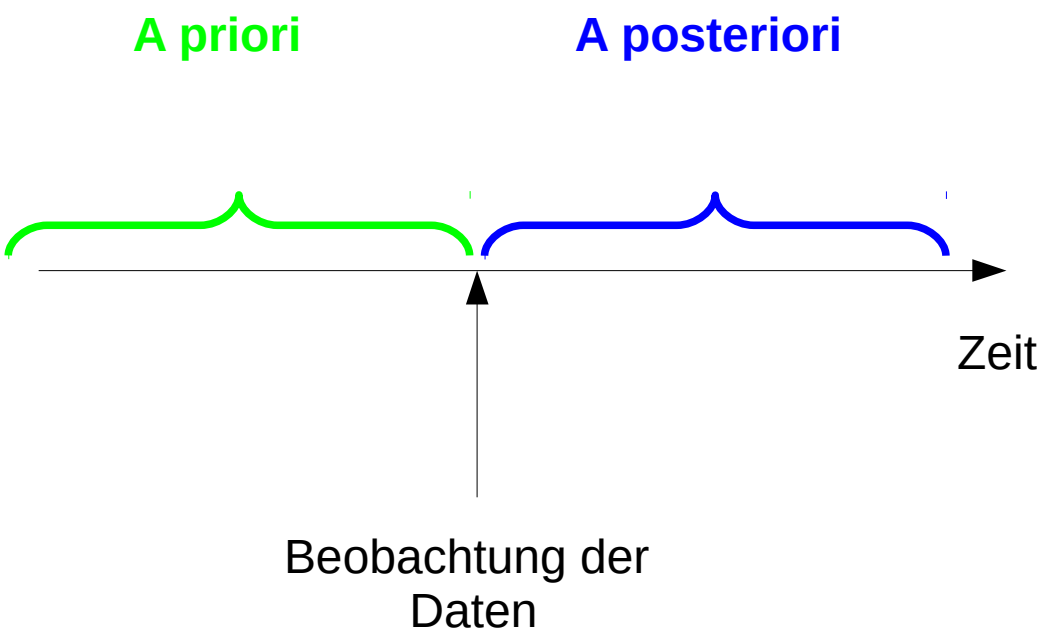
Was	bekannt	unbekannt
y_i	X	
x_1	X	
β_0		X
β_1		X
σ^2		X

Kontrollfrage 2: Unter der Annahme, dass bei der Zielgrösse und der erklärenden Variablen keine Daten fehlen, welcher Einteilung bei den Frequentisten entspricht dann die Bayes'sche Einteilung in bekannte und unbekannte Grössen?

Vorgehen bei Parameterschätzung

Bayesianer basieren Schätzungen von unbekannten Grössen auf der sogenannten **a posteriori Verteiung** der unbekannten Grössen gegeben die bekannten Grössen. Die a posteriori Verteilung wird mithilfe des **Satzes von Bayes** aufgrund der a priori Verteilung der unbekannten und aufgrund der Likelihood berechnet.

Die Bezeichnungen "a priori" und "a posteriori" beziehen sich immer auf den Zeitpunkt der Beobachtung der analysierten Daten. Die jeweiligen Verteilungen quantifizieren den Informationsstand zu den Unbekannten um jeweiligen Zeitpunkt. Dieses Konzept soll anhand der folgenden Grafik verdeutlicht werden.



Für unser Beispiel des einfachen linearen Modells, definieren wir zuerst den Vektor β als

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Die Beobachtungen y_i fassen wir ebenfalls in einem Vektor $\mathbf{y}$ zusammen. Die a posteriori Verteilung $f(\beta, \sigma^2|\mathbf{y})$ der Unbekannten β und σ^2 gegeben die Bekannten $\mathbf{y}$ lässt sich nun wie folgt berechnen

$$\begin{aligned} f(\beta, \sigma^2|\mathbf{y}) &= \frac{f(\beta, \sigma^2, \mathbf{y})}{f(\mathbf{y})} \\ &= \frac{f(\mathbf{y}|\beta, \sigma^2) f(\beta) f(\sigma^2)}{f(\mathbf{y})} \end{aligned} \tag{2}$$

In Gleichung (2) konnten wir die a posteriori Verteilung $f(\beta, \sigma^2|\mathbf{y})$ als Produkt der a priori Verteilungen ($f(\beta)$ und $f(\sigma^2)$) der unbekannten Grössen β und σ^2 und der Likelihood $f(\mathbf{y}|\beta, \sigma^2)$ ausdrücken. Der Faktor $f(\mathbf{y})^{-1}$ (Term im Nenner) entspricht der sogenannten Normalisierungskonstanten und ist nicht weiter von Interesse.

Kontrollfrage 3: Aus welchen Bestandteilen berechnen wir die a posteriori Verteiung der unbekannten Grössen gegeben die bekannten Grössen, wenn wir den Satz von Bayes anwenden?

Die a posteriori Verteilung $f(\beta, \sigma^2|\mathbf{y})$ ist in vielen Fällen nicht explizit darstellbar. Das war lange ein Problem, welches die Anwendung von Bayes'schen Analysen sehr einschränkte. Zwei Entwicklungen haben dieses Problem beseitigt.

1. In seinem Paper (Besag (1974)) zeigte Julian Besag, dass jede posteriori Verteilung durch eine Serie von Zufallszahlen aus den voll-bedingten Verteilungen bestimmt ist. Für unser Beispiel lauten die voll-bedingten Verteilungen: Bedingte Verteilung von β_0 gegeben alle anderen Grössen: $f(\beta_0|\beta_1, \sigma^2, \mathbf{y})$, bedingte Verteilung von β_1 gegeben alle anderen Grössen: $f(\beta_1|\beta_0, \sigma^2, \mathbf{y})$ und bedingte Verteilung von σ^2 gegeben alle anderen Grössen: $f(\sigma^2|\beta_0, \beta_1, \mathbf{y})$ (mehr Details dazu in einem späteren Abschnitt).
2. Die Entwicklung von effizienten Pseudo-Zufallszahlen-Generatoren auf dem Computer

Gibbs Sampler

Die Umsetzung der beiden oben aufgelisteten Punkte führt zu einer Prozedur, welche als **Gibbs Sampler** bezeichnet wird. Wenden wir den Gibbs Sampler auf einfaches lineares Regressionsmodell an, dann resultiert das folgende Vorgehen bei der Analyse. Unabhängig vom verwendeten Modell läuft die Konstruktion einer Gibbs Sampling Prozedur immer in den folgenden Schritten ab. Diese Schritte können für die meisten Analysen wie ein Kochbuchrezept verwendet werden.

1. Bestimmung der a priori Verteilungen für die unbekannten Grössen.
2. Bestimmung der Likelihood
3. Bestimmung der voll-bedingten Verteilungen

A priori Verteilungen

In unserem Bespiel handelt es sich dabei um $f(\beta)$ und $f(\sigma^2)$. In den meisten Fällen, wenn man das erste Mal eine bestimmte Art von Daten analysisern soll, empfielt es sich eine sogenannte uniformative a priori Verteilung zu wählen. Eine uninformative a priori Verteilung bedeutet einfach, dass deren Dichtewert überall gleich, also eine Konstante ist. Wenden wir zum Beispiel für die Unbekannte β eine uninformative a priori Verteilung an, dann bedeutet das, dass wir $f(\beta) = c$.

Alternativ zu der uniformativen a priori Verteilung gibt es auch a priori Verteiungen für bestimmte unbekannte Grössen, welche als de-facto Standard aktzeptiert sind. Ein Bespiel dafür ist die a priori Verteilung der unbekannten Restvarianz, welche üblicherweise als Inverse-Chi-Quadrat Verteilung angenommen wird.

Kontrollfrage 4: Welche Möglichkeiten zur Bestimmung einer a priori Verteilung gibt es?

Likelihood

Die Likelihood ist wie bei den Frequentisten als begingte Verteilung ($f(\mathbf{y}|\beta, \sigma^2)$) der Daten $\mathbf{y}$ gegeben die Parameter (β und σ^2). Falls keine Daten fehlen, dann ist die Bayes'sche Likelihood und die frequentistische Likelihood gleich.

Kontrollfrage 5: Welche bedingte Verteilung wird sowohl bei den Frequentisten als auch bei den Bayesianern als Likelihood bezeichnet?

Vollbedingte Verteilungen

Mit vollbedingten Verteilungen ist gemeint, dass für jede unbekannte Grösse die bedingte Verteilung gegeben alle anderen Grössen bestimmt wird. In unserem Bespiel des linearen Regressionsmodells haben wir drei unbekannte Grössen β_0, β_1 und σ^2. Somit haben wir auch drei vollbedingte Verteilungen

unbekannte Grösse	vollbedingte Verteilung	resultierende Verteilung
β_0	$f(\beta_0 \vert \beta_1, \sigma^2, \mathbf{y})$	$\mathcal{N}(\hat{\beta}_0, var(\hat{\beta}_0))$
β_1	$f(\beta_1 \vert \beta_0, \sigma^2, \mathbf{y})$	$\mathcal{N}(\hat{\beta}_1, var(\hat{\beta}_1))$
σ^2	$f(\sigma^2 \vert \beta_0, \beta_1, \mathbf{y})$	$\propto \chi^{-2}$

Kontrollfrage 6: Was versteht man unter vollbedingter Verteilung?

Aufgrund von Berechnungen, welche hier nicht gezeigt sind, können wir die oben aufgelisteten vollbedingten Verteilungen bestimmen. Die entsprechenden Verteilungen sind in der Kolonnen ganz rechts, welche mit "resultierende Verteilung" überschrieben ist, aufgelistet. Dabei steht $\mathcal{N}()$ für die Normalverteilung. Für die Erwartungswerte und Varianzen wird das Modell in Gleichung (1) leicht umformuliert.

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{x}\beta_1 + \epsilon \tag{3}$$

Aus dem obigen Modell bilden wir ein neues Modell, welches auf der rechten Seite der Gleichung nur von β_0 und ϵ abhängt. Da wir wissen, dass die Verteilung der Least Squares Schätzer eine Normalverteilung ist, werden wir diese für die Bestimmung der vollbedingten Verteilungen verwenden.

$$\mathbf{w}_0 = \mathbf{1}\beta_0 + \epsilon \tag{4}$$

wobei $\mathbf{w}_0 = \mathbf{y} - \mathbf{x}\beta_1$. Aufgrund des Modells in Gleichung (4) können wir den Least Squares Schätzer für β_0 aufstellen. Dieser lautet:

$$\hat{\beta}_0 = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{w}_0 \tag{5}$$

Die Varianz des Least Squares Schätzers für β_0 lautet:

$$var(\hat{\beta}_0) = (\mathbf{1}^T\mathbf{1})^{-1}\sigma^2 \tag{6}$$

Analog dazu berechnen wir den Least Squares Schätzer für β_0 und dessen Varianz.

$$\hat{\beta}_1 = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{w}_1 \tag{7}$$

wobei $\mathbf{w}_1 = \mathbf{y} - \mathbf{1}\beta_0$

$$var(\hat{\beta}_1) = (\mathbf{x}^T\mathbf{x})^{-1}\sigma^2 \tag{8}$$

Die resultiertende vollbedingte Verteilung von σ^2 wurde in obiger Tabelle mit $\propto \chi^{-2}$ angegeben. Das heisst, wir können diese Verteilung nur bis auf einen Proportionalitätsfaktor angeben und die Verteilung ist proportional zu einer inversen χ^2 Verteilung.

Umsetzung des Gibbs Samplers

Der Gibbs Sampler wird durch wiederholtes ziehen von Zufallszahlen aus den oben angegebenen vollbedingten Verteilungen umgesetzt. Das heisst, wir setzen für alle unbekannten Grössen sinnvolle Startwerte ein. Für β_0 und β_1 wählen wir 0 als Startwert und für σ^2 wählen wir die empirische Varianz von $\mathbf{y}$ als Startwert. Dann berechnen wir den Erwartungswert und die Varianz für die vollbedingte Verteilung von β_0. Aus dieser Verteilung ziehen wir einen neuen Wert für β_0. In einem zweiten Schritt berechnen wir den Erwartungswert und die Varianz für die vollbedingte Verteilung von β_1, wobei wir für β_0 schon den neuen Wert einsetzen. Aus der Verteilung für β_1 ziehen wir einen neuen Wert für β_1. Im dritten Schritt verfahren wir analog für σ^2. Danach beginnen wir die Schritte wieder bei β_0. Diese Schrittabfolge wiederholen wir 10000 mal und speichern alle gezogenen Werte für β_0, β_1 und σ^2. Die Bayes'schen Parameterschätzungen entsprechen dann den Mittelwerten der gespeicherten Werte.

Der folgende R-Codeblock soll die Umsetzung des Gibbs Samplers für β_0 und β_1 als Programm zeigen. Der Einfachheit halber wurde σ^2 konstant $\sigma^2 = 1$ angenommen.

```
# ### Startwerte für beta0 und beta1
beta <- c(0, 0)
# ### Bestimmung der Anzahl Iterationen
niter <- 10000
# ### Initialisierung des Vektors mit Resultaten
meanBeta <- c(0, 0)
for (iter in 1:niter) {
  # Ziehung des Wertes des Achsenabschnitts beta0
  w <- y - X[, 2] * beta[2]
  x <- X[, 1]
  xpxi <- 1/(t(x) %*% x)
  betaHat <- t(x) %*% w * xpxi
  # ### neue Zufallszahl fuer beta0
  beta[1] <- rnorm(1, betaHat, sqrt(xpxi))
  # Ziehung der Steigung beta1
  w <- y - X[, 1] * beta[1]
  x <- X[, 2]
  xpxi <- 1/(t(x) %*% x)
  betaHat <- t(x) %*% w * xpxi
  # ### neue Zufallszahl fuer beta1
  beta[2] <- rnorm(1, betaHat, sqrt(xpxi))
  meanBeta <- meanBeta + beta
}
# ### Ausgabe der Ergebnisse
cat(sprintf("Achsenabschnitt = %6.3f \n", meanBeta[1]/iter))
cat(sprintf("Steigung = %6.3f \n", meanBeta[2]/iter))
```

Antworten zu den Kontrollfragen

Antwort 1:

Bayesianer unterscheiden zwischen bekannten und unbekannten Grössen unabhängig, ob das Parameter oder Daten sind.

Antwort 2:

Die Einteilung in bekannte und unbekannte Grössen entspricht unter der Annahme, dass keine Daten fehlen, der frequentistischen Einteilung in Daten und Parameter. Dabei entsprechen die Daten den bekannten Grössen und die Parameter den unbekannten Grössen.

Antwort 3:

Die Bestandteile der a posteriori Verteiung der unbekannten Grössen gegeben die bekannten Grössen lauten

- a priori Verteilungen der unbekannten Grössen
- Likelihood
- Normalisierungskonstante

Antwort 4:

Wir haben zwei Möglichkeiten angeschaut.

1. uninformative a priori Verteilungen, d.h. die Dichte wird konstant angenommen, z.B. $f(\beta) = c$
2. für gewisse unbekannte Grössen gibt es de-facto Standards. Zum Beispiel für $f(\sigma^2)$ wird häufig eine inverse χ^2 Verteilung verwendet.

Antwort 5:

Die Likelihood entspricht der bedingten Verteilung der Daten gegeben die Parameter. In unserem Beispiel des einfachen linearen Modells war das $f(\mathbf{y}|\beta, \sigma^2)$.

Antwort 6:

Die vollbedingte Verteilung einer bestimmten unbekannten Grösse ist die bedingte Verteilung dieser unbekannten Grösse gegeben alle anderen Grössen. Zum Beispiel ist die vollbedingte Verteilung für β_0 gleich der bedingten Verteilung von β_0 gegeben alle die anderen Grössen also: $f(\beta_0|\beta_1, \sigma^2)$.

References

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological),* 36: 192–236. http://www.jstor.org/stable/2984812.