Bayesianische FDR (Teil 1)

Mareile Große Ruse

Seminar Multiples Testen

08.11.2010

Übersicht

- Wiederholung Bayes-Theorem
- False Discovery Rate

Übersicht

- Wiederholung Bayes-Theorem
- False Discovery Rate
- Empirischer Bayes-Ansatz

Übersicht

- Wiederholung Bayes-Theorem
- False Discovery Rate
- Empirischer Bayes-Ansatz
- Alternative: Nichtparametrisches Bayessches Modell

Übersicht

- Wiederholung Bayes-Theorem
- False Discovery Rate
- Empirischer Bayes-Ansatz
- Alternative: Nichtparametrisches Bayessches Modell
- Ausblick

Übersicht

- Wiederholung Bayes-Theorem
- False Discovery Rate
- Empirischer Bayes-Ansatz
- Alternative: Nichtparametrisches Bayessches Modell
- Ausblick

Bedingte Wahrscheinlichkeit

$(\Omega, \mathfrak{A}, \mathbb{P})$ W-Raum, $A, B \in \mathfrak{A}$, $\mathbb{P}(B) > 0$.

Definition

Die *bedingte Wahrscheinlichkeit von A gegeben B* ist definiert als

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

$$\underbrace{\mathbb{P}[A \mid B]}_{\text{a posteriori Wkeit von A}} = \frac{\mathbb{P}[B \mid A] \overbrace{\mathbb{P}(A)}^{\text{a priori Wkeit von A}}}{\mathbb{P}(B)}.$$

Bayes-Theorem für bedingte Wahrscheinlichkeiten

Theorem (Bayes-Theorem)

$A_1, \ldots, A_n \in \mathfrak{A}$ *mit* $\sum_{i=1}^{n} A_i = \Omega$ *und* $\mathbb{P}(A_i) > 0$.

$$\mathbb{P}\big[A_i \mid B\big] = \frac{\mathbb{P}\big[B \mid A_i\big]\mathbb{P}(A_i)}{\sum_{k=1}^{n} \mathbb{P}\big[B \mid A_k\big]\mathbb{P}(A_k)}$$

Bedingte Dichten

$\theta : (\Omega, \mathfrak{A}, \mathbb{P}) \to (S, \mathfrak{S})$, $(S, \mathfrak{S})$ Borelraum, μ σ-endl. Maß auf $(S, \mathfrak{S})$
$Z : (\Omega, \mathfrak{A}, \mathbb{P}) \to (\mathbb{R}, \mathfrak{B})$.
$f_{(Z,\theta)} : \mathbb{R} \times S \to \mathbb{R}_+$ sei $\lambda \otimes \mu$-Dichte von $\mathbb{P}^{(Z,\theta)}$. Dann

Theorem

$$f_\theta(\vartheta) = \int_{\mathbb{R}} f_{(Z,\theta)}(z, \vartheta) dz \quad \textit{ist } \mu\textit{-Dichte von } \mathbb{P}^\theta$$

$$f_Z(z) = \int_S f_{(Z,\theta)}(z, \vartheta) d\vartheta \quad \textit{ist } \lambda\textit{-Dichte von } \mathbb{P}^Z$$

$$f_{Z|\theta=\vartheta}(z) = \frac{f_{(Z,\theta)}(z, \vartheta)}{f_\theta(\vartheta)} \quad \textit{ist } \lambda\textit{-Dichte von } \mathbb{P}^{Z|\theta=\vartheta}$$

$$f_{\theta|Z=z}(\vartheta) = \frac{f_{(Z,\theta)}(z, \vartheta)}{f_Z(z)} \quad \textit{ist } \mu\textit{-Dichte von } \mathbb{P}^{\theta|Z=z}$$

Bayes-Theorem für Dichten

Theorem (Bayes-Theorem für Dichten)

Situation von oben. Dann

$$(i) \quad \boxed{f_{\theta|Z=z}(\theta) = \frac{f_{Z|\theta=\vartheta}(z) f_\theta(\vartheta)}{f_Z(z)}}$$

$$(ii) \quad \boxed{f_{Z|\theta=\vartheta}(z) = \frac{f_{\theta|Z=z}(\vartheta) f_Z(z)}{f_\theta(\vartheta)}}$$

Beispiel

$\theta : (\Omega, \mathfrak{A}) \to (\mathbb{N}, \mathfrak{P}(\mathbb{N}))$, μ Zählmaß auf $(\mathbb{N}, \mathfrak{P}(\mathbb{N}))$.
$Z : (\Omega, \mathfrak{A}) \to (\mathbb{R}, \mathfrak{B})$.
Es gelte:

$$\theta \sim Ber(p_1), \quad p_0 := 1 - p_1$$
$$Z \sim \mathcal{N}(0,1)(1-\theta) + \mathcal{N}(100,1)\theta$$

Gesucht: $\boxed{f_{\theta|Z=z}}$

Mit Bayes erhält man:

$$f_{\theta|Z=z}(\vartheta) = \frac{f_{Z|\theta=\vartheta}(z) f_\theta(\vartheta)}{f_Z(z)} = \begin{cases} \frac{p_0 \varphi_{0,1}(z)}{p_0 \varphi_{0,1}(z) + p_1 \varphi_{100,1}(z)}, & \text{für } \vartheta = 0, \\ \frac{p_1 \varphi_{100,1}(z)}{p_0 \varphi_{0,1}(z) + p_1 \varphi_{100,1}(z)}, & \text{für } \vartheta = 1 \end{cases}$$

False Discovery Rate (FDR)

Multiples Testproblem, n zu testende Hypothesen,
$\varphi = (\varphi_i, i = 1, \ldots, n)$ multiple Testprozedur.

V : (zufällige) Anzahl der Typ-1-Fehler

R : (zufällige) Anzahl von verworfenen Hypothesen insgesamt.

Definition

$$FDR(\varphi) = \mathbb{E}\left(\frac{V}{R \vee 1}\right)$$

heißt *False Discovery Rate (FDR)* von φ.

positive False Discovery Rate (pFDR)

Definition

Situation wie in obiger Definition.

$$\boxed{pFDR(\varphi) = \mathbb{E}\left[\frac{V}{R} \mid R > 0\right]}$$

heißt *positive False Discovery Rate (pFDR)* von φ.

Zweiklassen-Mischmodell

Definition (Zweiklassen-Mischmodell)

$(Z_1, H_1), \ldots, (Z_n, H_n)$ iid.
Für $i = 1, \ldots, n$: Z_i Teststatistik, H_i Indikatorvariable und

$$H_i \sim Ber(p_1)$$
$$Z_i | H_i \sim (1 - H_i) f_0 + H_i f_1$$

Bayesianische Interpretation

Theorem (Storey (2001))

Zweiklassen-Mischmodell, $\varphi = (\varphi_i, i = 1, \ldots, n)$ *multiple Testprozedur.*
φ_i *basierend auf Teststatistik* Z_i *mit Verwerfungsbereich* C*.*

$$pFDR(C) := pFDR(\varphi) = \mathbb{P}\big[H_1 = 0 \mid Z_1 \in C\big]$$

Lokale FDR

Analog zu $pFDR(C) = \mathbb{P}\big[H = 0 \mid Z \in C\big]$ wird definiert:

Definition

Lokale fdr: $$\boxed{fdr(z) := \mathbb{P}\big[H = 0 \mid Z = z\big]}$$

Bayes-Formel für Dichten:

$$fdr(z) = f_{H|Z=z}(0) = \frac{p_0 f_0(z)}{f_Z(z)}$$

Was kommt nun?

Im Zweiklassen-Mischmodell:

1. **Empirischer Bayes-Ansatz**;
   Bewertungskriterium: $\widehat{\mathbf{fdr}}(\mathbf{z})$
2. **Nichtparametrischer Bayes-Ansatz**;
   Bewertungskriterium: $\mathbb{E}\left[\mathbf{fdr}(\mathbf{z}) \mid \mathbf{Y}\right]$

Die *fdr* in empirischen Bayes-Methoden

Empirischer Bayes-Ansatz:

1. wähle p_0, f_0
2. schätze f aus den Daten
3. bilde $\widehat{fdr}(z_i) = \frac{p_0 f_0(z_i)}{\hat{f}(z_i)}$
4. ziehe Schlüsse aus dem Wert

Kritik: Subjektivität.

Der empirische Bayes-Ansatz

$n = 6033$ Gene.
Für jedes Gen: 50 Messungen in Normalgewebe, 52 in Tumorgewebe.

T_i = 2-Stichproben-t-Statistik für Gen i
F^t_{100} = Verteilungsfunktion zu t_{100}.

Zweiklassen-Mischmodell:

$$Z_i = \Phi^{-1}(F^t_{100}(T_i)) \qquad (z\text{-Score})$$
$$H_i \sim Ber(p_1)$$
$$Z_i | H_i \sim (1 - H_i) f_0 + H_i f_1$$

Gesucht:

$\mathbf{P}_{\text{emp}}(z_i)$:= Wahrscheinlichkeit, dass Gen mit z-Score z_i im Tumorgewebe andere Expression hat als im Normalgewebe.

Geschätzte fdr

Plan: Durchführung des obigen Punkteprogramms:

1. $f_0(z) := \varphi_{0,1}(z), \quad p_0 := 0.9.$
2. $\hat{f}$ Schätzung von f.
3. $$\widehat{fdr}(z) = \frac{0.9\varphi_{0,1}(z)}{\hat{f}(z)}$$
   $$= \text{Wahrscheinlichkeit für Nicht-Expression geg. } Z = z$$
   $$= 1 - \mathbf{P}_{\text{emp}}(z)$$
4. Schlüsse aus dem Wert ziehen:
   Entscheidungsregel:

Erkläre Expressionsunterschied für Gen $i \Leftrightarrow \widehat{fdr}(z_i) \leq 0.2$

Der nichtparametrische Ansatz

Setting:

- n zu untersuchende Gene
- Alle wichtigen Informationen eines Gens in sog. z-Score enthalten.
- Pro Gen i je ein Kontrollscore Z_i^{null} und ein Effekt-Score Z_i^{mix}
  $Z^{\text{null}} = (Z_i^{\text{null}}, i = 1, \ldots, n)$ **Null-Sample**,
  $Z^{\text{mix}} = (Z_i^{\text{mix}}, i = 1, \ldots, n)$ **Mix-Sample**
- Umnummerierung:
  $Z_i, i = 1, \ldots, n$ für Null-Sample;
  $Z_i, i = n+1, \ldots, 2n$ für Mix-Sample.

Fragestellung und Ansatz

Gesucht:

$\mathbf{P}(z_i)$ =Wahrscheinlichkeit, dass Gen i im Tumorgewebe andere Expression hat als im Normalgewebe.

Idee:

Verbinde a priori Kenntnisse mit in Daten enthaltenen Informationen

$$H_i = \begin{cases} 0, & \text{falls Gen } i \text{ ohne Veränderung} \\ 1, & \text{falls Gen } i \text{ mit Veränderung} \end{cases}$$

Verteilungsannahmen

1. $Z_1, \ldots, Z_n$ iid. und $Z_{n+1}, \ldots, Z_{2n}$ iid.
2. Null-Sample und Mix-Sample unabhängig

Verteilungsannahmen

1. $Z_1, \ldots, Z_n$ iid. und $Z_{n+1}, \ldots, Z_{2n}$ iid.
2. Null-Sample und Mix-Sample unabhängig
3. Für $i \leq n$ (Null-Sample)

$$Z_i \sim f_0(Z_i)$$

Verteilungsannahmen

1. $Z_1, \ldots, Z_n$ iid. und $Z_{n+1}, \ldots, Z_{2n}$ iid.
2. Null-Sample und Mix-Sample unabhängig
3. Für $i \leq n$ (Null-Sample)

$$Z_i \sim f_0(Z_i)$$

4. Für $i > n$ (Mix-Sample)

$$Z_i \sim f(Z_i) = p_0 f_0(Z_i) + (1 - p_0) f_1(Z_i)$$

Verteilungsannahmen

1. $Z_1, \ldots, Z_n$ iid. und $Z_{n+1}, \ldots, Z_{2n}$ iid.
2. Null-Sample und Mix-Sample unabhängig
3. Für $i \leq n$ (Null-Sample)

$$Z_i \sim f_0(Z_i)$$

4. Für $i > n$ (Mix-Sample)

$$Z_i \sim f(Z_i) = p_0 f_0(Z_i) + (1 - p_0) f_1(Z_i)$$

5. also $\forall\, i$: $Z_i | H_i \sim f_0(1 - H_i) + f_1 H_i$

Verteilungsannahmen

1. $Z_1, \ldots, Z_n$ iid. und $Z_{n+1}, \ldots, Z_{2n}$ iid.
2. Null-Sample und Mix-Sample unabhängig
3. Für $i \leq n$ (Null-Sample)

$$Z_i \sim f_0(Z_i)$$

4. Für $i > n$ (Mix-Sample)

$$Z_i \sim f(Z_i) = p_0 f_0(Z_i) + (1 - p_0) f_1(Z_i)$$

5. also $\forall\, i$: $Z_i | H_i \sim f_0(1 - H_i) + f_1 H_i$

$Z \sim Z^{\text{mix}}$. Bayes-Regel liefert:

$$\mathbb{P}\big[H = 1 \mid Z, f_0, f_1, p_0\big] = \frac{p_1 f_1(Z)}{p_0 f_0(Z) + p_1 f_1(Z)} = \frac{p_1 f_1(Z)}{f(Z)} \qquad (*)$$

Plan:

- Festlegen eines a priori Wahrscheinlichkeitsmodells für f_0, f_1, p_0.
- Bestimmung der a posteriori Verteilungen $\mathbb{P}^{f_0|Y}, \mathbb{P}^{f_1|Y}, \mathbb{P}^{p_0|Y}$ gegeben alle Daten (=: Y).
- $\mathbf{P}(z_i) := \mathbb{E}\big[\frac{p_1 f_1(z_i)}{f(z_i)} \mid Y\big] \approx \frac{\mathbb{E}\big[p_1|Y\big]\mathbb{E}\big[f_1(z_i)|Y\big]}{\mathbb{E}\big[p_0|Y\big]\mathbb{E}\big[f_0(z_i)|Y\big] + \mathbb{E}\big[p_1|Y\big]\mathbb{E}\big[f_1(z_i)|Y\big]}$
- Aus $\mathbf{P}(z_i)$ Schlüsse ziehen.

Notationen

$$N_0 = \mathcal{N}(b, \sigma_0^2)$$

$$N_1 = \frac{1}{2}\mathcal{N}(-\tilde{b}, \tilde{\sigma}^2) + \frac{1}{2}\mathcal{N}(\tilde{b}, \tilde{\sigma}^2)$$

$$G_0(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\mu_k}(\cdot), \qquad G_1(\cdot) = \sum_{k=1}^{\infty} \tilde{\omega}_k \delta_{\tilde{\mu}_k}(\cdot)$$

mit

$\mu_k \overset{iid.}{\sim} N_0$ und $\tilde{\mu}_k \overset{iid.}{\sim} N_1$,

$\sigma^2, \sigma_0^2, \tilde{\sigma}^2$ Invers-Gamma-verteilt , $b, \tilde{b}$ normalverteilt

$\omega_k, \tilde{\omega}_k$ unabh. zufällige reskalierte Gewichte aus $unif(0, 1)$

A priori Wahrscheinlichkeitsmodell

$$p_0 \sim unif(0.05, 1)$$

$$f_0(z) = \int \varphi_{\mu,\sigma^2}(z) dG_0(\mu) = \sum_{k=1}^{\infty} \omega_k \varphi_{\mu_k,\sigma^2}(z)$$

$$f_1(z) = \int \varphi_{\tilde{\mu},\sigma^2}(z) dG_1(\tilde{\mu}) = \sum_{k=1}^{\infty} \tilde{\omega}_k \varphi_{\tilde{\mu}_k,\sigma^2}(z)$$

MCMC im hierarchischen Modell (1)

Darstellung als hierarchisches Modell:

$$Z_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

$$\mu_i \sim \begin{cases} G_0, & \text{falls } r_i = 0, \\ G_1, & \text{falls } r_i = 1 \end{cases}$$

$$r_i \sim \begin{cases} \delta_0, & \text{für } i = 1, \ldots, n \\ Ber(p_1), & \text{für } i = n+1, \ldots, 2n \end{cases}$$

MCMC im hierarchischen Modell (2)

$\vartheta := (\sigma^2, \sigma_0^2, \tilde{\sigma}^2, b, \tilde{b}, p_0)$ und $\vartheta_\eta = \vartheta$ ohne Komponente η.

Eigenschaft von konjugierten a priori Verteilungen liefert:

$\mathbb{P}^{p_0|(Y,\vartheta_{p_0})}$: Beta-Verteilung

$\mathbb{P}^{\sigma^2|(Y,\vartheta_{\sigma^2})}, \mathbb{P}^{\sigma_0^2|(y,\vartheta_{\sigma_0^2})}, \mathbb{P}^{\tilde{\sigma}^2|(Y,\vartheta_{\tilde{\sigma}^2})}$: Invers-Gamma-Verteilungen

$\mathbb{P}^{b|(Y,\vartheta_b)}, \mathbb{P}^{\tilde{b}|(Y,\vartheta_{\tilde{b}})}$: Normal-Verteilungen

MCMC im hierarchischen Modell (3)

- Einige Überlegungen (μ_i sind diskret verteilt) liefern bedingte Verteilung der μ_i.
- erhalten so letztlich Realisierungen aus $\mathbb{P}^{G_0|Y}, \mathbb{P}^{G_1|Y}$
- damit Bestimmung von $\mathbb{E}\left[f_0 \mid Y\right], \mathbb{E}\left[f_1 \mid Y\right]$

Inferenz

Nach Idee oben:

$$\begin{aligned}\mathbf{P}(z_i) &= \mathbb{E}\Big[\frac{p_1 f_1(z_i)}{f(z_i)} \mid Y\Big] = \int \frac{p_1 f_1(z_i)}{f(z_i)} d\mathbb{P}^{(f_0,f_1,p_0)|Y} \\ &= 1 - \mathbb{E}\Big[fdr(z_i) \mid Y\Big] \\ &\approx \frac{\mathbb{E}\Big[p_1 \mid Y\Big]\mathbb{E}\Big[f_1(z_i) \mid Y\Big]}{\mathbb{E}\Big[p_0 \mid Y\Big]\mathbb{E}\Big[f_0(z_i) \mid Y\Big] + \mathbb{E}\Big[p_1 \mid Y\Big]\mathbb{E}\Big[f_1(z_i) \mid Y\Big]}\end{aligned}$$

Empirischer Ansatz:

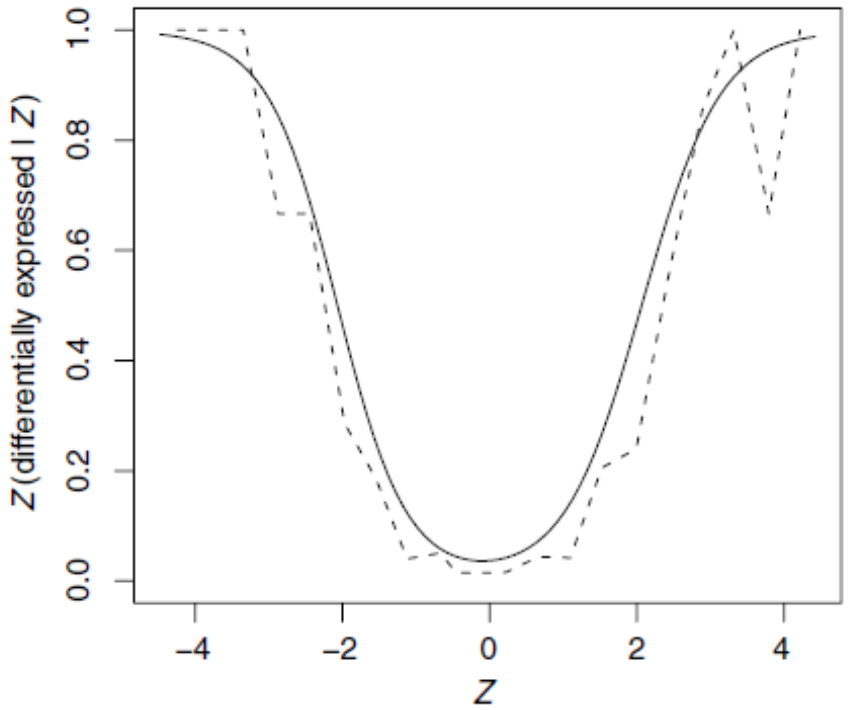
$$\mathbf{P}_{\text{emp}}(z_i) = \frac{\hat{p}_1 f_1(z_i)}{\hat{f}(z_i)} = 1 - \hat{fdr}(z_i)$$

Simulationsstudie - Wie gut ist $\mathbf{P}(z)$?

Die Realität sehe so aus:

$$f_0 = \mathcal{N}(0,1), f_1 = 0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1), p_0 = 0.8; n = 10\,000$$

- Teilen die Werte $z_1, \ldots, z_{10000}$ in Bins ein
- Ordnen Bin j die Anzahl $n_j = |\{k : z_k \in \text{Bin } j, H_k = 1\}|$ der Gene mit Veränderung zu
- Vergleich dieser Funktion (- - -) mit $\mathbf{P}(z)$ (—)



$\mathbf{P}(i)$ in Abhängigkeit vom wahren p_0

A posteriori Wahrscheinlichkeit für Expressionsunterschied in Abhängigkeit vom zugrunde liegenden wahren p_0 und dem z-Score.

Entscheidungsregel:

Gen mit Score z hat Expressionsunterschied $\Leftrightarrow \mathbf{P}(z) > 0.5$.

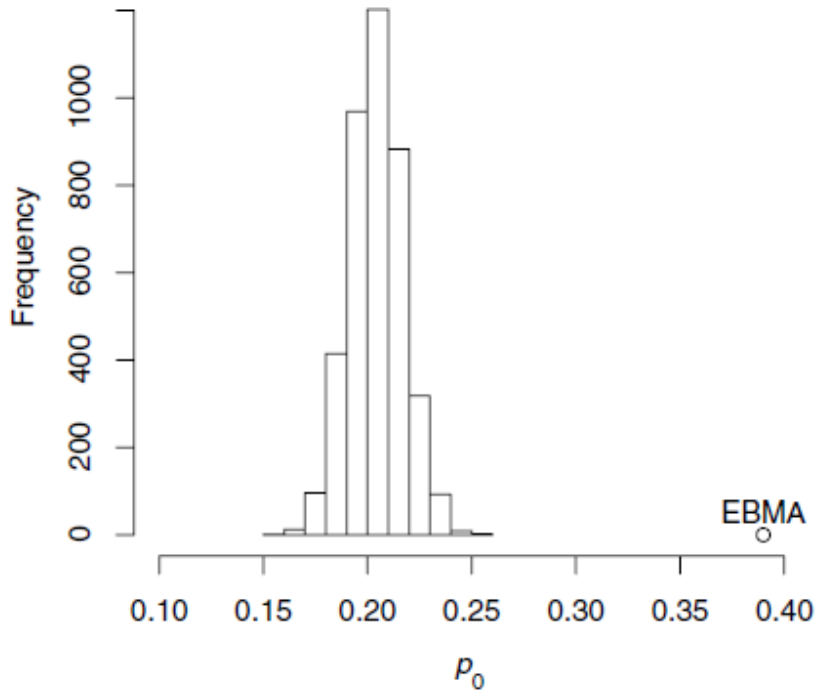
p_0	z										
	-5.00	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00	4.00	5.00
0.4	1.00	1.00	0.98	0.87	0.46	0.19	0.43	0.85	0.98	1.00	1.00
0.8	0.94	0.90	0.75	0.41	0.14	0.07	0.13	0.44	0.81	0.93	0.96
0.95	0.46	0.42	0.27	0.11	0.05	0.03	0.04	0.10	0.28	0.43	0.50

Empirischer und nichtparametrischer Ansatz in der Anwendung

Vergleich:

Dichte $f_{p_0|Y}$ und Punktschätzer

$\hat{p}_0 = 0.39$



Ausblick

- Verlustfunktionen
- Entscheidungsregel

Ausblick

- Verlustfunktionen
- Entscheidungsregel
- Optimalität

Ausblick

- Verlustfunktionen
- Entscheidungsregel
- Optimalität

Vielen Dank für eure Aufmerksamkeit!