Data Augmentation

Michael Obermeier

12. Mai 2005

Inhaltsverzeichnis

1 Markov Chain Monte Carlo

Markov Chain Monte Carlo

Situation: Gesucht sind die Parameter einer Verteilung mit Dichte $f(Z)$, die numerisch schwer zugänglich ist. $\longrightarrow$ posteriori-Verteilung von θ in Bayes-Inferenz

$$p(\theta|y,x) = \frac{p(y|x,\theta)p(\theta)}{\int p(y|x,\theta)p(\theta)d\theta}$$

Dabei: y Daten, x gegebene Kovariablen, θ unbekannter Parameter

Grundidee:

- Ziehen von pseudozufälligen Werten aus einer Markovkette $\{Z^{(t)} : t = 0, 1, 2, \ldots\}$, die gegen die *Zielverteilung* $f(Z)$ konvergiert.
- Zustände $Z^{(t)}$ entsprechen den gezogenen Zufallszahlen, die aus einer sogenannten Vorschlagsdichte stammen, welche nur von $Z^{(t-1)}$ abhängt.
- Für genügend großes t: $Z^{(t)}$ ist approximativ nach $f(Z)$ verteilt $\Longrightarrow$ $\{Z^{(t)}, Z^{(t+1)}, \ldots\}$ kann man als Sequenz von Zufallszahlen aus der Zielverteilung betrachten (t: Konvergenzzeit, burn-in)
- Aus diesen Werten: Schätzung der Verteilungseigenschaften

1.1 Metropolis-Hastings Algorithums

Metropolis-Hastings Algorithums (Metropolis 1953, Hastings 1970)

Erzeugt eine Markovkette, deren Glieder $Z^{(t)}$ unter geringen Voraussetzungen nach einer gewissen burn-in Phase als Zufallszahlen der gesuchten Verteilung aufgefasst werden können.

Algorithmus:

(1) Ziehe Kandidaten $\tilde{Z}$ aus der Vorschlagsdichte $g(Z^{(t-1)}, \tilde{Z})$

(2) $Z^{(t)} := \tilde{Z}$ mit der Akzeptanzwahrscheinlichkeit

$$\alpha(Z^{(t-1)}, Z^t) = \min\left\{\frac{g(\tilde{Z}, Z^{(t-1)}) \cdot f(\tilde{Z})}{g(Z^{(t-1)}, \tilde{Z}) \cdot f(Z^{(t)})}, 1\right\}$$

sonst: $Z^{(t)} := Z^{(t-1)}$

Wahl der Vorschlagsdichte: verschiedene Herangehensweisen ("Unabhängigkeits-Proposal", "Random Walk-Proposal",... siehe Jerak (2000))

Verallgemeinerungen des M-H Algorithmus

Z üblicherweise mehrdimensional $\Longrightarrow$ geeignetes Verfahren zum Aktualisieren: Unterteile Z in K logisch zusammenhängende Blöcke und aktualisiere die einzelnen Blöcke, gegeben alle anderen. Sei $f_{k|-k}(Z_k|Z_{-k})$ die bedingte Verteilung des k-ten Blocks bei gegebenen übrigen Dichten $\Longrightarrow$ Hybrid Algorithmus

- Sei $Z_{-k}^{(t)} = (Z_1^{(t)}, \ldots, Z_{k-1}^{(t)}, Z_{k+1}^{(t)}, \ldots Z_k^{(t)})'$
- (1) Ziehe $\widetilde{Z}_k$ aus Vorschlagsdichte $g_k(Z^{(t-1)}, \widetilde{Z}_k|Z_{-k}^{(t)})$

  (2) $Z^{(t)} = \widetilde{Z}_k$ mit Wahrscheinlichkeit

$$\alpha(Z_k^{(t-1)}, \widetilde{Z}_k) = \min\left\{\frac{g_k(\widetilde{Z}_k, Z^{(t-1)}|Z_{-k}^{(t)}) \cdot f_{k|-k}(\widetilde{Z}_k|Z_{-k}^{(t)})}{g_k(Z^{(t-1)}, \widetilde{Z}_k|Z_{-k}^{(t)}) \cdot f_{k|-k}(Z_k^{(t-1)}|Z_{-k}^{(t)})}; 1\right\}$$

sonst: $Z_k^{(t)} = Z_k^{(t-1)}$

1.2 Gibbs-Sampler

Gibbs-Sampler

- Direktes Ziehen von $Z_k^{(t)}$ aus den vollständig bedingten Dichten $f_{k|-k}(\widetilde{Z}_k|Z_{-k}^{(t)})$
- Somit:

$$g_k(Z^{(t-1)}, \widetilde{Z}_k|Z_{-k}^{(t)}) = f_{k|-k}(\widetilde{Z}_k|Z_{-k}^{(t)}),$$

  woraus folgt: Akzeptanzwahrscheinlichkeit $= 1$
- Also: Wert von Z in Schritt t:

$$\begin{aligned} Z_1^{(t)} &\sim P(Z_1|Z_2^{(t-1)}, Z_3^{(t-1)}, \ldots, Z_K^{(t-1)}) \\ Z_2^{(t)} &\sim P(Z_2|Z_1^{(t)}, Z_3^{(t-1)}, \ldots, Z_K^{(t-1)}) \\ &\vdots \\ Z_K^{(t)} &\sim P(Z_K|Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_{K-1}^{(t)}) \end{aligned}$$

$$\Longrightarrow \quad Z^{(t)} = (Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_K^{(t)})$$

1.3 Data Augmentation

Data Augmentation (Tanner & Wong, 1987)

Motivation:

- Gesucht ist die posteriori-Dichte $p(\theta|y)$ des Parameters θ, die jedoch nur schwer direkt zu berechnen ist
- Die beobachteten Daten y werden durch unbeobachtete Daten z so vermehrt (augmented), dass $p(\theta|(y,z))$ berechnet werden kann
- posteriori Dichte:

$$\begin{aligned} p(\theta|y) &= \int_Z p(\theta|(y,z))p(z|y)dz = \int_Z p(\theta|(y,z))dP(z|y) \\ &= \mathbb{E}_{P(Z|Y)}[p(\theta|z,y)] \end{aligned}$$

Idee:
1. Aus Approximation für $p(z|y)$ ziehe $z_1, \dots z_m$
2. Aktualisieren der Approximation für $p(\theta|y)$ durch $p(\theta|y) \approx \frac{1}{m}\sum_{j=1}^{m} p(\theta|z_j, y)$

Der Algorithmus

Idee:
- Ziehe θ aus $p(\theta|y)$
- Ziehe z aus $p(z|y) = \int_\Theta p(z|\theta,y)p(\theta|y)d\theta$

1. Wähle Startverteilung $p_0(\theta|y)$
2. Ziehe $z_1, \dots, z_m \sim p(z|y)$
   (a) Ziehe $\theta_1, \dots, \theta_m$ aus $p^{(t)}(\theta|y)$
   (b) Ziehe $z_1, \dots, z_m$ aus $p(z|\theta_i, y)$
3. Setze $p^{(t+1)}(\theta|y) = \frac{1}{m}\sum_{i=1}^{m} p(\theta|z_i, y)$
4. Wiederhole Schritt 2 und 3 bis zur Konvergenz

Data Augmentation, andere Herangehensweise

- Gegeben: Zufallsvektor $z = (u, v)$, dessen gemeinsame Verteilung schwer zu simulieren ist
- Sei $Z^{(t)} = (z_1^{(t)}, z_2^{(t)}, \dots, z_m^{(t)}) = \left((u_1^{(t)}, v_1^{(t)}), \dots, (u_m^{(t)}, v_m^{(t)})\right)$
- Aktualisieren dieses Vektors in zwei Schritten:
  1. Ziehen von $u_i^{(t+1)} \sim g(u|v_i^{(t)})$ (unabh. für $i = 1, 2, \dots m$)

2. Ziehen von: $v_i^{(t+1)} \sim \bar{h}(v|U^{(t+1)}) = \frac{1}{m}\sum_{i=1}^{m} h(v|u_i^{(t+1)})$

$\Longrightarrow$ Neuer Vektor $Z^{(t+1)} = \left((u_1^{(t+1)}, v_1^{(t+1)}), \ldots, (u_m^{(t+1)}, v_m^{(t+1)})\right)$

- Wiederholen des Algorithmus bis zur Konvergenz: $P(Z^{(t)}) \longrightarrow P(z)$ in Verteilung
- Für $m = 1$: Data Augmentation ist ein Spezialfall des Gibbs Sampling mit $K = 2$ und somit auch ein Spezialfall des Metropolis-Hastings Algorithmus.
- Für $p^{(t+1)}(\theta|y) = p(\theta|z_i, y)$: Data Augmentation entspricht m unabhängigen parallelen Läufen beim Gibbs Sampling mit jeweils $K = 2$.

2 Data Augmentation und Missing Data

Data Augmentation und Missing Data

$P(\theta|Y_{obs})$ schwer zu berechnen (Y_{obs}: beobachtete Daten) $\Longrightarrow$ Augmentieren mit Y_{mis} (nicht-beobachtete Daten) Algorithmus:

- **I(mputation)-Schritt**: $Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)})$ Ziehen von Werten, gegeben die beobachteten Werte und die aktuellen Parameterschätzungen
- **P(osterior)-Schritt**: $\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)})$ Neue Schätzungen werden durch Ziehung aus den aktualisierten posterioris unter Einbeziehung aller Werte $(Y_{obs}, Y_{mis}^{(t+1)})$ gewonnen

Man erhält eine Markov-Kette $\{(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, \ldots\}$, deren stationäre Verteilung $P(\theta, Y_{mis}|Y_{obs})$ ist. Speziell hat die Folge $\{\theta^{(t)} : t = 1, 2, \ldots\}$ als stationäre Verteilung $P(\theta|Y_{obs})$.

3 Multiple Imputation

Multiple Imputation (Rubin 1987)

Fehlende Werte werden durch mehrere simulierte Werte ersetzt: $Y_{mis}^{(1)}, Y_{mis}^{(2)}, , \ldots Y_{mis}^{(m)}$

- Erzeugen von $m > 1$ Versionen von Y_{mis}
- Analyse der m vervollständigten Datensätze mit Standardverfahren
- Auswertung durch Kombination der Ergebnisse (Variation zwischen den Ergebnissen,...)
- Vorteil: realistischere Schätzung der Varianz

Data Augmentation und Multiple Imputation

Um gültige Inferenzschlüsse ziehen zu können, müssen die Realisationen von $P(Y_{mis}|Y_{obs})$ unabhängig sein!

- Zwei Möglichkeiten:
  - verwende die Werte $Y_{mis}^{(k)}, Y_{mis}^{(2k)}, , \ldots Y_{mis}^{(mk)}$ aus dem Data Augmentation Algorithums

- simuliere m unabhängige Ketten der Länge k und verwende jeweils die letzten Werte $Y_{mis}^{(k)}$ jeder Kette

- Dabei: k muss groß genug sein, um approximativ Unabhängigkeit zu erreichen
- Insgesamt $k \cdot m$ Schritte Allerdings: m sehr klein (i.d.R. $3 \leq m \leq 5$), da die relative Effizienz eines Punktschätzers im Vergleich zu $m = \infty$

$$(1 + \frac{\gamma}{m})^{-1}$$

ist. (Mit γ: Anteil fehlender Werte, m: Anzahl Imputationen)

4 Ein Beispiel

Ein Beispiel: Cholesterinspiegel bei Herzinfarktpatienten

Bei 28 Herzinfarktpatienten am Pennsylvania Medical Center wurde jeweils zwei und vier Tage nach dem Infarkt der Cholesterinspiegel gemessen. Zusätzlich wurde er bei 19 Patienten noch am 14. Tag gemessen.

Y_1	Y_2	Y_3
270	218	156
236	234	-
210	214	242
142	116	-
280	200	-
272	276	256
160	146	142
220	182	216
	$\vdots$	

- Data Augmentation mit zugrundegelegter nicht-informativer priori für θ (inverse Wishart-Verteilung für θ: $\pi(\theta) \propto |\Sigma|^{-(\frac{p+1}{2})}$)
- (Hier im Beispiel) interessierende Parameter:
  - μ_3: durchschnittlicher Cholesterinspiegel nach 14 Tagen
  - $\delta_{13} = \mu_1 - \mu_3$: durchschnittliche Abnahme des Cholesterinspiegels vom 2. bis zu 14. Tag
  - $\tau_{13} = 100(\mu_1 - \mu_3)/\mu_1$: relative Abnahme des Cholesterinspiegels zwischen 2. und 14. Tag
- Simulation einer Kette der Länge 5100, wobei die ersten 100 als "burn-in-Werte" nicht betrachtet werden.
- Als Startwerte für θ (bzw. μ): ML-Schätzer aus den Y_{obs}

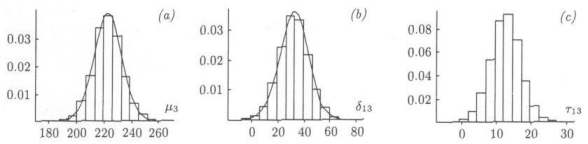


Abbildung 1: (a): μ_3, (b): δ_{13}, (c): τ_{13}

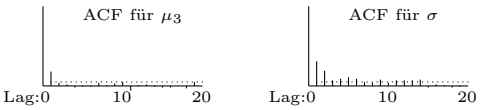
Histogramme der simulierten Werte

Mittelung über alle 5000 Werte der jeweiligen Parameter ergab die Schätzungen (bei zweimaligem Simulieren):

	μ_3	δ_{13}	τ_{13}
mean	222.2	31.8	12.4
95% intervall	(201.6, 244.0)	(8.9, 55.4)	(3.7, 20.9)
mean	222.4	31.4	12.3
95% intervall	(201.7, 242.6)	(8.9, 53.3)	(3.7, 20.3)

Multiple Imputation

- Realisationen von $P(Y_{mis}|Y_{obs})$ müssen unabhängig sein $\Longrightarrow$ Betrachtung der Autokorrelation zwischen den realisierten Werten einer Data Augmentation-Kette
- Hier:



- Bei einem Lag von 50 sicher keine Autokorrelation in den Daten: $\Longrightarrow$ 5 $(:= m)$ Ketten der Länge 50 $(:= k)$ (Startwerte gefunden mittels EM-Algorithmus) $\Longrightarrow$ Auffüllen des Datensatzes mit dem jeweils letzten Glied jeder Kette

Tabelle des aufgefüllten Datensatzes

beobachtete Daten			imputierte Werte für Y_3				
Y_1	Y_2	Y_3	1	2	3	4	5
270	218	156					
236	234	-	186	259	200	259	227
210	214	242					
142	116	-	238	50	116	133	197
280	200	-	187	190	186	222	169
272	276	256					
160	146	142					
220	182	216					
226	238	248					
242	288	-	243	264	295	234	215
186	190	168					
	⋮						

Punktschätzer und Kombination der Ergebnisse

	μ_3		δ_{13}		τ_{13}	
t	$\widehat{Q}^{(t)}$	$\sqrt{U^{(t)}}$	$\widehat{Q}^{(t)}$	$\sqrt{U^{(t)}}$	$\widehat{Q}^{(t)}$	$\sqrt{U^{(t)}}$
1	221.3	7.56	32.61	10.21	12.84	3.72
2	219.1	10.35	34.86	9.34	13.73	3.53
3	224.8	9.31	29.14	9.97	11.48	3.73
4	218.7	7.69	35.25	8.39	13.88	3.03
5	220.3	7.82	33.61	9.83	13.23	3.58

Dabei: $\widehat{Q}^{(t)}$: Mittelwert, $\sqrt{U^{(t)}}$: Standardfehler des Schätzers

	$\bar{Q}$	$\sqrt{T}$	FG	95%-Intervall
μ_3	220.8	9.02	517	(203.1, 238.6)
δ_{13}	33.09	9.94	760	(13.59, 52.60)
τ_{13}	13.03	3.68	595	(5.80, 20.26)

Mit:

$$\begin{aligned} \bar{Q} &= \frac{1}{m}\sum_{t=1}^{m}\widehat{Q}^{(t)} \qquad \text{"mittlerer Punktschätzer"} \\ T &= \bar{U} + (1+\frac{1}{m})B \qquad \text{Varianz dieses Schätzers} \\ &= \tfrac{1}{m}\textstyle\sum_{t=1}^{m} U^{(t)} + (1+\tfrac{1}{m})\cdot\Big(\tfrac{1}{m-1}\sum_{t=1}^{m}(\widehat{Q}^{(t)}-\bar{Q})^2\Big) \end{aligned}$$

Literatur

[1] Jerak, A. (2000). *Markov Chain Monte Carlo-Verfahren: Eine kurze Einführung*, http://www.statistik.lmu.de/ alex/forschung.

[2] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

[3] Tanner, M.A. und Wong, W.H. (1987). *The Calculation of Posterior Distributions by Data Augmentation*, Journal of the American Statistical Association.