Seminar über Statistik FS2008: Bayesian Statistics

Hierarchischer und Empirischer Bayes

Bruno Catarino
Rudolf Dünki

(Seminar vom 28.04.2008)

2. Mai 2008

1 Einführung

Zur Einführung ins Thema der empirischen und hierarchischen Bayes-Methoden wollen wir das folgende Beispiel betrachten [GCSR03]:
Gegeben sei ein experimentelles Design, in dem Beobachtungen aus mehreren Gruppen stammen, aber innerhalb der Gruppen unabhängig normalverteilt sind. Jede Gruppe ist durch einen eigenen Gruppenmittelwert charakterisiert und diese Gruppenmittelwerte seien ihrerseits auch unabhängig verteilt. Es seien nun N unabhängige Experimente ausgeführt worden, aus denen je n unabhängige Datensätze gewonnen worden sind. Jeder dieser Datensätze kann nun beschrieben werden durch

$$y_{ij} \sim N(\theta_j, \sigma^2) \tag{1}$$

wobei i = 1 ... n und j = 1 ... N. Für die Gruppenmittelwerte gilt offenbar

$$\bar{y}_{.j} = 1/n \sum_{i=1}^{n} y_{ij} \tag{2}$$

und für deren Varianz

$$\sigma_j^2 = 1/n \, \sigma^2 \;\; \forall j \tag{3}$$

In klassischen Bayes-Verfahren würde man eine a priori Verteilungen für die θ_j formulieren ($\theta \sim N(\mu, \tau^2)$) und anschliessend die θ_j aufgrund der a posteriori Verteilung schätzen. Was sind aber in diesem Beispiel plausible a posteriori Schätzer? Zwei Fälle sind evident:
1. $\hat{\theta}_j = \bar{y}_{.j}$, d.h einfach den Durchschnittswert des Experiments j. Die zugehörige a priori Verteilungen für die N Werte sind unabhängig und uniform auf (-∞,∞). Falls N eher gross und n eher klein ist, sind diese Werte nicht besonders glaubwürdig.
2. $\hat{\theta}_j = \bar{y}_{..} = \sum \bar{y}_{.j}/N$, d.h. der Gesamtdurchschschnitt. Die zugehörige a priori Verteilung ist uniform, aber es gilt $\theta_j = \theta \, \forall j$. Dieser Wert ist glaubwürdiger, aber negiert Unterschiede zwischen den Gruppen.
Ohne Spezifikation einer a priori Verteilung kann man das Problem frequentistisch mit varianzanalytischen Verfahren angehen. Dabei wird der Unterschied zwischen den Gruppen als sog. Zufallseffekt angesehen. Man würde die empirisch gefundenen Quadratsummen SS ('sums of squares') und mittleren

	df	SS	MS	$E(MS\|\sigma^2, \tau)$
Between groups	$N-1$	$\sum_i \sum_j (\overline{y}_{.j} - \overline{y}_{..})^2$	$SS/(N-1)$	$n\tau^2 + \sigma^2$
Within groups	$N(n-1)$	$\sum_i \sum_j (y_{ij} - \overline{y}_{.j})^2$	$SS/(N(n-1))$	σ^2
Total	$Nn-1$	$\sum_i \sum_j (y_{ij} - \overline{y}_{..})^2$	$SS/(Nn-1)$	

Abbildung 1:
Quadratsummen, mittlere Quadratsummen und erwartete mittlere Quadratsummen des Einführungsbeispiels (aus [GCSR03]).

Quadratsummen MS ('mean squares') sowohl innerhalb der Gruppen ('within groups') als auch zwischen den Gruppen ('between groups') ausrechnen. Ergibt ein F-Test, dass das Verhältnis der MS signifikant grösser als 1 ist, dann akzeptiert man Fall 1 als wahr. Andernfalls akzeptiert man Fall 2 (d.h. die Nullhypothese entspricht Fall 2).
Man kann also nur eine entweder-oder Entscheidung treffen. Als Alternative hiezu bietet sich die gewichtete Kombination an (Fall 3):

$$\hat{\theta}_j = \lambda \bar{y}_{.j} + (1-\lambda)\bar{y}_{..} \tag{4}$$

Diese hat die Form eines Bayes'schen a posteriori Schätzers, wobei die θ_j a priori iid verteilt sind. In obiger Formel taucht allerdings nicht der a priori Erwartungswert auf, sondern stattdessen das empirische Gesamtmittel $\bar{y}_{..}$. Dies ist eine Eigenheit der empirischen Bayes-Statistik: Die a priori Spezifika werden auch aus den Daten geschätzt.
Formal ist Fall 3 beschrieben durch

$$f(\theta_1, \ldots \theta_n | \alpha, \beta) = \Pi_{j=1}^{N} N(\alpha, \beta)|_{\theta_j} \tag{5}$$

und

$$f(\theta_1, \ldots \theta_n) = \int \Pi_{j=1}^{N} N(\alpha, \beta)|_{\theta_j} f(\alpha, \beta) d(\alpha, \beta) \tag{6}$$

D.h. die θ_j sind unabhängig, wenn (α, β) vorgegeben ist. Sie entstammen aber ihrerseits einer übergeordneten gemeinsamen Verteilung. Wir haben es also mit einer hierarchischen Struktur zu tun, wie in der Grafik illustriert.

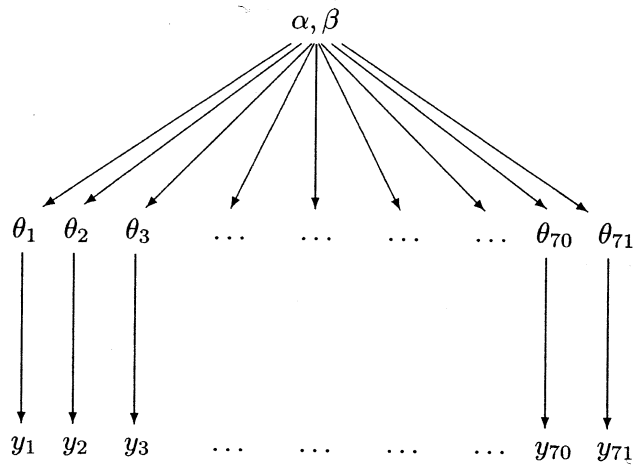

Figure 5.1 *Structure of the hierarchical model for the rat tumor example.*

Abbildung 2: Schema eines hierarchischen Bayes-Modells (aus [GCSR03]).

2 Hierarchische Bayesstatistik

2.1 Definitionen

Definition 1 (Hyperparameter)**.** Falls die A-priori-Verteilung bis auf einen Parameter τ bestimmt ist, also $\pi(\theta) = \pi(\theta, \tau)$ gilt, dann nennt man τ einen Hyperparameter.

Definition 2 (Hierarchisches Bayes'sches Modell)**.** Ein Hierarchisches Bayes'sches Modell ist ein Bayes'sches statistisches Modell $(f(x|\theta), \pi(\theta))$, dessen A-priori-Verteilung $\pi(\theta)$ in bedingte Verteilungen $\pi_1(\theta|\theta_1), \pi_2(\theta_1|\theta_2), \ldots, \pi_n(\theta_{n-1}|\theta_n)$ sowie in die Randverteilung $\pi_{n+1}(\theta_n)$ zerlegt werden kann, so dass

$$\pi(\theta) = \int_{\Theta_1 \times \cdots \times \Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2) \cdots \pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n) d\theta_1 \cdots d\theta_n$$

erfüllt ist. Die Parameter θ_i für $i = 1, \ldots, n$ nennt man die Hyperparameter der Stufe i.

Man beachte, dass auch nicht-Bayes'sche Modelle diese hierarchische Struktur vorweisen können.

In der Praxis braucht man selten über mehr als zwei Stufen zu gehen. D.h. man betrachtet

$$x|\theta \sim f(x|\theta), \qquad \theta|\theta_1 \sim \pi_1(\theta|\theta_1)$$

mit

$$\theta_1 \sim \pi_2(\theta_1) = \int_{\Theta_2 \times \cdots \times \Theta_n} \pi_2(\theta_1|\theta_2) \cdots \pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n) d\theta_2 \cdots d\theta_n.$$

Meistens handelt es sich bei der Verteilung erster Stufe π_1 um einen konjugierten Prior, dessen Parameter θ_1 die Verteilung π_2 hat.

2.2 Vorteile des hierarchischen Modells

Gründe, weshalb man hierarchische Modelle benutzt:

- Subjektivität vs. Objektivität: Das hierarchische Modell erlaubt uns das Modellieren der A-priori-Verteilung in einen Teil mit strukturellen Informationen (objektive Kriterien) und einen zweiten Teil mit subjektiven Informationen zu zerlegen.
- Das hierarchische Modell vereinfacht Simulationen (siehe Gibbs-sampling, evt. im nächsten Vortrag).
- Im noninformativen Fall kann man einen Kompromiss zwischen Jeffreys noninformativem Prior (als Prior zweiter Stufe) und dem konjugierten Prior (als Prior erster Stufe) eingehen.
- Das hierarchische Modell ist robuster als das nicht-hierarchische Bayes'sche Modell.
- Das hierarchische Modell bietet sich als natürliches Modell für Daten hierarchischer Struktur an. Dies ist beispielsweise der Fall, wenn man ein Experiment betrachtet, das ein Spezialfall eines allgemeineren Experimentes ist (siehe Meta-Analysis/Meta-Populationen).

In gewissen Situationen stehen Daten aus $N+1$ unabhängigen Versuchen zur Verfügung und der Modellparameter β variiert von Versuch zu Versuch,

ist also zufällig im frequentistischen Sinne. Die wahren Parameterwerte der $N+1$ Versuche seien $\beta_1, \ldots, \beta_{N+1}$. Wir fassen die β_i als iid Realisierungen einer Zufallsvariable mit Verteilung $\pi(\beta_i|\lambda)$ auf. Gegeben $\beta_1, \ldots, \beta_{N+1}$ seien dann die Daten $x_1, \ldots, x_{N+1}$ iid verteilt mit $x_i \sim f(x_i|\beta_i)$. Bei einem zweistufigen hierarchischen Modell wird für den Hyperparameter λ ein "Hyperprior" $\pi(\lambda)$ verwendet. Die soeben beschriebene Situation liegt im folgenden Beispiel vor:

Beispiel 1. *Wir betrachten Daten mit hierarchischer Struktur. Wir begleiten ein Kind sieben Jahre lang und führen jährlich einen IQ-Test durch. D.h. wir haben $x_i \sim \mathrm{N}(\beta_i, 10)$ jährliche und unabhängige Messungen des IQs für $i = 1, \ldots, 7$. Wir wollen nun die β_is bestimmen. Die IQ-Tests seien an das Alter des Kindes angepasst, d.h. man sollte theoretisch immer denselben Wert erhalten. Wir gehen also davon aus, dass die β_is den gleichen Erwartungswert λ haben, den "wahren" IQ-Wert.*

Die erste Stufe der a priori Verteilung wäre dann

$$\beta_i|\lambda \sim \mathrm{N}(\lambda, \sigma_\pi^2) \qquad (i = 1, \cdots, 7).$$

Nun wissen wir aber, dass das Kind zu einer (detailliert untersuchten) Klasse von Kindern gehört, deren IQ $\lambda \sim \mathrm{N}(\eta, \tau^2)$ ist, für bekannte η und τ. Diese Normalverteilung verwenden wir dann als die zweite Stufe der A-priori-Verteilung. Die noninformative Alternative dazu wäre $\pi_2(\lambda) = 1$ zu setzen.

In der empirischen Bayesstatistik würde man für den Hyperparameter λ keinen Prior spezifizieren, sondern λ direkt aus der Randverteilung $m(\cdot|\lambda)$ der Daten $x_1, \ldots, x_N$ schätzen:

$$m(x_1, \ldots, x_N|\lambda) = \prod_{i=1}^{N} \int f(x_i|\beta_i)\pi(\beta_i|\lambda)d\beta_i.$$

Dieses geschätzte $\hat{\lambda}$ würde man dann als a priori bekannt voraussetzen und den Posterior $\pi(\beta_{N+1}|x_{N+1})$ berechnen.

2.3 Bedingte Zerlegungen

Wir erinnern uns daran, dass die A-posteriori-Verteilung gegeben ist durch:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

Das folgende Lemma ist das Hauptrechenwerkzeug dieses Abschnittes.

Lemma 1. Sei $\theta|\theta_1 \sim \pi_1(\theta|\theta_1)$ mit $\theta_1 \sim \pi_2(\theta_1)$. Dann kann man die A-posteriori-Verteilung wie folgt zerlegen:

$$\pi(\theta|x) = \int_{\Theta_1} \pi(\theta|\theta_1, x)\pi(\theta_1|x)d\theta_1,$$

wobei

$$\pi(\theta|\theta_1, x) := \frac{f(x|\theta)\pi_1(\theta|\theta_1)}{m_1(x|\theta_1)},$$

$$m_1(x|\theta_1) := \int_\Theta f(x|\theta)\pi_1(\theta|\theta_1)d\theta,$$

$$\pi(\theta_1|x) := \frac{m_1(x|\theta_1)\pi_2(\theta_1)}{m(x)},$$

$$m(x) := \int_{\Theta_1} m_1(x|\theta_1)\pi_2(\theta_1)d\theta_1.$$

Weiter kann man diese Zerlegung auch auf die Momente der A-posteriori-Verteilung anwenden, d.h. für alle Funktionen h gilt

$$E^\pi[h(\theta)|x] = E^{\pi(\theta_1|x)}[E^{\pi_1}[h(\theta)|\theta_1, x]],$$

wobei

$$E^{\pi_1}[h(\theta)|\theta_1, x] := \int_\Theta h(\theta)\pi(\theta|\theta_1, x)d\theta.$$

Dieses Lemma folgt unmittelbar aus dem Bayes-Theorem und dem Satz von Fubini. Weiter hat es wichtige Konsequenzen für das Berechnen von Bayesschätzern. Es zeigt, dass man $\pi(\theta|x)$ simulieren kann, indem man zuerst θ_1 aus $\pi(\theta_1|x)$ und dann θ aus $\pi(\theta|\theta_1, x)$ erzeugt, falls die zwei bedingten Verteilungen einfach handhabbar sind.

Dieses Lemma gilt natürlich nur dann, wenn die verschiedenen Integrale auch definiert sind. Dies ist aber im Allgemeinen nicht so. Oft gilt $\int_{\Theta_1} \pi_2(\theta_1)d\theta_1 = \infty$. Folgendes Lemma gibt eine hinreichende Bedingung an dafür, dass die A-posteriori-Momente existieren.

Lemma 2. Sei $x|\theta \sim \mathrm{N}_d(\theta, \Sigma)$. Ist die Randverteilung

$$m(x) := \int_\Theta f(x|\theta)\pi(\theta)d\theta$$

endlich für alle $x \in \mathbb{R}^d$, dann existieren der Erwartungswert und die Varianz bezüglich der A-posteriori-Verteilung $\pi(\theta|x)$ immer.

Ein weiteres wichtiges Lemma für die Handhabung von Simulationen ist folgendes Lemma:

Lemma 3. Sei $\pi(\theta) = \int_{\Theta_1 \times \cdots \times \Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2) \cdots \pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n)d\theta_1 \cdots d\theta_n$ ein hierarchisches Modell. Dann gilt für die Verteilung von θ_i bedingt auf x und θ_j's für $j \neq i$

$$\pi(\theta_i|x, \theta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n) = \pi(\theta_i|\theta_{i-1}, \theta_{i+1}),$$

wobei $\theta_0 = \theta$ und $\theta_{n+1} = 0$.

Das Lemma sagt uns also, dass die "full conditionals" im hierarchischen Modell nur von den zwei lokalen Hyperparametern abhängen. Dies gibt uns einen Wink, dass hier Gibbs-sampler für die numerischen Berechnungen angebracht sind. Mehr dazu gibt es im nächsten Vortrag über Simulationen.

2.4 Nachteil des hierarchischen Modells

Bayesschätzer können meistens nicht explizit berechnet werden. Die Lemmas aus dem letzten Abschnitt erlauben uns aber Bayesschätzer numerisch zu approximieren, via Gibbs-sampler oder MCMC-Techniken.

3 Empirische Bayes Statistik

Die Grundidee der empirischen Bayesstatistik ist diese: Man schätzt die Hyperparameter einer bestimmten Stufe des hierarchischen Modells aus der Randverteilung der Daten und tut dann so, als wären diese Parameter a priori bekannt.

3.1 Rekap und Prärequisiten

Zum Anfang seien hier nochmals früher ausführlich vorgestellte Begriffe rekapituliert, auf die auch in der empirischen Bayes Statistik eine Bezug genommen wird. Im folgenden bezeichnet $f(x|\theta)$ die Dichte einer Zufallsvariablen X bedingt auf θ, $\pi(\theta)$ die a priori Verteilung von θ und $d(x)$ einen Schätzer für θ. Die folgenden Grössen treten auf:

- Der Verlust $L(\theta, d(x)) \geq 0$. Vielfach wird quadratischer Verlust angenommen:
$$L(\theta, d(x)) = (\theta - d(x))^2 \tag{7}$$
- Das Risiko R_d
$$R_d(\theta) = \int L(\theta, d(x)) f(x|\theta) dx \tag{8}$$
- $W_d(\pi)$, das erwartete Risiko bezüglich der a priori Verteilung $\pi(\theta)$, auch genannt Bayes-Risiko:
$$W_d(\pi) = \int \int L(\theta, d(x)) f(x|\theta) dx d\pi(\theta) \tag{9}$$
- Sei $d_\pi(x)$ dasjenige $d(x)$, das das erwartete Risiko minimiert. Unter quadratischem Verlust gilt
$$d_\pi(x) = \frac{\int \theta f(x|\theta) d\pi(\theta)}{\int f(x|\theta) d\pi(\theta)} \tag{10}$$
wobei der Nenner die marginale Dichte $m(x)$ darstellt:
$$m(x) = \int f(x|\theta) d\pi(\theta) \tag{11}$$
- Für den Erwartungswert und die Varianz der Marginalverteilung kann man schreiben
$$E_m(x) = E(E(x|\theta)) \tag{12}$$
(Satz vom iterierten Erwartungswert) und
$$var_m(x) = E(var(x|\theta)) + var(E(x|\theta)) \tag{13}$$
(Varianzzerlegungssatz)

3.2 Prinzip der empirischen Bayes Methode

3.2.1 Nichtparametrischer Fall

Wir betrachten wiederum $(n+1)$ unabängige Beobachtungen $x_1, \ldots, x_{n+1}$ mit zugehörigen bedingten Dichten $f(x_i|\theta_i)$. Das Problem ist nun einen Bayes Schätzer für θ_{n+1} zu finden, unter der Annahme, dass die θ_i iid sind mit $\theta_i \sim \pi(\theta_i)$, wobei π unbekannt ist. Das bedeutet, dass die Stichprobenverteilung bekannt ist, nicht aber die a priori Verteilung. Die Randverteilung $m(x)$ kann - quasi in Umkehr von (eq. 11) - benutzt werden, um die Dichte der Verteilung π zu schätzen, da die (bekannten) x_i als unabhängige Stichproben aus $m(x)$ angesehen werden können. Sei nun $\hat{\pi}(\theta)$ die geschätzte Dichte. Dann ergibt sich die empirische a posteriori Verteilung als

$$\tilde{\pi}(\theta_{n+1}|x_{n+1}) \propto f(x_{n+1}|\theta_{n+1})\hat{\pi}(\theta_{n+1}) \tag{14}$$

Beispiel 2. *Seien $x_i \sim Pois(\theta_i)$ und p_k die Anzahl Beobachtungen mit Wert k. Letzteres identifizieren wir mit $\hat{m}(k)$, der Schätzung der Randverteilung $m(k)$:*

$$p_k \approx m(k) = \int \frac{e^{-\theta}\theta^k}{k!}\pi(\theta)d\theta \tag{15}$$

Unter quadratischem Verlust ergibt sich $d_\pi(x)$ zu

$$d_\pi(x_{n+1}) = \frac{\int e^{-\theta}\theta^{x_{n+1}+1}\pi(\theta)d\theta}{\int e^{-\theta}\theta^{x_{n+1}}\pi(\theta)d\theta} \tag{16}$$

Dies ist aber gerade gleich

$$\frac{m(x_{n+1}+1)}{m(x_{n+1})}(x_{n+1}+1) \simeq \frac{p_{x_{n+1}+1}}{p_{x_{n+1}}+1}(x_{n+1}+1) \tag{17}$$

Der Ausdruck rechts ist der empirische Bayes-Schätzer. Das zusätzliche + 1 im Nenner kommt daher, dass p_{x_i} nach n Versuchen ausgewertet wird und die Beobachtung x_{n+1} in die Klasse $\{x_{n+1}\}$ fällt [MarLwi89]. In diesem Beispiel kann also auf die Schätzung $\hat{\pi}(\theta)$ verzichtet werden, da die zugrundeliegende Poissonverteilung das Problem wieder auf die Randverteilung zurückführt. Dies ist im allgemeinen aber nicht der Fall.

3.2.2 Parametrischer Fall

Wir nehmen wieder an, dass $\theta_1, \ldots, \theta_{n+1}$ iid sind. Der parametrische Fall zeichnet sich dadurch aus, dass zwar die Form der a priori Verteilung $\pi(\theta|\lambda)$ gegeben ist, nicht aber deren Parameterwerte. Für exponentiellen Familien ist dies z.B. die Konjugierte zur Likelihood. Die Randverteilung von x hängt vom unbekannten Parameter λ ab:

$$m(x|\lambda) = \int f(x|\theta)\pi(\theta|\lambda)d\theta. \tag{18}$$

Für $n+1$ unabhängige Beobachtungen $x_i \sim f(x_i|\theta_i)$ erhält man:

$$m(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^{n} \int f(x_i|\theta_i)\pi(\theta_i|\lambda)d\theta_i.$$

Man versucht λ durch $\hat{\lambda}(x_1, \ldots, x_n)$ aus den Daten $x_1, \ldots, x_n$ zu schätzen, zum Beispiel via Maximum Likelihood. Für die a posteriori Verteilung gilt dann

$$\pi(\theta_{n+1}|x_{n+1}) \approx \tilde{\pi}(\theta_{n+1}|x_{n+1}) \propto f(x_{n+1}|\theta_{n+1})\pi(\theta_{n+1}|\hat{\lambda}(x_1, \ldots, x_n)). \tag{19}$$

Im Fall von exponentiellen Familien und quadratischem Verlust gibt uns das folgende Lemma eine Rechtfertigung für dieses Vorgehen [Rob07]:

Lemma 4. Sei $x \sim f(x|\theta) = e^{\theta x - \psi(\theta)}h(x), x \in R^k$.
Wenn θ verteilt ist gemäss $\pi(\theta|\lambda), \lambda \in R^p$ und $\hat{\lambda}(x)$ die Lösung der zu $m(x|\lambda)$ gehörigen Likelihood-Gleichungen ist, dann gilt für den empirischen Bayes-Schätzer

$$\begin{aligned} d^{EB}(x) &= (\nabla_x \log m(x|\lambda))|_{\lambda=\hat{\lambda}(x)} - \nabla_x \log(h(x) \\ &= \nabla_x \log m(x|\hat{\lambda}(x)) - \nabla_x \log(h(x) \end{aligned}$$

Beweis: Zuerst einmal ist λ einfach ein unabhängiger Parameter. Damit gilt unter quadratischem Verlust

$$\begin{aligned} d(x|\lambda) &= 1/m(x|\lambda) \int \theta e^{\theta x - \psi(\theta)}h(x)\pi(\theta|\lambda)d\theta \\ &= 1/m(x|\lambda)\ \nabla_x(\int e^{\theta x - \psi(\theta)}\pi(\theta|\lambda)) - \nabla_x h(x)/h(x) \\ &= \nabla_x \log m(x|\lambda) - \nabla_x \log(h(x) \end{aligned}$$

Der zweite Teil des Lemmas ergibt sich mit der Kettenregel:

$\nabla_x \log m(x|\hat{\lambda}(x)) = \nabla_x \log m(x|\lambda))|_{\lambda=\hat{\lambda}(x)} + \nabla_\lambda m(x|\lambda)|_{\lambda=\hat{\lambda}(x)} \nabla_x \hat{\lambda}(x)$

Wegen der Maximum-Likelihood-Annahme ist der Term in ∇_λ aber gerade gleich 0.

Beispiel 3. *Wir betrachten wiederum den Fall $x_i \sim P(\theta_i)$. Jetzt setzen wir aber $\pi(\theta|\lambda) \sim Exp(\lambda)$. In dem Fall gilt*

$$m(x_i|\lambda) = \int_0^\infty f(x|\theta)\lambda e^{-\theta\lambda} d\theta = \frac{\lambda}{(\lambda+1)^{x_i+1}} \tag{20}$$

Also sind die $(x_i|\lambda) \sim Geo(\frac{\lambda}{\lambda+1})$ und $\hat{\lambda}_{ML} = 1/\bar{x}$, das Inverse des Durchschnitts von $x_1, \ldots, x_n$. Der empirische Bayes-Schätzer errechnet sich damit zu

$$d^{EB}(x_{n+1}) = (x_{n+1}+1)\frac{\bar{x}}{\bar{x}+1} \tag{21}$$

Bemerkung: Manchmal werden statt ML-Schätzer die Momentenschätzer verwendet um die Hyperparameter zu bestimmen, da man ML-Schätzer oft nicht geschlossen berechnen kann.

3.2.3 Lineare Bayes Schätzer

In der empirischen Bayes Statistik ist die Verteilung von (θ, x) nicht immer bekannt. Das erwartete Risiko kann somit i.a. nicht exakt bestimmt werden. Die Situation lässt sich u.U. ändern, wenn man sich von vornherein auf bestimmte Klassen von Schätzern beschränkt. Eine einfache Form von Bayes-Schätzern ergibt sich, wenn man lineare Bayes-Schätzer, also Schätzer der Klasse $d(x) = a + bx$ betrachtet und das erwartete Risiko $E(L(\theta, a+bx))$ wird durch Variation von $\{a, b\}$ minimiert. Unter quadratischem Verlust und der Annahme $E(x|\theta) = \theta$ ergibt sich damit [Grize03]

$$d(x) = var(\theta)/(var_m(x))x + (1 - var(\theta)/(var_m(x))E(\theta) \tag{22}$$

Hierbei bezeichnet $\text{var}_m(x)$ die Varianz der Marginalverteilung. Der lineare Bayes-Schätzer ist somit ein gewichtetes Mittel aus dem Datenwert x und dem Erwartungswert der a priori Verteilung. Für den empirischen Bayes Ansatz ist dies vorteilhaft, da nun nur noch Erwartungswerte und Varianzen aus

den Daten geschätzt werden müssen. Allerdings entspricht $d(x)$ nicht notwendigerweise $d_\pi(x)$, dem tatsächlichen Bayes-Schätzer. Der lineare Bayes Schätzer führt also nicht notwendigerweise zum richtigen Wert. Dafür vereinfacht sich die Parameterschätzung.

Beispiel 4. *Im Einführungsbeispiel stellte sich die Frage nach den 'richtigen' Klassenmittelwerten $y_{.j}$. Um dies mit Hilfe des linearen Bayes-Schätzers zu beantworten, gebrauchen wir die empirischen Werte*

$$\begin{array}{rcccll} \hat{E}(\theta) & = & \hat{E}_m(x) & = & \bar{y}_{..} & \\ \hat{var}(x|\theta) & = & \hat{\sigma}^2 & = & \frac{1}{N}\frac{1}{(n-1)}\Sigma_j\Sigma_i(y_{ij}-y.j)^2 & \\ \hat{var}_m(x) & = & (n\hat{\tau}^2+\hat{\sigma}^2) & = & \frac{1}{N-1}\Sigma_j(y_{.j}-y..)^2 & \textit{(cf. eq. 13)} \\ \hat{var}_\theta(\theta) & = & \hat{\tau}^2 & = & \max((\hat{var}_m(x)-\hat{\sigma}^2)/n,0) & \end{array}$$

Mit (eq. 22) findet man somit

$$d(\bar{y}_{.j}) = \frac{\hat{\tau}^2}{\hat{\sigma}^2/n+\hat{\tau}^2}\bar{y}_{.j} + \frac{\hat{\sigma}^2/n}{\hat{\sigma}^2/n+\hat{\tau}^2}\bar{y}_{..} \tag{23}$$

Die empirische Bayes Methode liefert uns also das Gewicht λ, das im Einführungsbeispiel gesucht war.

Literatur

[GCSR03] Andrew Gelman, John B. Carlin, Hal S. Stern, Donald B. Rubin: *Bayesian Data Analysis (2nd edition)* Texts in Statistical Science. Chapman & Hall CRC (2003)

[MarLwi89] J.S. Maritz and T. Lwin: *Empirical Bayes Methods (2nd edition)* Monographs on Statistics and Applied Probability 35, Chapman and Hall (1989)

[Rob07] Christian P. Robert *The Bayesian Choice (2nd edition)*, vol. 2, Springer Science and Business Media, New York (2007)

[Grize03] Y.L. Grize *Bayes Methoden.* Vorlesungsskript zum Modul 'Bayes Statistik' im NDK Statistik, Seminar für Statistik, ETH Zürich (2003). (Eine Verallgemeinerung findet sich in [Robb85])

[Robb85] Herbert Robbins, *Linear empirical Bayes estimation of means and variances*, Proc. Natl. Acad. Sci. USA, **82**, pp. 1571 -1574, (1985)