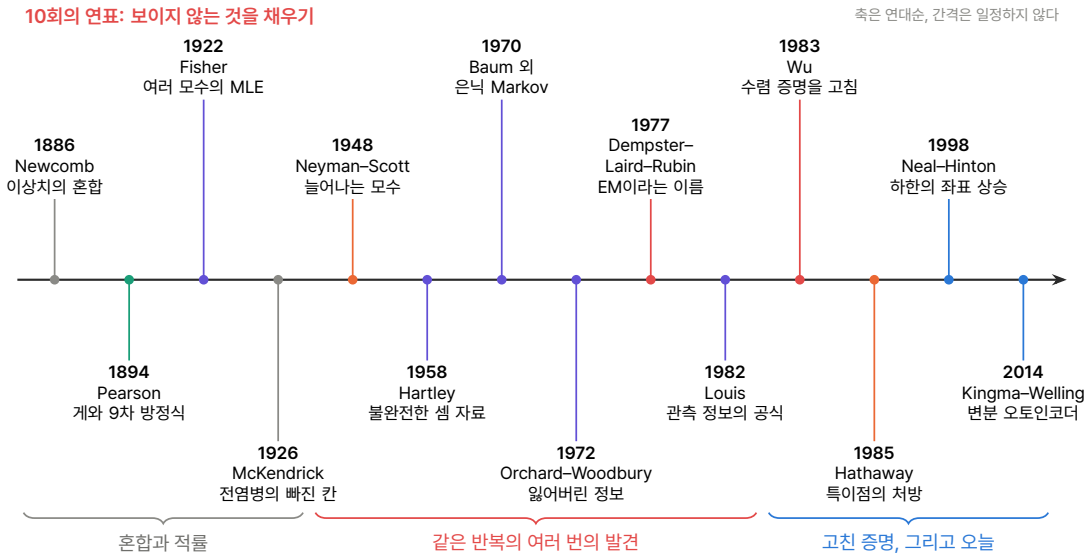
수리통계학 × 통계학의 역사

10회 강의노트

숨은 변수와 EM
보이지 않는 것을 채우는 알고리즘

1894 → 1977

Pearson은 1894년 게 1,000마리의 이마 폭에서 보이지 않는 두 무리를 찾으려고 9차 방정식을 손으로 풀었다. 83년 뒤 Dempster, Laird, Rubin은 같은 종류의 문제를 "보이지 않는 것을 기댓값으로 채우고 최대화하기"의 반복으로 풀어 EM이라 불렀다. 이 노트는 여러 모수의 가능도 이론을 완성하고, EM의 단조성과 수렴 속도를 증명한 뒤 정규 혼합에 쓴다.



다루는 것	정보 행렬 · 다변량 점근정규성 · 프로파일 가능도 · 불완전 자료 · EM 알고리즘 · 단조성(Gibbs) · ELBO · 잃어버린 정보와 수렴 속도 · Louis 공식 · 정규 혼합 · 특이점과 성분 수
원자료	Weldon의 나폴리 게 1,000마리(Pearson 1894, 묶음 자료) · Old Faithful 간헐천 272번 분출 · Dempster–Laird–Rubin의 유전 연관 자료
노트북	`sessions/10_em/10_em.ipynb` — 절·정리 번호가 같다
워크북	`10_em_workbook.pdf` (손으로 푸는 34문항) · 풀이판 `_workbook_sol.pdf`
선수 지식	1–9회 (특히 2회 Jensen·Gibbs, 4회 혼합분포, 7회 다변량 CLT, 9회 최대가능도·Wilks)

이 노트를 쓰는 법

이 노트는 다른 책 없이 읽도록 썼다. 쓰이는 정의와 정리는 모두 이 노트나 앞 회차의 노트에서 증명한다. 이 과정의 범위를 넘어 증명하지 않는 사실은 **증명 없이 쓰는 사실** 상자에 따로 밝힌다.

각 절은 **역사 → 정의·정리 → 증명 → 그림 → 연습문제** 순서다. 핵심 정리의 증명은 번호 붙은 단계 ① ② ③으로 쓰고, 단계마다 오른쪽 끝에 근거(앞 정리의 번호나 앞 회차의 정리)를 단다. 읽을 때는 근거를 가리고 "왜 이 단계가 성립하는가"를 먼저 말해 본다. 미분과 적분의 순서를 바꾸거나 전개의 나머지를 버리는 곳에는 **조건 확인** 단락이 있다. 나머지 정리는 **증명 전략**만 적고, 증명을 소문항으로 나눈 유도형 연습으로 돌린다. 이 연습은 건너뛰지 않고 손으로 쓴다.

근거에 붙은 [T1]–[T8]은 별책 **증명 도구 상자**의 도구 번호다. 이 회차에서는 [T1] 지시함수, [T4] 완전제곱, [T6] 미분과 적분의 교환, [T8] Taylor 전개를 쓴다.

연습문제는 소절 끝에 모여 있다. 난이도는 ★ 계산·확인, ★★ 증명·응용, ★★★ 도전이다. 모든 문제의 풀이가 부록 A에 있고, 문제 끝의 **풀이 ○○쪽**을 누르면 그 쪽으로 간다. 손으로 푸는 문제는 따로 **워크북**(계산 드릴, 빈칸 증명, 증명 문제, 퀄형)에 모았다.

상자의 뜻

> **정의** 대상을 처음 정하는 문장. 왼쪽 선이 보라다.

> **정리** 증명할 주장(정리·보조정리·따름정리·명제). 왼쪽 선이 빨강이다.

> **예제** 숫자를 넣어 끝까지 계산해 보는 곳. 왼쪽 선이 파랑이다.

> **가정 주의** 빼면 정리가 무너지는 가정. 왼쪽 선이 주황이다.

> **증명 없이 쓰는 사실** (예) 이 과정에서 증명하지 않고 쓰는 결과. 무엇이 참인지, 왜 그럴듯한지, 어디서 증명되는지를 밝힌다.

> **역사** 그 개념이 태어난 사람과 문제. 회색 선 옆의 작은 글씨다.

> **연습** 소절 끝의 연습문제. 가는 청록 선이고, 끝에 풀이가 있는 쪽 번호가 붙는다.

색의 뜻 수식과 그림에서 보라는 모수, 청록은 데이터·확률변수, 파랑은 통계량, 빨강은 핵심 결론이다. 노트북과 같은 색이다.

차례

0 들어가며 1

0.1 1–9회에서 가져오는 것 2

0.2 표기 2

1 여러 모수의 가능도 4

1.1 정보 행렬과 점근정규성 4

1.2 귀찮은 모수의 대가 5

1.3 프로파일 가능도 7

2 보이지 않는 자료의 가능도 10

2.1 완전 자료와 관측 자료 10

2.2 가능도의 분해 11

3 EM 알고리즘 13

3.1 알고리즘과 단조성 13

3.2 하한을 오르는 알고리즘 14

3.3 잃어버린 정보와 수렴 속도 15

4 정규 혼합의 EM 19

4.1 E 단계와 M 단계 19

4.2 이름표와 대칭 20

5 데이터로 — 게, 간헐천, 그리고 “may” 22

5.1 Weldon의 게와 Pearson의 9차 방정식 (1894) 22

5.2 Old Faithful 간헐천 23

5.3 The Federalist의 “may”, 다시 24

6 부숴보기 — 무한대의 가능도와 성분의 수 26

6.1 가능도가 무한대로 가는 점 . . . 26

6.2 성분은 몇 개인가 . . . 26

6.3 느린 수렴 . . . 27

7 고치기 — Neyman–Scott를 숨은 변수로 29

8 Beautiful insights — 하한과 온도 32

8.1 ELBO와 변분 오토인코더 . . . 32

8.2 k-평균은 온도 0의 EM이다 . . . 32

8.3 숨은 상태 — Baum–Welch에서 전문가 혼합까지 . . . 33

9 정리와 역사 노트 35

9.1 한 장 요약 . . . 35

9.2 Stigler의 명명 법칙 . . . 35

9.3 7 Pillars 지도 . . . 36

9.4 다음 회 — 충분성: 자료를 줄이는 법 (1920 → 1955) . . . 36

A 연습문제 풀이 38

0

들어가며

> **여는 질문.** 나폴리 만의 게 1,000마리를 쟀더니 이마 폭의 분포가 한쪽으로 기울어 있다. 두 무리가 섞여 있는 것은 아닐까? 그런데 어느 게가 어느 무리인지는 아무도 모른다. 이름표가 없는 자료에서 두 무리의 평균과 비율을 추정할 수 있는가?

4회 5.3절에서 이 질문을 남겨 두었다. 동물학자 Weldon은 나폴리 만의 게 1,000마리에서 "이마"의 폭과 몸길이의 비를 쟀다. 분포의 왜도는 -0.50으로 정규에서 뚜렷이 벗어나 있었다. 그는 진화가 일어나는 중이라면, 곧 한 종이 두 무리로 갈라지는 중이라면 이런 모양이 나올 것이라 생각했고, Karl Pearson에게 분석을 부탁했다. Pearson(1894)은 두 정규분포의 혼합에 다섯 개의 모수를 놓고, 자료의 처음 다섯 적률을 모형의 적률과 같게 하는 방정식을 9차 방정식 하나로 줄여 손으로 풀었다. 근이 둘 나왔고, 그는 여섯째 적률로 하나를 골랐다.

이 노트는 같은 문제를 다르게 푼다. 각 게가 어느 무리에서 왔는지를 "보이지 않는 자료"로 보면, 보이지 않는 것을 지금의 추정값으로 채우고(E), 채운 자료로 모수를 다시 추정하기(M)를 되풀이할 수 있다. 이 생각은 McKendrick(1926), Hartley(1958), Baum(1970) 등이 여러 번 따로 찾아냈고, Dempster, Laird, Rubin(1977)이 하나의 이론으로 묶어 **EM 알고리즘**이라 불렀다.

먼저 9회의 이론을 여러 모수로 완성한다. 정보 행렬, 다변량 점근정규성, 귀찮은 모수가 있을 때의 프로파일 가능도를 다룬다(1절). 2절은 불완전 자료의 가능도를 세우고, 3절은 EM이 가능도를 결코 줄이지 않음을 Gibbs 부등식으로 증명한다. 수렴 속도가 "잃어버린 정보의 비율"로 정해진다는 것도 3절에서 보인다. 4절은 정규 혼합의 EM을 유도하고, 5절은 Pearson의 게, Old Faithful 간헐천, *The Federalist*에 쓴다. 6절은 혼합 가능도가 무한대로 가는 특이점과 성분 수 검정의 어려움을 보고, 7절은 9회의 Neyman–Scott 문제를 EM으로 고친다. 8절에서는 같은 하한이 변분 오토인코더의 ELBO가 되고, k-평균이 "딱딱한" EM이 됨을 본다.

학습 목표

1. 정보 행렬과 다변량 점근정규성을 증명하고, 귀찮은 모수가 분산을 키움을 Schur 보수로 보인다.
2. 프로파일 가능도를 정의하고, 귀찮은 모수가 있을 때의 Wilks 정리를 증명한다.
3. 불완전 자료의 가능도를 세우고, EM의 단조성과 ELBO 분해를 Gibbs 부등식으로 증명한다.

4. 잃어버린 정보의 원리(Louis 공식)와 EM의 선형 수렴 속도를 유도하고 수치로 확인한다.
5. 정규 혼합의 EM을 유도·구현해 Pearson의 적률법과 비교하고, 특이점과 성분 수 검정의 함정을 본다.

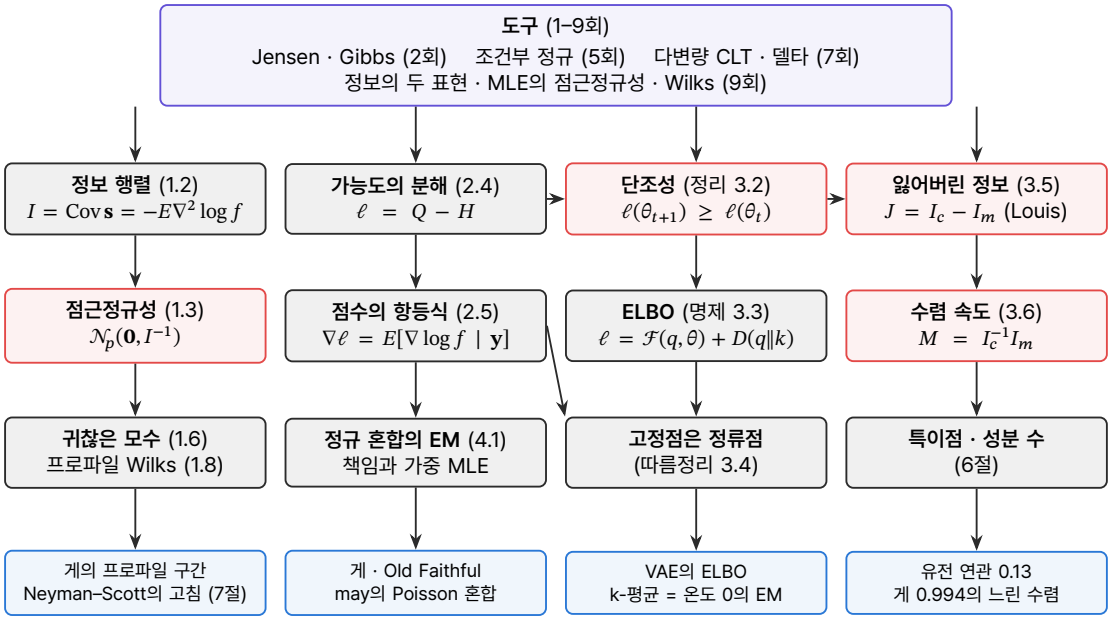


그림 0.1 이번 회차 정리들의 의존 관계. 화살표는 "증명에 쓰인다"는 뜻이다.

0.1 1–9회에서 가져오는 것

앞 회차	내용
2회 정리 5.9 · 8.3	Jensen 부등식, Gibbs 부등식 ($D(p\|q) \geq 0$, 등호는 $p = q$일 때만)
4회 5.3절 · 7.4절	정규의 혼합(Pearson의 게), 이름표 바꾸기, 두 정규의 혼합이 봉우리 하나일 수 있음(4회 연습 7.3)
7회 정리 5.2 · 5.4	다변량 중심극한정리, 다변량 델타 방법
9회 정리 2.2 · 3.1 · 5.2	정보의 두 표현, MLE의 점근정규성, Wilks 정리
9회 예제 7.2	Neyman–Scott: 모수가 자료와 함께 늘면 MLE가 일치하지 않는다

0.2 표기

모수는 벡터 $\theta \in \Omega \subset \mathbb{R}^p$이고, 한 관측의 점수 벡터는 $\mathbf{s}(x;\theta) = \nabla_\theta \log f(x;\theta)$, 로그가능도의 기울기와 헤세 행렬은 $\nabla\ell$, $\nabla^2\ell$로 쓴다. 2절부터는 **관측한 자료 $\mathbf{Y}$**와 **보이지 않는 자료 $\mathbf{Z}$**를 나누고,

둘을 합친 $(\mathbf{Y}, \mathbf{Z})$를 **완전 자료**라 부른다. 완전 자료의 밀도는 $f(\mathbf{y}, \mathbf{z}; \theta)$, 관측 자료의 밀도는 $g(\mathbf{y}; \theta)$, 관측이 주어질 때 숨은 자료의 조건부 밀도는 $k(\mathbf{z} \mid \mathbf{y}; \theta)$이다.

1

여러 모수의 가능도

> **역사** Pearson의 게는 모수가 다섯 개였다. Fisher(1922)는 **통계학의 수학적 기초**에서 Pearson의 적률법이 모수가 여럿일 때 비효율적일 수 있음을 지적하고, 최대가능도로 모든 모수를 한꺼번에 추정하자고 했다. 여러 모수의 정보는 행렬이 되고, 관심 없는 모수(**귀찮은 모수**, nuisance parameter)를 모르는 대가가 그 행렬의 역행렬에 나타난다. 9회 3절의 상자에서 미뤄 둔 증명을 여기서 한다.

1.1 정보 행렬과 점근정규성

정의 1.1 (점수 벡터와 정보 행렬) $\theta \in \mathbb{R}^p$일 때 한 관측의 **점수 벡터**는 $\mathbf{s}(x;\theta) = \nabla_\theta \log f(x;\theta)$이고, **정보 행렬**은 $p \times p$ 행렬 $I(\theta) = E_\theta\big[\mathbf{s}(X;\theta)\,\mathbf{s}(X;\theta)^{\prime}\big]$이다.

정리 1.2 (정보 행렬의 두 표현) 9회 2.1절의 정칙조건을 각 성분에 대해 가정하면 $E\,\mathbf{s} = \mathbf{0}$이고

$$I(\theta) = \operatorname{Cov}_\theta \mathbf{s}(X;\theta) = -E_\theta\big[\nabla^2_\theta \log f(X;\theta)\big].$$

iid 표본 n개의 정보 행렬은 $nI(\theta)$이다.

증명. **전략** 9회 정리 2.2의 증명을 성분별로 한다. $\int f\,dx = 1$을 θ_j로 한 번, 다시 θ_k로 한 번 미분하고 기댓값으로 읽는다 [T6]. 단계별 증명은 연습 1.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 39쪽에 있다.

$I(\theta)$는 공분산 행렬이므로 양의 준정부호다. 이 절에서는 참값에서 양의 정부호라고 가정한다. 그렇지 않으면 어떤 방향 $\mathbf{a}$에서 $\mathbf{a}'\mathbf{s} = 0$이어서, 그 방향으로 모수를 움직여도 자료의 분포가 (국소적으로) 변하지 않는다. 식별가능성이 국소적으로 깨진 것이다.

정리 1.3 (여러 모수의 점근정규성) 정칙조건(각 성분에 대한 9회 (R1)–(R5))과 $I(\theta_0) \succ 0$ 아래에서, $\hat\theta_n$이 가능도 방정식 $\nabla \ell(\theta) = \mathbf{0}$의 일치하는 근이면

$$\sqrt{n}\,(\hat\theta_n - \theta_0) \xrightarrow{\;D\;} \mathcal{N}_p\big(\mathbf{0},\; I(\theta_0)^{-1}\big).$$

증명. **전략** 가능도 방정식을 참값 주위에서 전개해 $\sqrt{n}\,(\hat\theta_n - \theta_0)$를 "곡률의 역행렬 × 점수의 합"으로 쓴다. 점수의 합에는 중심극한정리를, 곡률에는 대수의 법칙을 쓰고 Slutsky로 잇는다. 벡터 함수에는 성분마다 다른 중간점이 필요하므로 평균값 정리 대신 적분형 전개를 쓴다.

① $\mathbf{d} = \hat\theta_n - \theta_0$로 두면

$$\mathbf{0} = \nabla \ell(\hat\theta_n) = \nabla \ell(\theta_0) + H_n \mathbf{d}, \qquad H_n = \int_0^1 \nabla^2 \ell(\theta_0 + u\mathbf{d})\, du.$$ 미적분의 기본정리, 성분마다

② $n^{-1/2}\nabla \ell(\theta_0) = n^{-1/2}\sum_i \mathbf{s}(X_i;\theta_0) \xrightarrow{D} \mathcal{N}_p(\mathbf{0}, I(\theta_0))$이다. 점수는 iid이고 평균 $\mathbf{0}$, 공분산 $I(\theta_0)$이다. 정리 1.2, 7회 정리 5.2

③ $-n^{-1}\nabla^2 \ell(\theta_0) \xrightarrow{P} I(\theta_0)$이다. 성분마다 평균이 $I_{jk}(\theta_0)$인 iid의 평균이다. 정리 1.2, 7회 정리 1.6

④ (R5)의 셋째 미분 상계로, C가 p에만 의존하는 상수일 때

$$\left\| n^{-1}H_n - n^{-1}\nabla^2 \ell(\theta_0) \right\| \le \left(\tfrac{1}{n}\textstyle\sum_i M(X_i)\right) \cdot C\|\mathbf{d}\| \xrightarrow{P} 0.$$ $\mathbf{d} \xrightarrow{P} \mathbf{0}$, 7회 정리 1.6

⑤ $-n^{-1}H_n \xrightarrow{P} I(\theta_0)$이다. ③ ④, 7회 따름정리 1.4

⑥ 확률이 1로 가는 사건 위에서 $-n^{-1}H_n$이 가역이고 $(-n^{-1}H_n)^{-1} \xrightarrow{P} I(\theta_0)^{-1}$이다. 행렬의 역은 가역인 점에서 연속이다. 7회 정리 1.3

⑦ ①을 풀어 ②와 ⑥을 잇는다.

$$\sqrt{n}\,\mathbf{d} = \left(-n^{-1}H_n\right)^{-1} n^{-1/2}\nabla \ell(\theta_0) \xrightarrow{D} \mathcal{N}_p\!\left(\mathbf{0},\; I^{-1} I\, I^{-1}\right) = \mathcal{N}_p(\mathbf{0}, I^{-1}).$$

7회 정리 4.2, 5회 정리 3.8

조건 확인 (R1)–(R5)는 ②에서 점수의 평균이 $\mathbf{0}$이 되게(적분 기호 아래의 미분 [T6]), ④에서 참값의 한 근방에서 $|\partial^3 \log f| \le M(x)$, $EM(X) < \infty$가 되게 한다. 일치성 $\mathbf{d} \xrightarrow{P} \mathbf{0}$은 정리의 가정이다. $I(\theta_0) \succ 0$은 ⑥의 가역성에 쓰인다. ∎

> **따름정리 1.4** (신뢰영역과 함수) (a) $n(\hat\theta_n - \theta_0)' I(\hat\theta_n)(\hat\theta_n - \theta_0) \xrightarrow{D} \chi^2(p)$이므로, 이 이차형식이 $\chi^2_{p,\alpha}$ 이하인 θ_0들이 근사 $100(1-\alpha)\%$ 신뢰타원체다. (b) $g : \mathbb{R}^p \to \mathbb{R}$가 미분가능하고 $\nabla g(\theta_0) \neq \mathbf{0}$이면 $\sqrt{n}\big(g(\hat\theta_n) - g(\theta_0)\big) \xrightarrow{D} \mathcal{N}\big(0,\; \nabla g' I^{-1} \nabla g\big)$.

증명. **전략** (a)는 극한 분포 $\mathcal{N}_p(\mathbf{0}, I^{-1})$를 $I^{1/2}$로 표준화해 독립 표준정규의 제곱합으로 바꾸고, $I(\hat\theta_n)$을 Slutsky로 바꿔 넣는다. (b)는 다변량 델타 방법이다. 단계별 증명은 연습 1.7(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 39쪽에 있다.

1.2 귀찮은 모수의 대가

모수를 $\theta = (\psi, \lambda)$로 나눈다. $\psi \in \mathbb{R}^r$이 관심 모수, λ가 귀찮은 모수다. 정보 행렬과 그 역행렬도 같은 꼴로 나눠 $I = \left(\begin{smallmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{smallmatrix}\right)$, $I^{-1} = \left(\begin{smallmatrix} I^{\psi\psi} & I^{\psi\lambda} \\ I^{\lambda\psi} & I^{\lambda\lambda} \end{smallmatrix}\right)$로 쓴다.

보조정리 1.5 (블록 역행렬과 Schur 보수) $I \succ 0$이면 $I^{\psi\psi} = \left(I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}\right)^{-1}$이다. 또 모든 $\mathbf{u} \in \mathbb{R}^r$에 대해

$$\min_{\mathbf{v}} \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}' I \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} = \mathbf{u}'\left(I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}\right)\mathbf{u},$$

최소점은 $\mathbf{v} = -I_{\lambda\lambda}^{-1} I_{\lambda\psi}\mathbf{u}$이다.

증명. **전략** 이차형식을 $\mathbf{v}$에 대해 완전제곱한다. 남는 항이 Schur 보수이고, 블록 역행렬은 곱해서 확인한다.

① $I_{\lambda\lambda} \succ 0$이므로 가역이다. $S = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}$로 둔다. — $I \succ 0$의 주대각 블록

② I가 대칭이므로 $I_{\psi\lambda} = I_{\lambda\psi}'$이고

$$\begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}' I \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} = \mathbf{u}' I_{\psi\psi}\mathbf{u} + 2\mathbf{v}' I_{\lambda\psi}\mathbf{u} + \mathbf{v}' I_{\lambda\lambda}\mathbf{v}.$$ — 블록 곱셈

③ $\mathbf{w} = \mathbf{v} + I_{\lambda\lambda}^{-1} I_{\lambda\psi}\mathbf{u}$로 두면 ②의 오른쪽은 $\mathbf{u}' S\mathbf{u} + \mathbf{w}' I_{\lambda\lambda}\mathbf{w}$이다. — 완전제곱, 전개해 확인 [T4]

④ $\mathbf{w}' I_{\lambda\lambda}\mathbf{w} \geq 0$이고 등호는 $\mathbf{w} = \mathbf{0}$, 곧 $\mathbf{v} = -I_{\lambda\lambda}^{-1} I_{\lambda\psi}\mathbf{u}$일 때뿐이다. 따라서 최솟값은 $\mathbf{u}' S\mathbf{u}$이다. — $I_{\lambda\lambda} \succ 0$

⑤ $\mathbf{u} \neq \mathbf{0}$이면 $\mathbf{u}' S\mathbf{u}$는 영이 아닌 벡터에서의 I의 이차형식이므로 양수다. 곧 $S \succ 0$이고 가역이다. — ④, $I \succ 0$

⑥ $I\begin{pmatrix} S^{-1} \\ -I_{\lambda\lambda}^{-1} I_{\lambda\psi} S^{-1}\end{pmatrix} = \begin{pmatrix}\mathbf{1}_r \\ \mathbf{0}\end{pmatrix}$이므로 이것이 I^{-1}의 첫 블록 열이고, $I^{\psi\psi} = S^{-1}$이다. — 블록 곱셈(연습 1.2)

■

정리 1.6 (귀찮은 모수는 분산을 키운다) $I \succ 0$이면 $I^{\psi\psi} \succeq I_{\psi\psi}^{-1}$이고, 등호는 $I_{\psi\lambda} = \mathbf{0}$일 때뿐이다. 곧 $\boldsymbol{\lambda}$를 모를 때 $\hat{\boldsymbol{\psi}}$의 점근분산 $I^{\psi\psi}/n$은 $\boldsymbol{\lambda}$를 알 때의 $I_{\psi\psi}^{-1}/n$보다 작지 않다.

증명. **전략** Schur 보수 S가 $I_{\psi\psi}$보다 작음을 보이고, 양의 정부호 행렬에서 역행렬을 취하면 순서가 뒤집힘을 쓴다.

① 모든 $\mathbf{u}$에서 $\mathbf{u}' I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi}\mathbf{u} = (I_{\lambda\psi}\mathbf{u})' I_{\lambda\lambda}^{-1} (I_{\lambda\psi}\mathbf{u}) \geq 0$이다. — $I_{\lambda\lambda}^{-1} \succ 0$

② 따라서 $S = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi} \preceq I_{\psi\psi}$이다. — ①

③ $A \succeq B \succ 0$이면 $A^{-1} \preceq B^{-1}$이다. $B^{-1/2} A B^{-1/2} \succeq \mathbf{1}$의 고유값은 모두 1 이상이므로 그 역행렬 $B^{1/2} A^{-1} B^{1/2}$의 고유값은 모두 1 이하이고, 양쪽에 $B^{-1/2}$를 곱해도 순서가 유지된다. — 고유분해, 합동변환

④ $I^{\psi\psi} = S^{-1} \succeq I_{\psi\psi}^{-1}$이다. — 보조정리 1.5, ② ③

⑤ 등호이면 $S = I_{\psi\psi}$이므로 모든 $\mathbf{u}$에서 ①의 값이 0이고, $I_{\lambda\lambda}^{-1} > 0$이므로 $I_{\lambda\psi}\mathbf{u} = \mathbf{0}$, 곧 $I_{\lambda\psi} = \mathbf{0}$이다. 거꾸로 $I_{\lambda\psi} = \mathbf{0}$이면 $S = I_{\psi\psi}$이다. ①

■

$I_{\psi\lambda} = \mathbf{0}$이면 두 모수가 **직교한다**고 한다. 정규의 (μ, σ^2)은 직교한다(9회 연습 3.4). 그래서 σ^2을 몰라도 $\bar{X}$의 점근분산은 그대로다. 감마의 (α, β)는 직교하지 않는다. $\alpha = 2$에서 β를 알면 $\hat{\alpha}$의 점근분산이 $1/\psi'(2) = 1.55$를 n으로 나눈 것이지만, 모르면 $6.90/n$으로 4.4배가 된다(9회 연습 3.3). 정규 혼합에서는 비율과 평균, 분산이 서로 강하게 얽혀 있어서 이 대가가 아주 크다(5절).

1.3 프로파일 가능도

정의 1.7 (프로파일 가능도) $\boldsymbol{\psi}$를 고정하고 귀찮은 모수에 대해 최대화한

$$\ell_p(\boldsymbol{\psi}) = \sup_{\boldsymbol{\lambda}} \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \ell\big(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}\big)$$

을 $\boldsymbol{\psi}$의 **프로파일 로그가능도**라 한다. $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$는 $\boldsymbol{\psi}$를 고정했을 때의 MLE다.

ℓ_p의 최대점은 $\hat{\boldsymbol{\psi}}$이다(전체 최대점의 $\boldsymbol{\psi}$ 성분). 산의 능선을 따라 걸으며 높이를 재는 것과 같다. 정규 혼합처럼 모수가 많은 모형에서 한 모수의 불확실성을 보는 가장 간단한 그림이다.

정리 1.8 (귀찮은 모수가 있을 때의 Wilks 정리) 정리 1.3의 조건과 $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ 아래에서, 제약된 MLE $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}_0}$도 일치하면

$$\Lambda = 2\big[\ell(\hat{\boldsymbol{\theta}}_n) - \ell_p(\boldsymbol{\psi}_0)\big] \xrightarrow{D} \chi^2(r).$$

증명. **전략** MLE 주위의 이차 전개로 로그가능도를 이차형식으로 바꾼다. 귀찮은 모수에 대한 최대화는 그 이차형식의 최소화이고, 보조정리 1.5가 답을 준다.

① $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_n$에서 $\nabla \ell = \mathbf{0}$이므로, $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\| \le K/\sqrt{n}$인 $\boldsymbol{\theta}$들 위에서 균등하게

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) - \tfrac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' I(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_P(1).$$ 적분형 Taylor [T8], 정리 1.3의 ③–⑤

② H_0 아래에서 $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 = O_P(n^{-1/2})$이고, 제약된 최대점 $(\boldsymbol{\psi}_0, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}_0})$도 $\hat{\boldsymbol{\theta}}$에서 $O_P(n^{-1/2})$ 안에 있다. 정리 1.3, 같은 전개

③ 따라서 $\boldsymbol{\lambda}$에 대한 최대화는 ①의 이차형식을 $\mathbf{v} = \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}$에 대해 최소화하는 것이다. $\mathbf{u} = \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}}$로 두면

$$\ell_p(\boldsymbol{\psi}_0) = \ell(\hat{\boldsymbol{\theta}}) - \tfrac{n}{2} \min_{\mathbf{v}} \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}' I \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix} + o_P(1).$$ ①②

④ 최솟값은 $\mathbf{u}'S\mathbf{u}$이고 $S = (I^{\psi\psi})^{-1}$이다. 보조정리 1.5

⑤ $\Lambda = 2[\ell(\hat\theta) - \ell_p(\boldsymbol\psi_0)] = n(\hat{\boldsymbol\psi} - \boldsymbol\psi_0)' S (\hat{\boldsymbol\psi} - \boldsymbol\psi_0) + o_P(1)$이다. ③④

⑥ $\sqrt{n}(\hat{\boldsymbol\psi} - \boldsymbol\psi_0) \xrightarrow{D} \mathcal{N}_r(\mathbf{0}, I^{\psi\psi})$이다. 정리 1.3의 앞 r개 성분

⑦ $\mathcal{N}_r(\mathbf{0}, I^{\psi\psi})$를 $S^{1/2}$로 표준화하면 $\Lambda \xrightarrow{D} \chi^2(r)$이다. 따름정리 1.4(a)의 논증, 7회 정리 4.2

조건 확인 ①의 균등한 전개에는 정리 1.3의 셋째 미분 상계가 참값의 한 근방 전체에서 필요하다. 제약된 MLE의 일치성은 정리의 가정이고, 이것이 ②를 준다. ∎

자유도는 "고정한 모수의 수" r이다. 귀찮은 모수는 자유도를 쓰지 않는다. 그 대가는 이미 $S = (I^{\psi\psi})^{-1} \preceq I_{\psi\psi}$, 곧 더 넓은 구간으로 치렀다.

예제 1.9 (정규 평균의 프로파일 가능도는 t 검정이다) $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$에서 μ를 고정하면 $\hat\sigma^2_\mu = \frac{1}{n}\sum (x_i - \mu)^2 = \hat\sigma^2 + (\bar x - \mu)^2$이고 $\ell_p(\mu) = -\frac{n}{2}\log \hat\sigma^2_\mu - \frac{n}{2}$이다. 따라서

$$\Lambda(\mu_0) = n \log\!\left(1 + \frac{(\bar x - \mu_0)^2}{\hat\sigma^2}\right) = n \log\!\left(1 + \frac{t^2}{n-1}\right), \qquad t = \frac{\bar x - \mu_0}{s/\sqrt{n}}.$$

Λ는 $|t|$의 증가함수이므로 가능도비 검정은 6회의 t 검정과 같은 기각역을 갖는다. 다만 임계값을 $\chi^2(1)$에서 가져오면 근사, $t(n-1)$에서 가져오면 정확하다. 9회 예제 6.2의 Michelson 구간이 이 계산이었다.

가정 주의 프로파일 가능도는 진짜 가능도가 아니다. 귀찮은 모수를 그 MLE로 바꿔 넣었으므로 프로파일 점수 $\frac{d}{d\psi}\ell_p$의 기댓값이 정확히 0이 아니고, 그 치우침은 귀찮은 모수 하나당 $O(1)$씩 쌓인다. 귀찮은 모수의 수가 고정되어 있으면 $\sqrt{n}$ 척도에서 사라지지만, 9회의 Neyman–Scott처럼 자료와 함께 늘면 프로파일 가능도의 최대점이 일치하지 않는다. 7절에서 귀찮은 모수를 "숨은 변수"로 바꿔 이 문제를 푼다.

연습문제

연습 1.1 (정규의 정보 행렬 ★) $\mathcal{N}(\mu, \sigma^2)$을 (μ, σ)로 모수화하면 정보 행렬이 $\mathrm{diag}(1/\sigma^2,\ 2/\sigma^2)$임을 보여라. 9회 연습 2.2의 변환 규칙($\sigma^2 \mapsto \sigma$)과 맞는가? 풀이 38쪽

연습 1.2 (블록 역행렬 ★) 보조정리 1.5의 블록 역행렬 공식을 $I \cdot I^{-1}$의 첫 블록 열을 곱해 확인하라. 풀이 38쪽

연습 1.3 (변동계수 ★★) 정규 표본에서 변동계수 $c = \sigma/\mu$의 MLE $\hat c = \hat\sigma/\bar X$의 점근분산이 $\frac{c^2(c^2 + 1/2)}{n}$임을 따름정리 1.4(b)로 보여라. 풀이 38쪽

연습 1.4 (감마의 프로파일 가능도 ★★) 감마(α,β)에서 α를 고정한 β의 MLE가 $\bar{x}/\alpha$이고, $\ell_p(\alpha)=n\big[(\alpha-1)\overline{\log x}-\log\Gamma(\alpha)-\alpha\log(\bar{x}/\alpha)-\alpha\big]$임을 보여라. $\ell_p'(\alpha)=0$이 9회 예제 4.4의 방정식 $\log\alpha-\psi(\alpha)=\log\bar{x}-\overline{\log x}$임을 확인하라. 풀이 38쪽

연습 1.5 (두 표본의 평균 차 ★★) $X_i\overset{\text{iid}}{\sim}\mathcal{N}(\mu_1,\sigma^2)$ $(i\le m)$, $Y_j\overset{\text{iid}}{\sim}\mathcal{N}(\mu_2,\sigma^2)$ $(j\le n)$에서 $\psi=\mu_1-\mu_2$의 프로파일 가능도비가 $(m+n)\log\!\big(1+\frac{t^2}{m+n-2}\big)$임을 보여라. t는 합동분산을 쓴 두 표본 t 통계량이다(6회). 풀이 38쪽

연습 1.6 (정보 행렬의 두 표현 증명 ★ 유도형) 정리 1.2를 다음 순서로 증명하라. 미분과 적분의 순서를 바꾸는 곳마다 9회의 정칙조건 가운데 무엇을 쓰는지 밝혀라.

(a) $\int f(x;\theta)\,dx=1$을 θ_j로 미분해 $E_\theta s_j=0$을 보여라.

(b) (a)의 식 $\int s_j f\,dx=0$을 다시 θ_k로 미분해 $E\big[\frac{\partial^2\log f}{\partial\theta_j\partial\theta_k}\big]+E[s_j s_k]=0$을 보여라.

(c) (a)로 $E\,\mathbf{s}\mathbf{s}'=\operatorname{Cov}\mathbf{s}$임을 보이고 두 표현을 얻어라.

(d) iid 표본의 점수는 $\sum_i\mathbf{s}(X_i;\theta)$이다. 독립인 벡터의 합의 공분산이 더해짐을 써서 정보 행렬이 $nI(\theta)$임을 보여라.

풀이 39쪽

연습 1.7 (신뢰영역과 함수의 증명 ★★ 유도형) 따름정리 1.4를 다음 순서로 증명하라. $I=I(\theta_0)$로 쓴다.

(a) $\mathbf{W}\sim\mathcal{N}_p(\mathbf{0},I^{-1})$이면 $I^{1/2}\mathbf{W}\sim\mathcal{N}_p(\mathbf{0},\mathbf{1})$이고 $\mathbf{W}'I\mathbf{W}\sim\chi^2(p)$임을 보여라(5회 정리 3.8, 정리 3.17).

(b) $\sqrt{n}(\hat\theta_n-\theta_0)\xrightarrow{D}\mathbf{W}$와 $I(\hat\theta_n)\xrightarrow{P}I$에서 (a)의 이차형식이 $\chi^2(p)$로 분포수렴함을 보여라. $I(\cdot)$의 연속성은 어디서 쓰는가?

(c) (b)의 식에서 θ_0을 미지수로 보고 근사 신뢰타원체를 얻어라.

(d) 7회 정리 5.4(다변량 델타 방법)를 g에 써서 따름정리의 (b)를 보여라. $\nabla g(\theta_0)\neq\mathbf{0}$은 어디에 필요한가?

풀이 39쪽

2

보이지 않는 자료의 가능도

> **역사** "보이지 않는 것을 채워 넣고 다시 추정한다"는 생각은 여러 번 따로 태어났다. McKendrick(1926)은 전염병 자료에서 관측되지 않은 칸을 채우며 모수를 고쳐 갔고, Healy와 Westmacott(1956)는 실험 설계에서 빠진 관측값을 같은 방법으로 메웠다. Hartley(1958)는 불완전한 셈 자료의 가능도를 이렇게 최대화했고, Baum과 동료들(1970)은 은닉 Markov 모형에서 같은 반복을 찾아 음성 인식의 바탕을 놓았다. Orchard와 Woodbury(1972), Sundberg(1974)는 "잃어버린 정보"의 원리를 정리했다. Dempster, Laird, Rubin(1977)은 이 모든 것이 하나의 알고리즘임을 보이고 EM이라는 이름을 붙였다. 그들의 첫 예제는 Rao(1965)의 교과서에 나온 유전 연관 자료였다.

2.1 완전 자료와 관측 자료

정의 2.1 (불완전 자료의 가능도) 완전 자료 $(\mathbf{Y}, \mathbf{Z})$의 밀도를 $f(\mathbf{y}, \mathbf{z}; \theta)$라 하자. $\mathbf{Y} = \mathbf{y}$만 관측하면 가능도는 숨은 자료를 적분(이산이면 합)해 없앤

$$L(\theta) = g(\mathbf{y}; \theta) = \int f(\mathbf{y}, \mathbf{z}; \theta)\, d\mathbf{z}, \qquad \ell(\theta) = \log g(\mathbf{y}; \theta)$$

이다. 관측이 주어질 때 숨은 자료의 조건부 밀도는 $k(\mathbf{z} \mid \mathbf{y}; \theta) = f(\mathbf{y}, \mathbf{z}; \theta) / g(\mathbf{y}; \theta)$이다. $\ell_c(\theta) = \log f(\mathbf{y}, \mathbf{z}; \theta)$를 **완전 자료 로그가능도**라 한다.

어려움은 적분이 로그 안에 있다는 데 있다. 완전 자료가 있으면 ℓ_c는 대개 다루기 쉽다(지수족이면 충분통계량의 함수다). 관측 자료의 ℓ은 합의 로그라서 미분이 얽힌다.

예제 2.2 (숨은 자료의 네 얼굴)

(a) **혼합분포.** $Z_i \in \{1, \dots, K\}$는 관측 Y_i가 나온 성분이다. $P(Z_i = k) = \pi_k$, $Y_i \mid Z_i = k \sim f_k(\cdot; \theta_k)$. 관측 가능도는 $\prod_i \sum_k \pi_k f_k(y_i; \theta_k)$이고, 이름표를 알면 $\ell_c = \sum_i \sum_k \mathbb{1}\{z_i = k\}\big[\log \pi_k + \log f_k(y_i; \theta_k)\big]$로 성분마다 따로 떨어진다.

(b) **유전 연관.** 197마리의 동물이 네 범주에 $\mathbf{y} = (125, 18, 20, 34)$로 나뉘었고, 범주의 확률은 $\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$이다. 첫 범주를 확률 $\frac{1}{2}$과 $\frac{\theta}{4}$의 두 칸 (x_1, x_2)로 나눴다고 보면 숨은 자료는 x_2이고, 완전 자료에서 θ는 $x_2 + y_4$번 성공, $y_2 + y_3$번 실패한 이항 모형의 비율이 된다.

(c) **중도절단.** 수명 T_i를 시각 c까지만 지켜보면 $T_i > c$인 개체는 "c보다 길다"는 것만 안다.

숨은 자료는 절단된 개체의 실제 수명이다.

(d) **임의효과.** $X_{ij} = \mu_i + \varepsilon_{ij}$에서 사람마다의 평균 μ_i를 관측하지 못하는 확률변수 $\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(m, \tau^2)$로 본다. 9회의 Neyman–Scott 문제를 이렇게 바꾸면 모수의 수가 n과 함께 늘지 않는다(7절).

예제 2.3 (유전 연관 자료의 MLE) 관측 로그가능도는 $\ell(\theta) = 125\log(2+\theta) + 38\log(1-\theta) + 34\log\theta +$상수이고, 가능도 방정식 $\frac{125}{2+\theta} - \frac{38}{1-\theta} + \frac{34}{\theta} = 0$에 분모를 곱하면 $197\theta^2 - 15\theta - 68 = 0$이 된다(연습 2.1). 양의 근은 $\hat\theta = \frac{15+\sqrt{53809}}{394} = 0.6268$이다. 닫힌 꼴이 있으므로 EM이 필요하지 않지만, 무엇이 일어나는지 숫자로 하나하나 확인할 수 있어서 3절의 시험대로 쓴다.

2.2 가능도의 분해

보조정리 2.4 (가능도의 분해) $k(\mathbf{z} \mid \mathbf{y}; \theta) > 0$인 모든 $\mathbf{z}$에서 $\ell(\theta) = \log f(\mathbf{y}, \mathbf{z}; \theta) - \log k(\mathbf{z} \mid \mathbf{y}; \theta)$이다. 따라서 임의의 θ'에 대해

$$\ell(\theta) = Q(\theta \mid \theta') - H(\theta \mid \theta'), \qquad \begin{aligned} Q(\theta \mid \theta') &= E_{\theta'}\big[\log f(\mathbf{y}, \mathbf{Z}; \theta) \mid \mathbf{y}\big], \\ H(\theta \mid \theta') &= E_{\theta'}\big[\log k(\mathbf{Z} \mid \mathbf{y}; \theta) \mid \mathbf{y}\big]. \end{aligned}$$

기댓값은 $\mathbf{Z} \sim k(\cdot \mid \mathbf{y}; \theta')$에 대한 것이다.

증명. 첫 등식은 k의 정의 $f = g \cdot k$에 로그를 취한 것이다. 왼쪽은 $\mathbf{z}$에 의존하지 않으므로 양변을 $k(\mathbf{z} \mid \mathbf{y}; \theta')$로 평균해도 그대로다. ∎

Q는 "지금의 추정값 θ'로 숨은 자료를 추측했을 때의 완전 자료 로그가능도의 기댓값"이다. 3절의 EM은 이 Q를 θ에 대해 최대화한다. 분해의 쓸모는 다음 두 사실에서 나온다. H는 $\theta = \theta'$에서 최대이고(Gibbs, 3절), 점수는 완전 자료 점수의 조건부 기댓값이다.

명제 2.5 (점수의 항등식) 적분과 미분의 순서를 바꿀 수 있으면

$$\nabla\ell(\theta) = E_\theta\big[\nabla_\theta \log f(\mathbf{y}, \mathbf{Z}; \theta) \mid \mathbf{y}\big] = \nabla_\theta Q(\theta \mid \theta')\Big|_{\theta' = \theta}.$$

증명. **전략** 관측 밀도 g를 완전 밀도의 적분으로 쓰고 적분 기호 아래에서 미분한다 [T6]. 그다음 $f/g = k$로 읽으면 조건부 기댓값이 된다. 단계별 증명은 연습 2.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 40쪽에 있다.

명제는 "보이지 않는 자료의 점수는, 볼 수 있었다면의 점수를 지금 아는 것으로 평균한 것"이라고 말한다. 혼합분포에서는 $\frac{\partial \ell}{\partial \mu_k} = \sum_i r_{ik} \frac{y_i - \mu_k}{\sigma_k^2}$이다. 여기서 $r_{ik} = P(Z_i = k \mid y_i)$는 관측 i가

성분 k에서 왔을 사후확률이다(연습 2.4). 이름표를 알면 성분 k의 관측들만 더할 것을, 모르니 사후확률로 나눠 더한다.

연습문제

연습 2.1 (유전 연관의 이차방정식 ★) 예제 2.3의 가능도 방정식이 $197\theta^2 - 15\theta - 68 = 0$이 됨을 보이고 $\hat\theta = 0.6268$을 확인하라. 풀이 39쪽

연습 2.2 (숨은 칸의 조건부분포 ★) 유전 연관 자료에서 $x_2 \mid \mathbf{y} \sim \mathrm{b}\big(125,\ \frac{\theta}{2+\theta}\big)$임을 보이고, 완전 자료의 MLE가 $\frac{x_2+34}{x_2+72}$임을 보여라. 풀이 39쪽

연습 2.3 (중도절단된 지수분포 ★★) $T_i \overset{\text{iid}}{\sim} \mathrm{Exp}$ (평균 θ)를 시각 c에서 절단해 $y_i = \min(T_i, c)$와 절단 여부를 본다. 절단되지 않은 개체가 r개이면 관측 가능도가 $\theta^{-r} e^{-\sum y_i/\theta}$이고 MLE가 $\sum y_i / r$임을 보여라. 무기억성(4회)으로 $T_i \mid T_i > c$의 분포가 $c + \mathrm{Exp}(\theta)$임을 쓰라. 풀이 39쪽

연습 2.4 (혼합의 점수 ★★) 정규 혼합에서 $\frac{\partial \ell}{\partial \mu_k} = \sum_i r_{ik} \frac{y_i - \mu_k}{\sigma_k^2}$를 직접 미분으로 보이고, 명제 2.5로도 보여라. 풀이 40쪽

연습 2.5 (묶음 자료 ★★) Pearson의 게는 0.004 폭의 칸으로 묶여 있다. 정규 $\mathcal{N}(\mu, \sigma^2)$에서 칸 $(a, b]$에 든 관측의 조건부 평균이 $\mu + \sigma \frac{\varphi(\alpha) - \varphi(\beta)}{\Phi(\beta) - \Phi(\alpha)}$ ($\alpha = \frac{a-\mu}{\sigma}$, $\beta = \frac{b-\mu}{\sigma}$)임을 보여라. 묶음 자료의 EM에서 E 단계가 계산하는 양이다. 풀이 40쪽

연습 2.6 (점수의 항등식 증명 ★★ 유도형) 명제 2.5를 다음 순서로 증명하라.

(a) $g(\mathbf{y};\theta) = \int f(\mathbf{y},\mathbf{z};\theta)\,d\mathbf{z}$를 적분 기호 아래에서 미분해 $\nabla g = \int (\nabla \log f)\, f\, d\mathbf{z}$를 보여라. 이 교환에 필요한 조건을 [T6]의 말로 적어라.

(b) $\nabla \ell = \nabla g / g$에 (a)를 넣고 $f/g = k$를 써서 첫 등식을 얻어라.

(c) $Q(\theta \mid \theta')$에서 가중치 $k(\cdot \mid \mathbf{y};\theta')$는 θ에 의존하지 않는다. 이것으로 둘째 등식을 얻어라.

(d) 유전 연관 자료(예제 2.3)의 완전 자료 점수는 $\frac{x_2+34}{\theta} - \frac{38}{1-\theta}$이다. 그 조건부 기댓값이 관측 점수 $\frac{125}{2+\theta} - \frac{38}{1-\theta} + \frac{34}{\theta}$와 같음을 확인하라.

풀이 40쪽

3

EM 알고리즘

3.1 알고리즘과 단조성

알고리즘 3.1 (EM (Dempster–Laird–Rubin, 1977)) 시작값 θ_0에서 $t = 0, 1, 2, \dots$에 대해 다음을 되풀이한다.

E 지금의 추정값으로 숨은 자료의 조건부분포 $k(\cdot \mid \mathbf{y}; \theta_t)$를 구하고 $Q(\theta \mid \theta_t) = E_{\theta_t}[\log f(\mathbf{y}, \mathbf{Z}; \theta) \mid \mathbf{y}]$를 만든다.

M $\theta_{t+1} = \arg\max_\theta Q(\theta \mid \theta_t)$.

$\ell(\theta_{t+1}) - \ell(\theta_t)$이 충분히 작아지면 멈춘다.

완전 자료 모형이 지수족이면 $\log f$는 충분통계량에 대해 선형이므로, E 단계는 "충분통계량의 조건부 기댓값을 구하기"이고 M 단계는 "그 기댓값을 완전 자료의 MLE 공식에 넣기"다. 유전 연관 자료에서는 E 단계가 $x_2^{(t)} = 125 \cdot \frac{\theta_t}{2+\theta_t}$, M 단계가 $\theta_{t+1} = \frac{x_2^{(t)}+34}{x_2^{(t)}+72}$이다(연습 2.2). 한 바퀴의 그림이 그림 3.1이다.

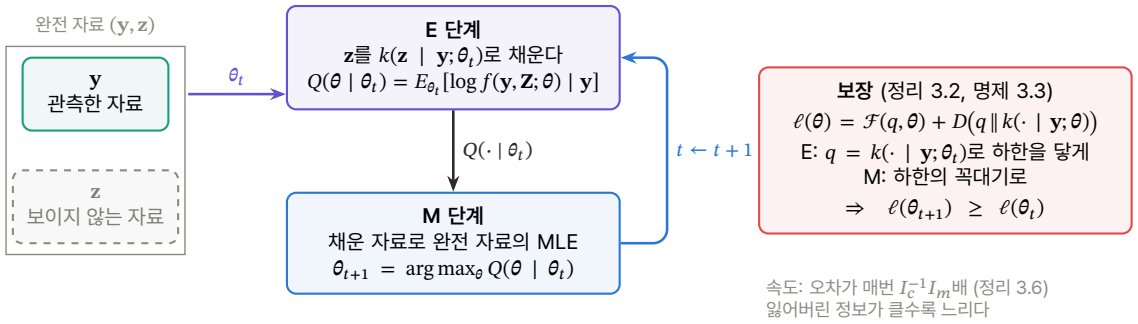


그림 3.1 EM의 한 바퀴. 완전 자료 가운데 $\mathbf{z}$가 보이지 않는다. E 단계는 지금의 추정값으로 $\mathbf{z}$의 조건부분포를 구해 Q를 만들고, M 단계는 Q를 최대화한다. 오른쪽은 이 반복이 가능도를 줄이지 않는 이유와 그 속도다.

정리 3.2 (EM은 가능도를 줄이지 않는다) EM의 반복은 $\ell(\theta_{t+1}) \geq \ell(\theta_t)$를 만족한다. 등호는 $Q(\theta_{t+1} \mid \theta_t) = Q(\theta_t \mid \theta_t)$이고 $k(\cdot \mid \mathbf{y}; \theta_{t+1}) = k(\cdot \mid \mathbf{y}; \theta_t)$일 때뿐이다.

증명. **전략** 보조정리 2.4로 가능도의 증가를 Q의 증가와 H의 변화로 나눈다. Q는 M 단계가 늘리고, H는 Gibbs 부등식으로 제자리에서 가장 크다.

① 보조정리 2.4를 $\theta' = \theta_t$로 두고 θ_{t+1}과 θ_t에 쓴 두 식을 빼면

$$\ell(\theta_{t+1}) - \ell(\theta_t) = \big[Q(\theta_{t+1} \mid \theta_t) - Q(\theta_t \mid \theta_t)\big] - \big[H(\theta_{t+1} \mid \theta_t) - H(\theta_t \mid \theta_t)\big]. \qquad \text{보조정리 2.4}$$

② 첫 괄호는 0 이상이다. θ_{t+1}은 $Q(\cdot \mid \theta_t)$의 최대점이다. M 단계

③ 둘째 괄호를 기댓값 하나로 쓴다.

$$H(\theta \mid \theta_t) - H(\theta_t \mid \theta_t) = E_{\theta_t}\!\left[\log \frac{k(\mathbf{Z} \mid \mathbf{y}; \theta)}{k(\mathbf{Z} \mid \mathbf{y}; \theta_t)} \,\middle|\, \mathbf{y}\right] = -D\big(k(\cdot \mid \mathbf{y}; \theta_t) \,\big\|\, k(\cdot \mid \mathbf{y}; \theta)\big).$$

H와 D의 정의

④ $D \geq 0$이므로 둘째 괄호는 0 이하이다. Gibbs 부등식, 2회 정리 8.3

⑤ 따라서 $\ell(\theta_{t+1}) - \ell(\theta_t) \geq 0$이다. ①②④

⑥ 등호는 두 괄호가 모두 0일 때, 곧 $Q(\theta_{t+1} \mid \theta_t) = Q(\theta_t \mid \theta_t)$이고 $k(\cdot \mid \mathbf{y}; \theta_{t+1}) = k(\cdot \mid \mathbf{y}; \theta_t)$일 때뿐이다. Gibbs의 등호 조건

조건 확인 보조정리 2.4의 첫 등식은 $k(\mathbf{z} \mid \mathbf{y}; \theta) > 0$인 곳에서 성립한다. $k(\cdot \mid \mathbf{y}; \theta_t) > 0$인 곳에서 $k(\cdot \mid \mathbf{y}; \theta_{t+1}) > 0$이어야 ①의 평균을 취할 수 있다. 숨은 자료의 받침이 모수에 의존하지 않으면 성립한다. Q와 H는 유한하다고 가정한다. ③의 계산은 9회 보조정리 1.4와 같다. ■

증명에서 M 단계가 한 일은 Q를 **늘린** 것뿐이다. 그래서 Q를 완전히 최대화하지 않고 늘리기만 해도(예: Newton 한 걸음) 단조성은 그대로다. 이것을 **일반화 EM**(GEM)이라 한다.

3.2 하한을 오르는 알고리즘

명제 3.3 (증거 하한 (ELBO)) $\mathbf{z}$ 위의 아무 확률밀도 q에 대해

$$\ell(\theta) = \mathcal{F}(q, \theta) + D\big(q \,\big\|\, k(\cdot \mid \mathbf{y}; \theta)\big), \qquad \mathcal{F}(q, \theta) = E_q\big[\log f(\mathbf{y}, \mathbf{Z}; \theta)\big] - E_q\big[\log q(\mathbf{Z})\big].$$

따라서 $\ell(\theta) \geq \mathcal{F}(q, \theta)$이고, 등호는 $q = k(\cdot \mid \mathbf{y}; \theta)$일 때다.

증명. **전략** 완전 밀도를 $f = g \cdot k$로 쓰고 아무 분포 q로 평균한 뒤, 양변에서 $E_q \log q$를 뺀다. 남는 항이 Kullback–Leibler 발산이다. 단계별 증명은 연습 3.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 41쪽에 있다.

이 명제는 EM을 한 함수 $\mathcal{F}(q,\theta)$를 번갈아 최대화하는 좌표 상승으로 보여 준다(Neal–Hinton, 1998). E 단계는 θ_t를 고정하고 q에 대해 최대화한다. 답은 $q = k(\cdot \mid \mathbf{y};\theta_t)$이고, 하한이 그 점에서 가능도에 닿는다. M 단계는 q를 고정하고 θ에 대해 최대화한다. $E_q \log q$는 θ에 의존하지 않으므로 이것은 $Q(\theta \mid \theta_t)$의 최대화다. 가능도 아래에 그 점에서 닿는 하한을 깔고, 하한의 꼭대기로 옮기고, 그 점에서 다시 하한을 까는 것이다(그림 3.2). 8절에서 q를 신경망으로 근사하면 변분 오토인코더가 된다.

따름정리 3.4 (EM의 고정점은 정류점이다) $\theta_t \to \theta^*$ (내점)이고 Q가 두 인수에 대해 연속미분 가능하면 $\nabla \ell(\theta^*) = \mathbf{0}$이다.

증명. **전략** M 단계의 1계 조건에서 $t \to \infty$로 극한을 취하고, 명제 2.5로 그 극한을 관측 점수로 읽는다. 단계별 증명은 연습 3.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 41쪽에 있다.

가정 주의 정류점은 극대점이 아닐 수도 있다. EM은 가능도를 늘리기만 하므로 극소점에 머무는 일은 드물지만(9회 4절의 Newton과 다르다), 국소 최대점이나 안장점에 멈출 수 있다. 원논문의 수렴 증명에는 틀린 단계가 있었고, Wu(1983)가 정확한 조건을 정리했다. 실전에서는 여러 시작점에서 돌려 가능도가 가장 높은 것을 고른다.

3.3 잃어버린 정보와 수렴 속도

숨은 자료를 보지 못해서 정보를 얼마나 잃었는가? 완전 자료의 정보에서 관측 자료의 정보를 뺀 것이다. 이 직관을 정확히 쓴 것이 다음 정리다.

정리 3.5 (잃어버린 정보의 원리 (Orchard–Woodbury 1972, Louis 1982)) 관측 정보 $J(\theta) = -\nabla^2 \ell(\theta)$는

$$J(\theta) = I_c(\theta \mid \mathbf{y}) - I_m(\theta \mid \mathbf{y})$$

로 쓰인다. 여기서 $I_c(\theta \mid \mathbf{y}) = -E_\theta\big[\nabla^2 \log f(\mathbf{y},\mathbf{Z};\theta) \mid \mathbf{y}\big]$는 완전 자료 정보의 조건부 기댓값이고,

$$I_m(\theta \mid \mathbf{y}) = -E_\theta\big[\nabla^2 \log k(\mathbf{Z} \mid \mathbf{y};\theta) \mid \mathbf{y}\big] = \operatorname{Cov}_\theta\big(\nabla \log f(\mathbf{y},\mathbf{Z};\theta) \mid \mathbf{y}\big)$$

는 **잃어버린 정보**다(둘째 등식이 Louis 공식).

증명. **전략** 보조정리 2.4의 첫 등식 $\ell = \log f - \log k$를 두 번 미분하고 숨은 자료의 조건부분포로 평균한다. 둘째 등식은 k가 $\mathbf{z}$의 밀도라는 것과 정리 1.2의 두 표현이다.

① 모든 $\mathbf{z}$에서 $\ell(\theta) = \log f(\mathbf{y},\mathbf{z};\theta) - \log k(\mathbf{z} \mid \mathbf{y};\theta)$이다. 보조정리 2.4

② 두 번 미분하면 $-\nabla^2 \ell(\theta) = -\nabla^2 \log f(\mathbf{y},\mathbf{z};\theta) + \nabla^2 \log k(\mathbf{z} \mid \mathbf{y};\theta)$이다. ①

③ 왼쪽은 $\mathbf{z}$에 의존하지 않으므로 $\mathbf{Z} \sim k(\cdot \mid \mathbf{y};\theta)$로 평균해도 그대로다. 따라서 $J(\theta) = I_c(\theta \mid \mathbf{y}) - I_m(\theta \mid \mathbf{y})$이다. I_c, I_m의 정의

④ θ마다 $k(\cdot \mid \mathbf{y};\theta)$는 $\mathbf{z}$의 밀도이므로 정리 1.2를 이 모형에 쓸 수 있다. $E[\nabla \log k \mid \mathbf{y}] = \mathbf{0}$이고 $-E[\nabla^2 \log k \mid \mathbf{y}] = \mathrm{Cov}(\nabla \log k \mid \mathbf{y})$이다. 정리 1.2

⑤ $\nabla \log k(\mathbf{z} \mid \mathbf{y};\theta) = \nabla \log f(\mathbf{y},\mathbf{z};\theta) - \nabla \ell(\theta)$이고 둘째 항은 $\mathbf{z}$에 대해 상수다. ①을 한 번 미분

⑥ 상수를 빼도 공분산은 같으므로 $I_m(\theta \mid \mathbf{y}) = \mathrm{Cov}\big(\nabla \log f(\mathbf{y},\mathbf{Z};\theta) \mid \mathbf{y}\big)$이다. ④⑤

조건 확인 ④에서 $\int k\,d\mathbf{z} = 1$을 적분 기호 아래에서 두 번 미분한다 [T6]. 정리 1.2의 정칙조건이 숨은 자료의 조건부 모형에서도 성립한다고 가정한다. ∎

> **정리 3.6** (EM의 수렴 속도 (DLR 1977)) $\theta_t \to \hat\theta$ (내점, $J(\hat\theta) \succ 0$)이고 Q가 두 번 연속미분가능하며 $I_c = I_c(\hat\theta \mid \mathbf{y}) \succ 0$이면
>
> $$\theta_{t+1} - \hat\theta = M(\theta_t - \hat\theta) + o\big(\|\theta_t - \hat\theta\|\big), \qquad M = I_c^{-1} I_m.$$
>
> M의 고유값은 모두 $[0,1)$에 있다. 모수가 하나이면 $M = I_m / I_c$는 **잃어버린 정보의 비율**이다.

증명. **전략** M 단계의 1계 조건 $\nabla_1 Q(\theta_{t+1} \mid \theta_t) = \mathbf{0}$을 고정점 $(\hat\theta,\hat\theta)$ 주위에서 일차 전개하고, 두 미분 행렬을 $-I_c$와 I_m으로 읽는다. ∇_1, ∇_2는 Q의 첫 인수와 둘째 인수에 대한 미분이고, 아래 미분은 모두 $(\hat\theta,\hat\theta)$에서 잰다.

① $\nabla_1 Q(\hat\theta \mid \hat\theta) = \nabla \ell(\hat\theta) = \mathbf{0}$이다. 명제 2.5, 따름정리 3.4

② M 단계의 조건을 일차 전개한다. $\delta_t = \theta_t - \hat\theta$로 쓰면

$$\mathbf{0} = \nabla_1 Q(\theta_{t+1} \mid \theta_t) = \nabla_1^2 Q\,\delta_{t+1} + \nabla_2 \nabla_1 Q\,\delta_t + o\big(\|\delta_{t+1}\| + \|\delta_t\|\big).$$ $Q \in C^2$, ①, [T8]

③ $\nabla_1^2 Q(\hat\theta \mid \hat\theta) = E_{\hat\theta}[\nabla^2 \log f(\mathbf{y},\mathbf{Z};\hat\theta) \mid \mathbf{y}] = -I_c$이다. Q의 정의, [T6]

④ 보조정리 2.4로 $Q(\theta \mid \theta') = \ell(\theta) + H(\theta \mid \theta')$이므로

$$\nabla_1 Q(\theta \mid \theta') = \nabla \ell(\theta) + \int \nabla \log k(\mathbf{z} \mid \mathbf{y};\theta)\, k(\mathbf{z} \mid \mathbf{y};\theta')\, d\mathbf{z}.$$ 보조정리 2.4

⑤ ④를 θ'로 미분하고 $\theta = \theta' = \hat\theta$로 두면, $\nabla k = k \nabla \log k$이므로 $\nabla_2 \nabla_1 Q = E[\nabla \log k \nabla \log k' \mid \mathbf{y}] = I_m$이다. 정리 3.5의 ④

⑥ ②에 ③⑤를 넣으면 $I_c \delta_{t+1} = I_m \delta_t + o(\|\delta_{t+1}\| + \|\delta_t\|)$이다. I_c가 가역이므로 $\delta_{t+1} = O(\|\delta_t\|)$이고, 나머지항은 $o(\|\delta_t\|)$이다. 곧 $\delta_{t+1} = M\delta_t + o(\|\delta_t\|)$이다. $I_c \succ 0$

⑦ $M = I_c^{-1}(I_c - J) = \mathbf{1} - I_c^{-1} J$이다. M은 대칭행렬 $I_c^{-1/2} I_m I_c^{-1/2}$와 닮음이므로 고유값이 실수이고, $I_m \succeq 0$이므로 0 이상이다. $\mathbf{1} - M$은 $I_c^{-1/2} J I_c^{-1/2} \succ 0$과 닮음이므로 고유값이 1보다 작다. 정리 3.5, $J \succ 0$

조건 확인 Q가 $(\hat\theta, \hat\theta)$ 근처에서 두 번 연속미분가능해야 ②의 전개가 성립한다. ③과 ⑤에서는 미분과 조건부 기댓값의 순서를 바꾼다 [T6]. ∎

정리는 EM의 성격을 한 문장으로 말한다. **숨은 자료가 정보를 많이 가져갈수록 EM은 느리다.** Newton은 이차 수렴하지만(9회 연습 4.2), EM은 선형 수렴이고 그 비율이 잃어버린 정보의 비율이다. 대신 EM은 헤세 행렬을 구하지 않고, 걸음마다 가능도가 늘며, 모수의 제약(확률의 합이 1, 분산이 양수)을 M 단계가 저절로 지킨다.

> **예제 3.7** (유전 연관 자료의 EM) $\theta_0 = 0.5$에서 EM은 $0.6082,\ 0.6243,\ 0.6265,\ 0.6268, \ldots$ 로 간다. 오차의 비 $\frac{|\theta_{t+1}-\hat\theta|}{|\theta_t-\hat\theta|}$는 $0.1465,\ 0.1346,\ 0.1330,\ 0.1328,\ 0.1328$로 모인다. 정리 3.5로 계산하면 $\hat\theta = 0.6268$에서 관측 정보 $J = 377.5$, 완전 자료 정보 $I_c = 435.3$, 잃어버린 정보 $I_m = 57.8$이고, $I_m/I_c = 0.1328$이 정확히 그 비다(DLR의 표 1). I_m은 Louis 공식으로 $\mathrm{Var}(x_2 \mid \mathbf{y})/\hat\theta^2$이다(연습 3.2). $\hat\theta$의 표준오차는 $1/\sqrt{377.5} = 0.0515$이다.

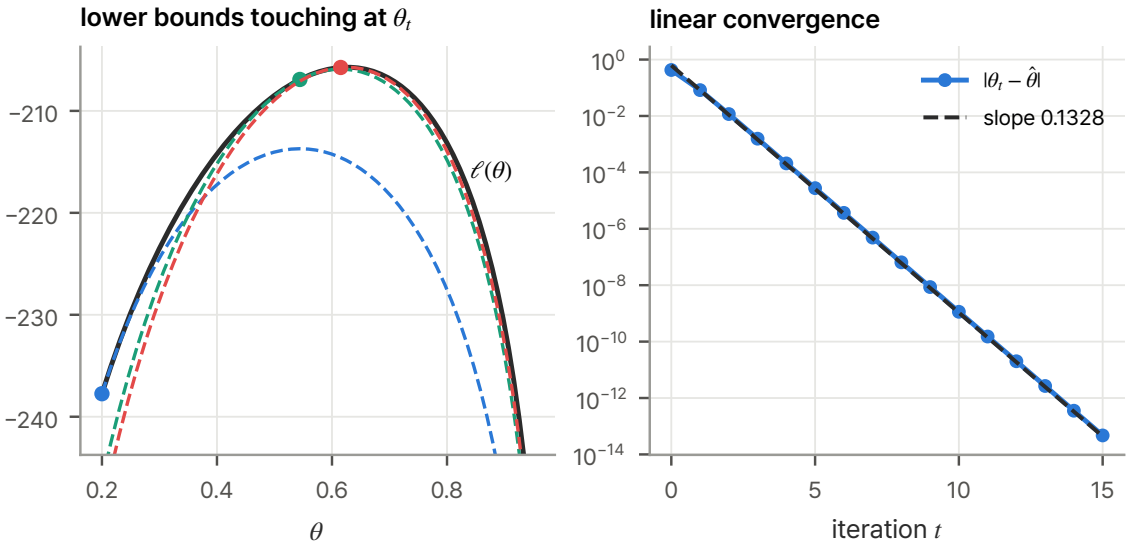


그림 3.2 유전 연관 자료의 EM. 왼쪽: 로그가능도 $\ell(\theta)$(검정)와 세 반복점에서의 하한 $\mathcal{F}(q_t, \theta) = Q(\theta \mid \theta_t) - H(\theta_t \mid \theta_t)$(색). 각 하한은 θ_t에서 ℓ에 닿고, 그 꼭대기가 다음 반복점이다. 오른쪽: 오차 $|\theta_t - \hat\theta|$의 로그. 기울기가 $\log 0.1328$인 직선(점선)을 따른다.

연습문제

연습 3.1 (H는 제자리에서 최대 ★) $H(\theta \mid \theta') \le H(\theta' \mid \theta')$를 Jensen 부등식으로 직접 보여라.

풀이 40쪽

연습 3.2 (유전 연관의 수렴 속도 ★★) 유전 연관 자료에서 EM 사상 $\theta \mapsto \frac{125\theta/(2+\theta)+34}{125\theta/(2+\theta)+72}$의 $\hat\theta$에서의 미분계수가 $I_m/I_c = 0.1328$임을 확인하라. Louis 공식 $I_m = \mathrm{Var}(x_2 \mid \mathbf{y})/\theta^2$도 확인하라.

풀이 40쪽

연습 3.3 (중도절단의 EM ★★★) 연습 2.3에서 EM 반복이 $\theta_{t+1} = \frac{1}{n}\big(\sum y_i + (n-r)\theta_t\big)$임을 보여라. 고정점이 MLE $\sum y_i / r$이고, 수렴 비율이 절단된 비율 $(n-r)/n$임을 보여라. 이것이 I_m/I_c와 같음을 확인하라.

풀이 41쪽

연습 3.4 (일반화 EM ★) M 단계 대신 $Q(\theta_{t+1} \mid \theta_t) \ge Q(\theta_t \mid \theta_t)$인 아무 θ_{t+1}을 고르면 정리 3.2의 결론이 그대로임을 보여라.

풀이 41쪽

연습 3.5 (증거 하한의 증명 ★ 유도형) 명제 3.3를 다음 순서로 증명하라.

(a) 모든 $\mathbf{z}$에서 $\log f(\mathbf{y},\mathbf{z};\theta) = \ell(\theta) + \log k(\mathbf{z} \mid \mathbf{y};\theta)$임을 보이고, 양변을 $\mathbf{Z} \sim q$로 평균하라.

(b) 양변에서 $E_q \log q(\mathbf{Z})$를 빼서 $\mathcal{F}(q,\theta) = \ell(\theta) - D\big(q \,\|\, k(\cdot \mid \mathbf{y};\theta)\big)$를 얻어라.

(c) Gibbs 부등식(2회 정리 8.3)으로 부등식과 등호 조건을 얻어라.

(d) (c)에서 $\theta = \theta_t$를 고정하고 q에 대해 $\mathcal{F}$를 최대화하면 답이 무엇인지 적어라. 이것이 E 단계다.

풀이 41쪽

연습 3.6 (고정점은 정류점: 증명 ★★ 유도형) 따름정리 3.4를 다음 순서로 증명하라.

(a) θ_{t+1}이 내점이면 M 단계의 1계 조건 $\nabla_1 Q(\theta_{t+1} \mid \theta_t) = \mathbf{0}$이 성립하는 이유를 적어라.

(b) $\nabla_1 Q$가 두 인수에 대해 연속이고 $(\theta_{t+1}, \theta_t) \to (\theta^*, \theta^*)$이므로 $\nabla_1 Q(\theta^* \mid \theta^*) = \mathbf{0}$임을 보여라.

(c) 명제 2.5로 $\nabla \ell(\theta^*) = \mathbf{0}$을 결론지어라.

(d) 예제 4.4의 점은 EM의 고정점이면서 정류점이다. 정류점이 최대점일 필요가 없음을 이 예로 설명하라.

풀이 41쪽

4

정규 혼합의 EM

4.1 E 단계와 M 단계

K개 정규 성분의 혼합 $g(y;\theta)=\sum_{k=1}^{K}\pi_k\,\varphi(y;\mu_k,\sigma_k^2)$에서 $\theta=(\pi_1,\dots,\pi_K,\mu_1,\dots,\mu_K,\sigma_1^2,\dots,\sigma_K^2)$이고 $\sum\pi_k=1$이다. 숨은 자료는 이름표 Z_i다(예제 2.2(a)).

명제 4.1 (정규 혼합의 EM) E 단계는 **책임**(responsibility)

$$r_{ik}=P_{\theta_t}(Z_i=k\mid y_i)=\frac{\pi_k\,\varphi(y_i;\mu_k,\sigma_k^2)}{\sum_{j}\pi_j\,\varphi(y_i;\mu_j,\sigma_j^2)}\qquad(\text{오른쪽은 }\theta_t\text{의 값})$$

을 구하는 것이고, $n_k=\sum_i r_{ik}$로 두면 M 단계는

$$\pi_k=\frac{n_k}{n},\qquad \mu_k=\frac{1}{n_k}\sum_i r_{ik}\,y_i,\qquad \sigma_k^2=\frac{1}{n_k}\sum_i r_{ik}\,(y_i-\mu_k)^2$$

이다(σ_k^2의 식에는 새 μ_k를 쓴다).

증명. **전략** 완전 자료 로그가능도는 지시변수 $1\{Z_i=k\}$에 대해 선형이다. E 단계는 지시변수를 그 조건부 기댓값인 책임으로 바꾸고, M 단계는 π의 항과 성분별 항으로 나뉜 Q를 따로 최대화한다.

① 완전 자료 로그가능도는 다음과 같다.

$$\ell_c(\theta)=\sum_i\sum_k 1\{Z_i=k\}\Big[\log\pi_k-\tfrac{1}{2}\log(2\pi\sigma_k^2)-\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\Big].$$

예제 2.2(a)

② 관측이 서로 독립이므로 $E_{\theta_t}[1\{Z_i=k\}\mid \mathbf{y}]=P_{\theta_t}(Z_i=k\mid y_i)=r_{ik}$이다.

지시함수 [T1], 1회 정리 5.5

③ $Q(\theta\mid\theta_t)=\sum_k n_k\log\pi_k+\sum_k\sum_i r_{ik}\big[-\tfrac{1}{2}\log(2\pi\sigma_k^2)-\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\big]$이다. 첫 항은 π에만, 나머지는 성분별 모수에만 의존한다.

①②, 기댓값의 선형성

④ $\sum_k\pi_k=1$ 아래에서 $\sum_k n_k\log\pi_k\le\sum_k n_k\log\frac{n_k}{n}$이고, 등호는 $\pi_k=n_k/n$일 때다.

Gibbs 부등식, 2회 정리 8.3

⑤ 성분 k의 항을 μ_k로 미분하면 $\sum_i r_{ik}(y_i-\mu_k)/\sigma_k^2=0$이므로 $\mu_k=\frac{1}{n_k}\sum_i r_{ik}y_i$이다.

1계 조건, 오목성

⑥ σ_k^2로 미분하면 $-\frac{n_k}{2\sigma_k^2}+\frac{1}{2\sigma_k^4}\sum_i r_{ik}(y_i-\mu_k)^2=0$이므로 $\sigma_k^2=\frac{1}{n_k}\sum_i r_{ik}(y_i-\mu_k)^2$이다.

9회 예제 1.2(c)의 계산

■

이름표를 알면 성분 k의 MLE는 그 성분에 속한 관측들의 평균과 분산이다. 모르니, 각 관측을 책임에 따라 성분들에 "나눠 준" 다음 같은 계산을 한다. n_k는 성분 k의 실질적인 관측 수다. 여러 변수이면 평균은 같은 꼴이고 공분산 행렬은 $\Sigma_k=\frac{1}{n_k}\sum_i r_{ik}(\mathbf{y}_i-\boldsymbol{\mu}_k)(\mathbf{y}_i-\boldsymbol{\mu}_k)'$이다.

알고리즘 4.2 (정규 혼합의 EM (수치적으로 안전한 판))

1. 시작값: k-평균(8절)의 결과나 무작위로 고른 관측을 평균으로, 전체 공분산을 공분산으로, $\pi_k=1/K$.
2. E 단계: $a_{ik}=\log\pi_k+\log\varphi(\mathbf{y}_i;\boldsymbol{\mu}_k,\Sigma_k)$를 로그로 계산하고 $\log r_{ik}=a_{ik}-\operatorname{logsumexp}_j a_{ij}$. 이때 $\ell=\sum_i \operatorname{logsumexp}_j a_{ij}$도 함께 얻는다.
3. M 단계: 명제 4.1.
4. ℓ의 증가가 정한 값보다 작아지면 멈춘다. 여러 시작점에서 돌려 ℓ이 가장 큰 것을 고른다.

$\operatorname{logsumexp}_j a_j=a^*+\log\sum_j e^{a_j-a^*}$ $(a^*=\max_j a_j)$는 지수를 취하기 전에 가장 큰 값을 빼서 넘침과 밑돌기를 막는다. 성분에서 멀리 떨어진 관측은 φ가 10^{-300}보다 작아져 그냥 계산하면 0이 되고, 책임이 0/0이 된다(연습 4.4).

4.2 이름표와 대칭

명제 4.3 (이름표 바꾸기) 성분의 번호를 아무 순열 σ로 바꾼 θ^σ는 $\ell(\theta^\sigma)=\ell(\theta)$를 만족한다. 따라서 K개 성분이 서로 다른 최대점 하나는 $K!$개의 최대점으로 나타난다.

증명. $\sum_k \pi_k\varphi(y;\mu_k,\sigma_k^2)$는 더하는 순서에 무관하다. ■

추정에는 해가 없다. 어느 성분을 "첫째"로 부를지 정하면(예: $\mu_1<\mu_2$) 된다. 그러나 표준오차나 사후분포(16회)를 계산할 때는 이 대칭이 문제를 일으킨다. 더 미묘한 대칭도 있다.

예제 4.4 (같은 곳에서 출발하면 갈라지지 않는다) 모든 성분을 같은 값 (μ, σ^2)에서 출발시키면 모든 관측의 책임이 $r_{ik} = \pi_k$로 y_i에 무관하다. M 단계는 모든 성분에 같은 가중평균 $\bar{y}$와 가중분산 $\hat{\sigma}^2$을 준다. EM은 영원히 "한 정규"에 머문다. 이 점 $(\bar{y}, \hat{\sigma}^2)$은 실제로 혼합 로그가능도의 정류점이다(연습 4.2). 대개 안장점이어서 가능도를 늘리는 방향이 있지만, EM은 대칭을 스스로 깨지 못한다. 시작값을 다르게 잡아야 하는 이유다.

연습문제

연습 4.1 (M 단계의 유도 ★★) 명제 4.1의 σ_k^2 식을 Q를 미분해 유도하라. 여러 변수일 때 Σ_k의 식은 어떻게 되는가? ($\log\det\Sigma$의 미분은 Σ^{-1}이다.) 풀이 41쪽

연습 4.2 (같은 출발점 ★★) 예제 4.4의 점에서 혼합 로그가능도의 기울기가 모든 모수 방향으로 0임을 보여라. (힌트: 명제 2.5와 $r_{ik} = \pi_k$.) 풀이 42쪽

연습 4.3 (Poisson 혼합 ★★) 두 Poisson의 혼합 $\pi\,\mathrm{Poi}(\lambda_1) + (1-\pi)\mathrm{Poi}(\lambda_2)$의 E 단계와 M 단계를 유도하라. M 단계의 λ_k는 책임으로 무게를 단 평균이다. 풀이 42쪽

연습 4.4 (logsumexp ★) $a = (-1000, -1001)$일 때 $\log(e^{a_1} + e^{a_2})$를 그대로 계산하면 무엇이 나오는가? logsumexp로 계산하면? 두 성분 모두에서 멀리 떨어진 관측의 책임을 계산할 때 이것이 왜 필요한지 설명하라. 풀이 42쪽

5

데이터로 — 게, 간헐천, 그리고 “may”

5.1 Weldon의 게와 Pearson의 9차 방정식 (1894)

예제 5.1 (자료) Weldon의 1,000마리는 몸길이에 대한 이마 폭의 비를 0.004 폭의 칸 29개에 묶어 기록했다(R 패키지 `mixdist`의 `pearson`). 칸의 가운데 값으로 계산하면 평균 0.6447, 표준편차 0.0191, 왜도 -0.50, 초과첨도 0.06이다. 왼쪽 꼬리가 길다. 한 정규를 묶음 자료의 가능도로 맞추면 로그가능도가 두 정규의 혼합보다 26.4 낮다($2\Delta\ell = 52.7$). 자료는 한 정규보다 혼합을 강하게 지지한다. 다만 이 비교의 영분포는 $\chi^2(3)$이 아니다(6.2절).

예제 5.2 (Pearson의 적률법) 혼합 $\pi\mathcal{N}(\mu_1,\sigma_1^2)+(1-\pi)\mathcal{N}(\mu_2,\sigma_2^2)$의 처음 다섯 적률을 표본의 것과 같게 놓으면 미지수 다섯의 다항 연립방정식이 된다. Pearson은 이것을 9차 방정식 하나로 줄여 손으로 풀었고, 쓸 만한 해 두 개를 얻었다(첫 성분의 비율 0.4145와 0.5328). 같은 연립방정식을 수치로 풀면 다음 두 해가 나온다.

	π	μ_1	σ_1	μ_2	σ_2	6차 표준적률
해 1	0.424	0.6311	0.0180	0.6547	0.0124	16.17
해 2	0.533	0.6350	0.0191	0.6558	0.0115	16.41
표본						15.52

비율이 Pearson의 손 계산과 조금 다른 것은 묶음을 다루는 방식(칸의 가운데 값, 열린 양끝 칸) 때문이다. 다섯 적률로는 두 해를 가를 수 없다. Pearson은 여섯째 적률이 표본에 더 가까운 해 1을 골랐다. 우리 계산에서도 해 1이 더 가깝다.

예제 5.3 (최대가능도와 EM) 칸의 가운데 값에 칸의 도수를 무게로 주고 명제 4.1의 EM을 돌리면, 서로 다른 네 시작점이 모두 같은 점 $\hat\pi = 0.433$, $\hat\mu_1 = 0.6317$, $\hat\sigma_1 = 0.0183$, $\hat\mu_2 = 0.6546$, $\hat\sigma_2 = 0.0126$으로 간다. 적률법의 해 1과 가깝다. 그런데 로그가능도의 증가가 10^{-12} 아래로 내려갈 때까지 1,200–1,900번을 반복해야 한다. MLE에서 EM 사상의 야코비 행렬(정리 3.6의 M)을 수치로 구하면 가장 큰 고유값이 0.9936이다. 한 번 반복할 때마다 오차가 0.64%씩만 줄어서, 한 자릿수를 얻는 데 약 360번이 든다. 두 성분이 거의 겹쳐 있어서 이름표가 정보를 거의 다 가져갔기 때문이다. 묶음을 정확히 반영한 가능도 $\sum_j n_j \log P(a_j < Y \le b_j)$를 직접

최대화하면 $\hat\pi = 0.453$, $\hat\mu_1 = 0.6326$, $\hat\sigma_1 = 0.0186$, $\hat\mu_2 = 0.6547$, $\hat\sigma_2 = 0.0125$로, 가운데 값을 쓴 것과 거의 같다.

예제 5.4 (비율의 프로파일 가능도) 묶음 가능도에서 π를 고정하고 나머지 넷을 최대화한 프로파일 가능도(정의 1.7)는 아주 평평하다(그림 5.1). 정리 1.8의 95% 구간 $\{\pi : 2[\ell(\hat\theta) - \ell_p(\pi)] \le 3.84\}$은 $[0.175,\ 0.687]$이다. Pearson의 두 해 0.41과 0.53은 모두 그 안에 있다(프로파일 가능도비 0.05와 0.30). 자료는 “두 무리가 섞여 있다”는 것은 강하게 말하지만, **얼마나** 섞여 있는지는 거의 말하지 않는다. 귀찮은 모수 넷의 대가다(정리 1.6). Pearson이 두 해 가운데 하나를 고르기 어려웠던 것도, 혼합 분석 프로그램 MIX(Macdonald)의 예제가 이 자료에서 비율을 0.5로 고정해 두는 것도 같은 이유다. EM이 느린 것은 이 평평함의 다른 얼굴이다(정리 3.6).

5.2 Old Faithful 간헐천

예제 5.5 (분출 시간의 두 무리) 미국 Yellowstone 국립공원의 Old Faithful 간헐천에서 잰 272번의 분출 시간(분, `datasets::faithful`)은 2분 근처와 4.5분 근처에 봉우리가 있다. 두 정규의 혼합을 EM으로 맞추면

$$\hat\pi = 0.348, \quad \hat\mu_1 = 2.019, \quad \hat\sigma_1 = 0.236, \quad \hat\mu_2 = 4.273, \quad \hat\sigma_2 = 0.437$$

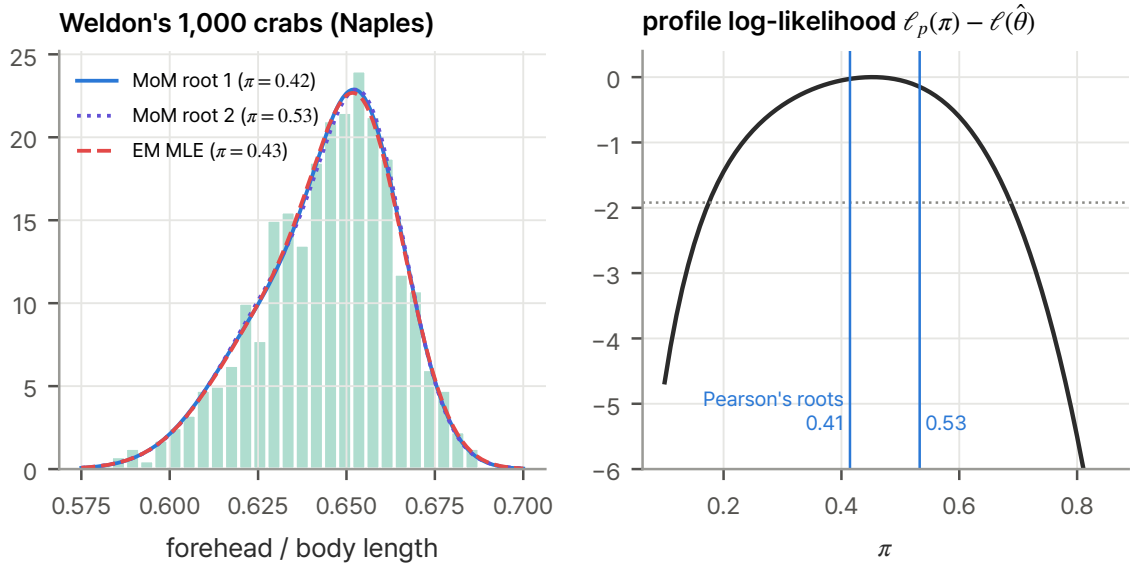


그림 5.1 왼쪽: Weldon의 게 1,000마리(막대, 밀도로 맞춤)와 세 적합. 적률법의 두 해(파랑, 보라)와 EM의 MLE(빨강)가 거의 겹친다. 오른쪽: 첫 성분의 비율 π의 프로파일 로그가능도(묶음 가능도). 가로 점선 위가 95% 구간 $[0.175,\ 0.687]$이고, 세로선은 Pearson의 두 해다.

이고, $\ell = -276.36$이다(한 정규보다 145 높다). 29번의 반복으로 수렴하고, EM 사상의 가장 큰 고유값은 0.589다. 두 무리가 잘 떨어져 있어 이름표의 대부분이 자료에서 거의 확실히 정해지므로 잃어버린 정보가 적다. 관측 정보(수치 헤세 행렬)의 역행렬로 만든 표준오차는 $(0.029,\ 0.026,\ 0.023,\ 0.034,\ 0.027)$이다. 이 결과는 일반 목적의 최적화(BFGS)로 가능도를 직접 최대화한 것과 소수 넷째 자리까지 같다. 분출 시간과 다음 분출까지의 대기 시간을 함께 쓰면 2차원 혼합이 되고, 두 무리의 비율은 0.644와 0.356이다(그림 5.2).

5.3 The Federalist의 “may”, 다시

예제 5.6 (두 Poisson의 혼합 대 음이항) 9회 예제 6.3에서 Madison의 글 262조각의 *may* 개수는 Poisson보다 흩어져 있었고, 음이항이 잘 맞았다. 음이항은 “조각마다 위험률이 감마분포를 따르는” Poisson의 혼합이다(4회 5.2절). 위험률이 두 값 중 하나라고 보면 두 Poisson의 혼합이 되고, EM(연습 4.3)으로 $\hat{\pi} = 0.697$, $\hat{\lambda}_1 = 0.279$, $\hat{\lambda}_2 = 1.524$를 얻는다. 기대도수는 $(155.4,\ 64.9,\ 25.5,\ 10.7,\ 3.9,\ 1.2,\ 0.3)$으로 관측 $(156,\ 63,\ 29,\ 8,\ 4,\ 1,\ 1)$에 잘 맞는다. 로그가능도는 -291.52로, 모수가 셋인데도 모수가 둘인 음이항(-291.26)보다 조금 낮다. 자료는 “주제가 두 종류”보다 “주제마다 조금씩 다르다”를 약간 더 좋아한다. 두 모형 모두 Poisson(-302.90)보다는 훨씬 낫다.

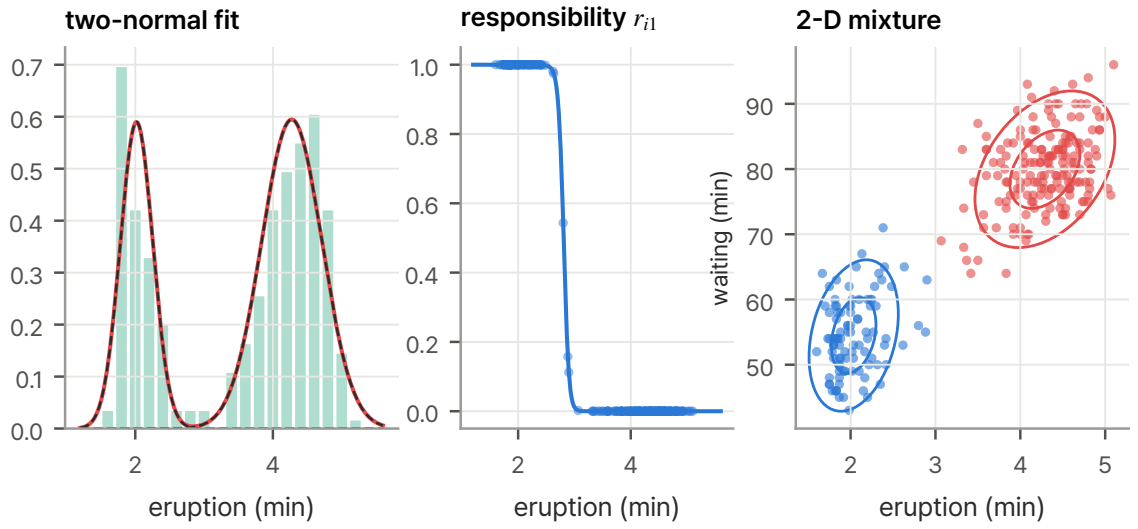


그림 5.2 Old Faithful. 왼쪽: 분출 시간의 히스토그램과 EM 적합(빨강), 두 성분(점선). 가운데: 짧은 무리에 속할 책임 r_{i1}. 경계 근처의 관측만 0과 1 사이의 값을 갖는다. 오른쪽: 분출 시간과 대기 시간의 2차원 정규 혼합. 타원은 각 성분의 1, 2 표준편차 등고선이다.

연습문제

연습 5.1 (적률의 방정식 ★★) 혼합 $\pi\mathcal{N}(\mu_1,\sigma_1^2)+(1-\pi)\mathcal{N}(\mu_2,\sigma_2^2)$의 k차 원점 적률이 $\pi m_k(\mu_1,\sigma_1)+(1-\pi)m_k(\mu_2,\sigma_2)$임을 보이고, $m_1,\dots,m_4$를 쓰라($m_k(\mu,\sigma)=E(\mu+\sigma Z)^k$).
풀이 42쪽

연습 5.2 (여섯째 적률로 고르기 ★) 예제 5.2의 두 해의 6차 표준적률이 16.17과 16.41임을 노트북으로 확인하라. 다섯 적률을 맞춘 두 해를 여섯째 적률로 가르는 것이 왜 약한 기준인가? (6차 적률의 표준오차를 생각하라.) 풀이 42쪽

연습 5.3 (묶음 가능도 ★★) 칸 $(a_j,b_j]$의 도수가 n_j이면 가능도가 $\prod_j P_\theta(a_j<Y\le b_j)^{n_j}$임을 설명하고, 칸이 좁아지면 이것이 가운데 값에서의 밀도로 만든 가능도와 상수배만 다름을 보여라. 풀이 42쪽

연습 5.4 (2차원 Old Faithful ★) 분출 시간과 대기 시간의 2차원 혼합을 노트북으로 맞추고, 두 성분의 상관계수를 구하라. 한 성분 안에서도 두 변수가 상관되어 있는가? 풀이 42쪽

6

부숴보기 — 무한대의 가능도와 성분의 수

6.1 가능도가 무한대로 가는 점

> **명제 6.1** (혼합 가능도는 유계가 아니다) 분산이 성분마다 다른 정규 혼합에서 $\sup_{\theta} \ell(\theta) = \infty$ 이다.

증명. **전략** 한 성분의 평균을 관측 하나에 놓고 그 분산을 0으로 보낸다. 나머지 관측은 다른 성분이 받쳐 주어 로그밀도가 아래로 유계로 남는다. 단계별 증명은 연습 6.4(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 43쪽에 있다.

엄밀한 뜻의 MLE, 곧 가능도의 최댓값은 없다. 한 성분이 관측 하나에 달라붙어 분산을 0으로 줄이면 가능도는 끝없이 커진다(그림 6.1). 그래도 Kiefer(1978)는 내부에 일치하는 근이 있음을 보였다(9회 정리 1.6의 혼합판). EM이 실제로 찾는 것은 이 근 근처의 국소 최대점이다. 다만 반복 중에 한 성분의 n_k가 1 근처로 줄면 그 성분은 특이점으로 끌려간다. 처방은 셋이다. (1) 모든 성분의 분산을 같게 둔다. 그러면 가능도가 유계다(연습 6.1). (2) $\min_k \sigma_k / \max_k \sigma_k \geq c > 0$로 제약한다(Hathaway, 1985). (3) 분산에 사전분포나 벌점을 준다(16회).

6.2 성분은 몇 개인가

한 정규($K = 1$)와 두 정규의 혼합($K = 2$)을 가능도비로 비교하고 싶다. 모수의 차이는 셋(π, 둘째 평균, 둘째 분산)이므로 $\chi^2(3)$을 쓰고 싶어진다. 그러나 Wilks 정리(9회 정리 5.2, 정리 1.8)의 조건이 두 군데서 깨진다.

- 귀무가설 "$K = 1$"은 $\pi = 0$ (경계, 9회 연습 6.3) 또는 $(\mu_1, \sigma_1) = (\mu_2, \sigma_2)$로 쓰인다. 한 모형이 모수공간의 여러 곳에 나타난다.
- $\pi = 0$이면 (μ_2, σ_2)가 가능도에 나타나지 않는다. 귀무가설 아래에서 식별되지 않는 모수가 있다. 그래서 정보 행렬이 특이하고, 테일러 전개의 바탕이 무너진다.

Hartigan(1985)은 평균이 유계가 아니면 이 가능도비가 귀무가설 아래에서 $n \to \infty$일 때 확률적으로 무한대로 간다는 것을 보였다(아주 느리게). 실전의 처방은 모수적 부트스트랩이다. 귀무가설 모형에서 자료를 여러 번 만들어 같은 가능도비를 계산하고, 그 분포에서 관측값의 위치를 읽는다

(15회). $n = 272$인 표준정규 자료로 1,000번 해 보면 가능도비의 95% 분위수는 약 9.6이고, $\chi^2(3)$의 임계값 7.81을 넘는 비율은 약 0.10이다. $\chi^2(3)$을 그대로 쓰면 크기 0.05의 검정이 실제로는 0.10의 검정이 되어 성분을 너무 자주 더한다. Old Faithful의 290이나 게의 52.7은 어느 기준으로도 압도적이다.

6.3 느린 수렴

정리 3.6의 비율을 세 예제에서 비교하면 EM의 성격이 드러난다. 유전 연관은 0.13, Old Faithful은 0.59, Pearson의 게는 0.994다(그림 6.2). 게에서 비율 π를 10^{-6}의 정밀도로 얻으려면 1,700번 넘게 돌려야 한다. 겹친 성분은 이름표의 정보가 대부분 사라진 경우이므로, 느린 수렴은 알고리즘의 결함이라기보다 자료가 모수를 잘 정하지 못한다는 신호다(예제 5.4). EM을 가속하는 방법(Aitken 가속, 몇 번의 EM 뒤에 Newton으로 바꾸기 등)이 있지만, 표준오차가 크다는 사실은 바뀌지 않는다.

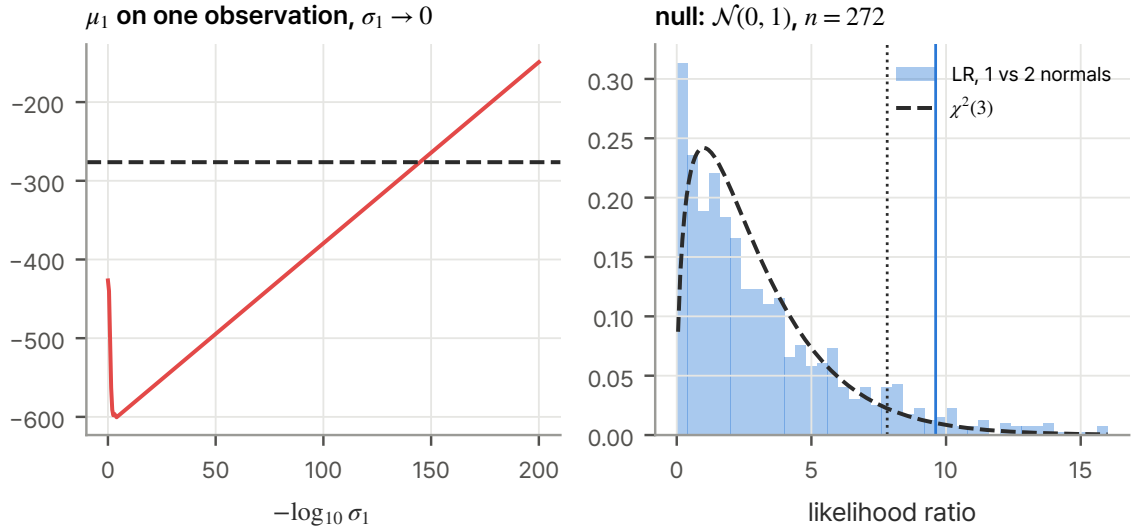


그림 6.1 왼쪽: Old Faithful의 한 관측에 첫 성분의 평균을 놓고($\pi = \frac{1}{2}$, 둘째 성분은 한 정규의 적합) σ_1을 줄이면, 로그가능도가 처음에는 크게 줄다가 $-\log \sigma_1$의 속도로 늘어 $\sigma_1 \approx 10^{-145}$에서 EM의 국소 최대점(점선)을 넘고 무한대로 간다. 특이점은 실재하지만 터무니없이 먼 곳에 있다. 오른쪽: 표준정규 자료($n = 272$)에서 한 정규 대 두 정규 혼합의 가능도비의 분포와 $\chi^2(3)$ 밀도(점선). 오른쪽 꼬리가 더 두껍다.

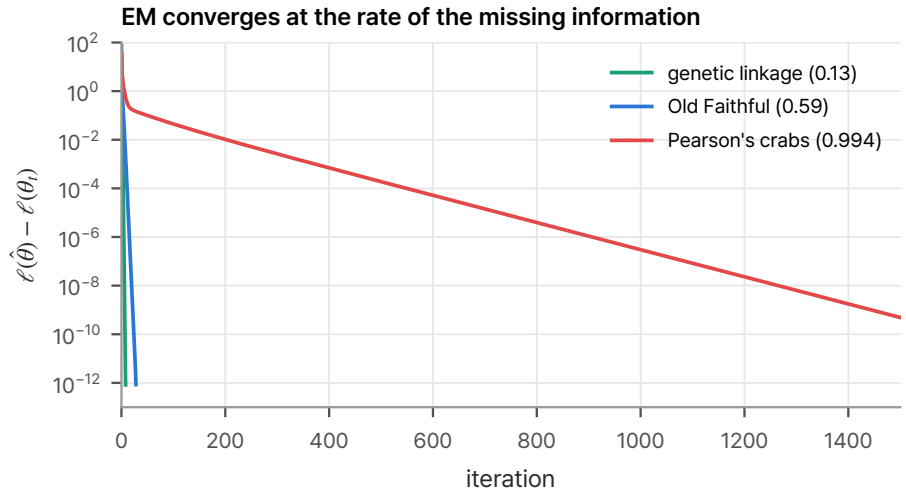


그림 6.2 세 예제에서 EM의 수렴. 세로축은 최종 로그가능도와의 차의 로그. 직선의 기울기가 정리 3.6의 비율(유전 연관 0.13, Old Faithful 0.59, 게 0.994)로 정해진다.

연습문제

연습 6.1 (같은 분산이면 유계 ★★) 모든 성분의 분산이 같은 σ^2인 혼합에서 서로 다른 관측값이 K개보다 많으면 로그가능도가 위로 유계임을 보여라. (힌트: $\sigma \to 0$이면 어떤 성분 평균에도 놓이지 않는 관측의 밀도가 0으로 간다.) 풀이 43쪽

연습 6.2 (분산의 비 제약 ★★★) 서로 다른 관측값이 2개보다 많을 때, $\sigma_1/\sigma_2 \in [c, 1/c]$로 제약하면 두 성분 혼합의 로그가능도가 유계임을 보여라. (힌트: $\sigma_1 \to 0$이면 $\sigma_2 \to 0$이다.) 풀이 43쪽

연습 6.3 (성분 수의 정보 기준 ★) Old Faithful의 분출 시간에 $K = 1, 2, 3$의 정규 혼합을 맞추고 AIC $= -2\ell + 2p$와 BIC $= -2\ell + p \log n$을 비교하라(노트북). 어느 K를 고르는가? 풀이 43쪽

연습 6.4 (혼합 가능도는 유계가 아니다: 증명 ★ 유도형) 두 성분 정규 혼합 $\pi\varphi(y;\mu_1,\sigma_1^2) + (1-\pi)\varphi(y;\mu_2,\sigma_2^2)$에서 명제 6.1를 다음 순서로 증명하라.

(a) 혼합 밀도는 두 항 각각보다 크거나 같음을 보여라.

(b) $\mu_1 = y_1$로 두고 $\pi \in (0,1)$, μ_2, $\sigma_2 > 0$을 고정한다. (a)를 관측 1에는 첫 항으로, 나머지 관측에는 둘째 항으로 써서 $\ell(\theta) \ge \log \frac{\pi}{\sqrt{2\pi}\,\sigma_1} + \sum_{i \ge 2} \log\big((1-\pi)\varphi(y_i;\mu_2,\sigma_2^2)\big)$를 보여라.

(c) $\sigma_1 \to 0$이면 오른쪽이 $+\infty$로 감을 보여라. 둘째 합이 σ_1에 의존하지 않는다는 것이 어디에 쓰이는가?

(d) 두 성분의 분산이 같다는 제약 아래에서는 왜 (b)의 논증이 막히는지 설명하라(연습 6.1).

풀이 43쪽

7

고치기 — Neyman–Scott를 숨은 변수로

9회 예제 7.2에서 n명을 두 번씩 잰 자료 $X_{i1}, X_{i2} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$의 MLE $\hat\sigma^2$은 $\sigma^2/2$로 갔다. 모수 $\mu_1, \dots, \mu_n$이 자료와 함께 늘어서 정리 1.3의 가정이 깨졌기 때문이다. 고치는 방법은 관점을 바꾸는 것이다. 사람마다의 평균을 "추정할 모수"가 아니라 "보이지 않는 확률변수"로 본다.

정의 7.1 (임의효과 모형) $\mu_i \overset{\text{iid}}{\sim} \mathcal{N}(m, \tau^2)$이고, μ_i가 주어지면 $X_{i1}, X_{i2} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$이다. 모수는 (m, τ^2, σ^2) 셋뿐이고 μ_i는 숨은 자료다(예제 2.2(d)).

명제 7.2 (임의효과 모형의 MLE) $D_i = X_{i1} - X_{i2}$, $S_i = X_{i1} + X_{i2}$로 두면 $D_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 2\sigma^2)$, $S_i \overset{\text{iid}}{\sim} \mathcal{N}(2m,\ 4\tau^2 + 2\sigma^2)$이고 $\{D_i\}$와 $\{S_i\}$는 독립이다. 따라서 $\hat\tau^2 \ge 0$이 되는 한 MLE는

$$\hat\sigma^2 = \frac{1}{2n}\sum_i D_i^2, \qquad \hat m = \frac{\bar S}{2}, \qquad \hat\tau^2 = \frac{1}{4}\Big(\frac{1}{n}\sum_i (S_i - \bar S)^2 - 2\hat\sigma^2\Big)$$

이고, $\hat\sigma^2 \xrightarrow{P} \sigma^2$이다.

증명. **전략** 숨은 μ_i를 적분해 없애면 각 쌍이 이변량 정규가 된다. 차와 합으로 바꿔 서로 독립인 두 정규 표본으로 나누고, 각 표본의 분산 MLE에 불변성을 쓴다. 단계별 증명은 연습 7.4(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 44쪽에 있다.

$\hat\sigma^2$이 9회 연습 7.3의 "고친" 추정량과 정확히 같다. 사람마다의 평균을 확률변수로 보았더니, 그것과 무관한 정보(D_i)만으로 σ^2을 추정하게 된 것이다. 이 결과는 μ_i가 실제로 정규분포를 따르지 않아도 맞다. D_i에는 μ_i가 들어 있지 않기 때문이다.

예제 7.3 (EM으로 풀기) 닫힌 꼴이 있지만, μ_i를 숨은 자료로 놓은 EM도 한 줄짜리다. E 단계에서 $\mu_i \mid X_{i1}, X_{i2}$는 정규분포

$$\mu_i \mid \mathbf{x}_i \sim \mathcal{N}(\tilde\mu_i,\ v), \qquad v = \Big(\frac{1}{\tau^2} + \frac{2}{\sigma^2}\Big)^{-1}, \qquad \tilde\mu_i = v\Big(\frac{m}{\tau^2} + \frac{X_{i1}+X_{i2}}{\sigma^2}\Big)$$

이고(연습 7.2), M 단계는 $m = \bar{\tilde\mu}$, $\tau^2 = \overline{(\tilde\mu_i - m)^2} + v$, $\sigma^2 = \frac{1}{2n}\sum_i\sum_j\big[(X_{ij} - \tilde\mu_i)^2 + v\big]$이다. $n = 2000$, $\sigma^2 = 4$, $\tau^2 = 9$로 만든 자료에서 EM은 $\hat\sigma^2 = 3.93$(표준오차 $\sqrt{2\sigma^4/n} = 0.13$)으로 수렴하고, 같은 자료의 "고정효과" MLE는 1.97이다(그림 7.1).

E 단계의 $\tilde\mu_i = m + \frac{\tau^2}{\tau^2 + \sigma^2/2}(\bar X_i - m)$는 각 사람의 평균 $\bar X_i$를 전체 평균 쪽으로 끌어당긴다(연습 7.3).

측정 오차가 클수록(σ^2이 클수록) 더 많이 끌어당긴다. Galton(5회)의 평균으로의 회귀와 같은 모양이고, 16회의 James–Stein 축소 추정량이 이것을 "모수를 확률변수로 보지 않아도" 이득이 된다는 역설로 바꾼다.

연습문제

연습 7.1 (차와 합 ★★) 명제 7.2에서 $\operatorname{Var} S_i = 4\tau^2 + 2\sigma^2$를 확인하고, $\hat\tau^2 < 0$이 되는 확률이 0이 아님을 설명하라. 그때 MLE는 무엇인가? (경계, 9회 연습 6.3) 풀이 44쪽

연습 7.2 (E 단계 ★★) $\mu \sim \mathcal{N}(m, \tau^2)$이고 μ가 주어지면 $X_1, X_2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$일 때 $\mu \mid X_1, X_2$의 분포가 예제 7.3의 정규분포임을 보여라(5회 정리 C, 또는 지수의 완전제곱). 풀이 44쪽

연습 7.3 (축소 ★) $\tilde\mu_i = m + B(\bar X_i - m)$, $B = \dfrac{\tau^2}{\tau^2 + \sigma^2/2}$임을 보여라. $\tau^2 \to 0$과 $\tau^2 \to \infty$에서 $\tilde\mu_i$는 무엇이 되는가? 풀이 44쪽

연습 7.4 (임의효과 모형의 MLE 증명 ★★ 유도형) 명제 7.2를 다음 순서로 증명하라.

(a) $X_{ij} = \mu_i + \varepsilon_{ij}$로 쓰고 μ_i, ε_{i1}, ε_{i2}가 독립인 정규임을 써서 (X_{i1}, X_{i2})가 평균 (m, m), 분산 $\tau^2 + \sigma^2$, 공분산 τ^2인 이변량 정규임을 보여라.

(b) (D_i, S_i)가 이변량 정규이고 $\operatorname{Cov}(D_i, S_i) = 0$임을 보여 독립을 결론지어라(5회 정리 3.8, 정리 3.11). 두 분산을 구하라.

(c) 가능도가 $\{D_i\}$ 부분과 $\{S_i\}$ 부분의 곱임을 보이고, 평균이 0인 정규 표본과 평균이 미지인 정규 표본의 분산 MLE를 각각 적어라.

(d) $(2\sigma^2,\ 4\tau^2 + 2\sigma^2)$가 (σ^2, τ^2)의 일대일 변환임을 보이고 불변성(9회 정리 1.3)으로 $\hat\sigma^2$, $\hat\tau^2$

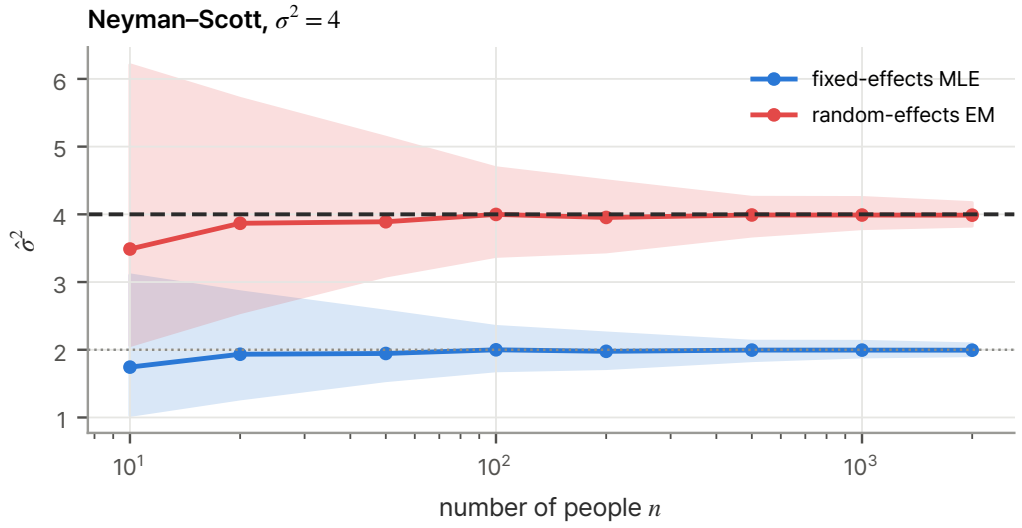


그림 7.1 Neyman–Scott 자료($\sigma^2 = 4$)에서 사람 수 n에 따른 두 추정량. 고정효과 MLE(파랑)는 $\sigma^2/2 = 2$로, 임의효과 모형의 EM(빨강)은 참값 4로 간다. 띠는 반복 200번의 10–90% 범위다.

을 얻어라. $\hat{\tau}^2 \geq 0$이라는 조건은 어디서 필요한가?

(e) $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$을 Khinchin의 대수의 법칙(7회 정리 1.6)으로 보여라.

풀이 44쪽

8

Beautiful insights — 하한과 온도

8.1 ELBO와 변분 오토인코더

명제 3.3의 하한 $\mathcal{F}(q,\theta)$는 기계학습에서 **증거 하한**(evidence lower bound, ELBO)이라 불린다. 생성 모형 $p_\theta(\mathbf{x}\mid\mathbf{z})$가 신경망이고 숨은 변수가 $\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{1})$이면, 관측 가능도 $\int p_\theta(\mathbf{x}\mid\mathbf{z})p(\mathbf{z})\,d\mathbf{z}$도, E 단계의 사후분포 $k(\mathbf{z}\mid\mathbf{x};\theta)$도 계산할 수 없다. 변분 오토인코더(VAE; Kingma–Welling 2014)는 두 가지를 바꾼다.

- E 단계의 정확한 사후분포 대신, 관측 $\mathbf{x}$를 받아 정규분포 $q_\phi(\mathbf{z}\mid\mathbf{x})=\mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}),\operatorname{diag}\boldsymbol{\sigma}_\phi^2(\mathbf{x}))$를 내놓는 신경망(인코더)을 쓴다. 관측마다 E 단계를 푸는 대신 하나의 함수로 "모든 관측의 E 단계를 한꺼번에" 근사하므로 **상각**(amortized) 추론이라 부른다.
- E와 M을 번갈아 하는 대신 $\mathcal{F}(q_\phi,\theta)$를 (ϕ,θ)에 대해 함께 확률적 경사상승으로 올린다. $\mathbf{z}=\boldsymbol{\mu}_\phi(\mathbf{x})+\boldsymbol{\sigma}_\phi(\mathbf{x})\odot\varepsilon$, $\varepsilon\sim\mathcal{N}(\mathbf{0},\mathbf{1})$로 쓰면(재모수화) 기댓값의 기울기를 표본으로 어림할 수 있다.

ELBO를 $E_q\log p_\theta(\mathbf{x}\mid\mathbf{z})-D(q_\phi(\cdot\mid\mathbf{x})\|p)$로 다시 쓰면(연습 8.1), 첫 항은 "복원"의 로그가능도, 둘째 항은 인코더가 사전분포에서 너무 멀어지지 않게 하는 벌점이다. 명제 3.3에서 보았듯이 하한과 진짜 로그가능도의 차는 $D(q\|k)$이다. 인코더가 정확한 사후분포를 내놓지 못하는 만큼 하한이 느슨하다. EM은 이 차를 매번 0으로 만드는 특수한 경우다.

8.2 k-평균은 온도 0의 EM이다

명제 8.1 (딱딱한 EM) 모든 성분의 비율이 같고 공분산이 $\sigma^2\mathbf{1}$로 같은 정규 혼합에서, $\sigma\to 0$이면 책임은

$$r_{ik}=\frac{\exp\!\left(-\|\mathbf{y}_i-\boldsymbol{\mu}_k\|^2/(2\sigma^2)\right)}{\sum_{j}\exp\!\left(-\|\mathbf{y}_i-\boldsymbol{\mu}_j\|^2/(2\sigma^2)\right)}\longrightarrow \mathbf{1}\!\left\{k=\arg\min_{j}\|\mathbf{y}_i-\boldsymbol{\mu}_j\|\right\}$$

(가장 가까운 평균이 하나뿐일 때)로 가고, M 단계의 평균은 각 무리에 배정된 관측의 평균이 된다. 이것이 k-평균 알고리즘(Lloyd 1957, 출판 1982)이다.

증명. **전략** 책임의 분자와 분모를 같은 지수 항으로 나눠 거리의 차만 남긴다. $\sigma\to 0$에서 각 항은 차의 부호에 따라 0이나 무한대로 간다. 단계별 증명은 연습 8.4(유도형)에서 직접 쓴다. 완성된

증명은 부록 A의 풀이 45쪽에 있다.

책임은 거리의 음수를 $2\sigma^2$로 나눈 softmax다. σ^2이 9회 8.3절의 온도 T의 역할을 한다. 온도를 0으로 내리면 "부드러운" 배정이 "딱딱한" 배정이 된다. k-평균은 그래서 빠르고 단순하지만, 모든 무리가 같은 크기의 둥근 모양이라고 가정한 것과 같다. Old Faithful처럼 기울어진 타원 모양의 무리에서는 경계 근처의 관측을 EM과 다르게 배정한다(노트북). k-평균이 매 단계 목적함수 $\sum_i \min_k \|\mathbf{y}_i - \boldsymbol{\mu}_k\|^2$를 줄이는 것은 정리 3.2의 딱딱한 판이다(연습 8.2). EM의 좋은 시작점으로 k-평균의 결과를 쓰는 것이 표준이다(알고리즘 4.2).

8.3 숨은 상태 — Baum–Welch에서 전문가 혼합까지

혼합분포의 이름표가 관측마다 독립이 아니라 Markov 연쇄를 따르면 **은닉 Markov 모형**이 된다. 음성의 소리 조각(관측) 뒤에 숨은 음소(상태)가 있고, 음소는 앞의 음소에 의존한다. Baum과 동료들(1970)의 알고리즘은 이 모형의 EM이다. E 단계의 사후확률을 전향–후향 재귀로 효율적으로 계산한 것이 전부다. 1970년대부터 2010년대 초까지 음성 인식 시스템의 중심이었고, 기계번역의 단어 정렬(IBM 모형, 1993)도 숨은 정렬을 EM으로 추정했다. 4회 8절의 전문가 혼합(mixture of experts)도 원래는 EM으로 학습되었다(Jordan–Jacobs 1994). 오늘의 큰 언어 모형에서 전문가를 고르는 게이트는 경사하강으로 학습하지만, "보이지 않는 배정"을 확률로 다룬다는 생각은 같다.

연습문제

연습 8.1 (ELBO의 두 얼굴 ★★) $f(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$일 때 $\mathcal{F}(q, \theta) = E_q \log p_\theta(\mathbf{x} \mid \mathbf{Z}) - D(q\|p)$임을 보여라. $q = \mathcal{N}(\mu, s^2)$, $p = \mathcal{N}(0,1)$이면 $D(q\|p) = \frac{1}{2}(\mu^2 + s^2 - 1 - \log s^2)$이다. 풀이 45쪽

연습 8.2 (k-평균은 내려간다 ★★) k-평균의 두 단계(가장 가까운 평균에 배정, 배정된 관측의 평균으로 옮김)가 각각 $W = \sum_i \|\mathbf{y}_i - \boldsymbol{\mu}_{c(i)}\|^2$를 늘리지 않음을 보이고, 유한 번 안에 멈춤을 보여라. 풀이 45쪽

연습 8.3 (온도와 책임 ★) 두 성분 1차원 혼합($\pi = \frac{1}{2}$, 같은 σ)에서 책임 r_{i1}이 y_i에 대한 로지스틱 함수 $\big(1 + \exp(-\frac{\mu_1 - \mu_2}{\sigma^2}(y_i - \frac{\mu_1 + \mu_2}{2}))\big)^{-1}$임을 보여라. $\sigma \to 0$이면 계단함수가 된다. 풀이 45쪽

연습 8.4 (딱딱한 EM의 증명 ★ 유도형) $d_{ij} = \|\mathbf{y}_i - \boldsymbol{\mu}_j\|$로 쓰고 가장 가까운 평균 k^*가 하나뿐이라 하자. 명제 8.1을 다음 순서로 증명하라.

(a) 분자와 분모를 $\exp\big(-d_{ik}^2/(2\sigma^2)\big)$로 나눠 $r_{ik} = \big(\sum_j \exp(-(d_{ij}^2 - d_{ik}^2)/(2\sigma^2))\big)^{-1}$임을 보여라.

(b) $k = k^*$이면 $j \neq k^*$인 항이 모두 0으로 가서 $r_{ik^*} \to 1$임을 보여라.

(c) $k \neq k^*$이면 $j = k^*$인 항이 무한대로 가서 $r_{ik} \to 0$임을 보여라.

(d) M 단계의 평균 $\frac{1}{n_k}\sum_i r_{ik}\mathbf{y}_i$의 극한이 무리 k에 배정된 관측의 평균임을 보여라. 무리가 비어 있으면 어떻게 되는가?

(e) 가장 가까운 평균이 둘이면($d_{i1} = d_{i2} < d_{ij},\ j \geq 3$) 극한의 책임은 무엇인가?

풀이 45쪽

9

정리와 역사 노트

9.1 한 장 요약

정리	내용	조건
정보 행렬 (1.2)	$E\mathbf{s}=\mathbf{0}, I=\operatorname{Cov}\mathbf{s}=-E\nabla^2\log f$	정칙조건
점근정규성 (1.3)	$\sqrt{n}(\hat\theta-\theta_0)\xrightarrow{D}\mathcal{N}_p(\mathbf{0},I^{-1})$	$I\succ 0$, 일치하는 근
귀찮은 모수 (1.6)	$I^{\psi\psi}=(I_{\psi\psi}-I_{\psi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi})^{-1}\geq I_{\psi\psi}^{-1}$	직교이면 등호
프로파일 Wilks (1.8)	$2[\ell(\hat\theta)-\ell_p(\psi_0)]\xrightarrow{D}\chi^2(r)$	귀찮은 모수의 수 r 고정
분해·점수 (2.4, 2.5)	$\ell=Q-H,\ \nabla\ell=E[\nabla\log f\mid\mathbf{y}]$	—
단조성 (3.2)	$\ell(\theta_{t+1})\geq\ell(\theta_t)$ (Gibbs)	—
ELBO (3.3)	$\ell=\mathcal{F}(q,\theta)+D(q\Vert k)$, EM은 좌표 상승	—
잃어버린 정보 (3.5)	$J=I_c-I_m,\ I_m=\operatorname{Cov}(\nabla\log f\mid\mathbf{y})$	—
수렴 속도 (3.6)	$\theta_{t+1}-\hat\theta\approx I_c^{-1}I_m(\theta_t-\hat\theta)$	내점, $J\succ 0$

이번 회차의 중심은 두 식이다. 가능도의 분해 $\ell=Q-H$에서 H가 Gibbs 부등식으로 제자리에서 최대라는 사실이 EM의 단조성을 주고, 같은 분해를 두 번 미분한 것이 잃어버린 정보의 원리와 수렴 속도를 준다. 숨은 자료가 정보를 많이 가져갈수록 EM은 느리고, 가능도는 평평하며, 표준오차는 크다. 같은 사실의 세 얼굴이다.

9.2 Stigler의 명명 법칙

> **EM은 몇 번 발견되었나**
>
> - Simon Newcomb(1886)는 천문 관측의 이상치를 다루려고 정규분포 몇 개의 혼합을 가정했다. 그리고 각 관측의 "무게"를 계산해 가중평균을 다시 구하는 반복을 썼다. 오늘의 눈으로 보면 혼합의 EM이다.
> - 같은 반복은 전염병(McKendrick 1926), 실험의 빠진 값(Healy–Westmacott 1956), 불완전한 셈 자

> 료(Hartley 1958), 은닉 Markov 모형(Baum 외 1970), 지수족의 불완전 자료(Orchard–Woodbury 1972, Sundberg 1974)에서 따로따로 나타났다.
> - Dempster–Laird–Rubin(1977)이 이름을 붙이고 일반 이론을 세웠다. 그들의 수렴 증명에는 틀린 곳이 있었고, Wu(1983)가 고쳤다. 이름을 얻은 사람이 처음 찾은 사람이 아니고, 처음 증명한 사람도 완전히 증명하지 못했다.

9.3 7 Pillars 지도

이번 회차도 셋째 기둥 **Likelihood** 위에 있다. 9회가 가능도를 "보이는 자료의 확률"로 다루었다면, 10회는 보이지 않는 것을 적분해 없앤 가능도를 다뤘다. 둘째 기둥 **Information**은 더 정교해졌다. 정보는 행렬이 되었고, 숨은 자료 때문에 잃은 정보를 정확히 셀 수 있게 되었다. 다섯째 기둥 **Regression**도 잠깐 보였다. 임의효과 모형의 E 단계는 Galton의 회귀처럼 각 사람의 평균을 전체 평균 쪽으로 끌어당겼다.

9.4 다음 회 — 충분성: 자료를 줄이는 법 (1920 → 1955)

EM의 M 단계는 "충분통계량의 기댓값"만 있으면 되었다. 정규 혼합에서는 $\sum r_{ik}$, $\sum r_{ik} y_i$, $\sum r_{ik} y_i^2$가 전부였다. 자료 전체 대신 몇 개의 요약만 알아도 모수에 대한 정보를 하나도 잃지 않는다는 것이 Fisher(1920, 1922)의 충분성이다. 11회는 Fisher가 Eddington과 두 표준편차 추정량을 놓고 벌인 논쟁에서 출발한다. 인수분해 정리, Rao–Blackwell과 Lehmann–Scheffé로 최소분산 불편추정량을 찾고, 8회의 독일 전차 추정량이 왜 최선인지 보인다. 지수족이 왜 특별한지(9회 2절의 Cramér–Rao 한계에 닿는 유일한 모형)도 거기서 본다.

참고문헌

- Pearson, K. Contributions to the mathematical theory of evolution. *Phil. Trans. R. Soc. A* 185 (1894), 71–110. · 자료: R 패키지 `mixdist`의 `pearson`(Macdonald).
- Newcomb, S. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics* 8 (1886), 343–366.
- McKendrick, A. G. Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.* 44 (1926), 98–130.
- Hartley, H. O. Maximum likelihood estimation from incomplete data. *Biometrics* 14 (1958), 174–194.
- Baum, L. E., Petrie, T., Soules, G., Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* 41 (1970), 164–171.
- Orchard, T., Woodbury, M. A. A missing information principle: theory and applications. *Proc. 6th Berkeley Symp.* 1 (1972), 697–715. · Sundberg, R. Maximum likelihood theory for incomplete data from an

exponential family. *Scand. J. Statist.* 1 (1974), 49–58.

- Dempster, A. P., Laird, N. M., Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *JRSS B* 39 (1977), 1–38.
- Wu, C. F. J. On the convergence properties of the EM algorithm. *Ann. Statist.* 11 (1983), 95–103. · Louis, T. A. Finding the observed information matrix when using the EM algorithm. *JRSS B* 44 (1982), 226–233.
- Neal, R. M., Hinton, G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* (1998), 355–368.
- Kiefer, N. M. Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica* 46 (1978), 427–434. · Hathaway, R. J. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* 13 (1985), 795–800.
- Hartigan, J. A. A failure of likelihood asymptotics for normal mixtures. *Proc. Berkeley Conf. in Honor of J. Neyman and J. Kiefer* 2 (1985), 807–810.
- Rao, C. R. *Linear Statistical Inference and Its Applications*. Wiley, 1965.
- Lloyd, S. P. Least squares quantization in PCM. *IEEE Trans. Inform. Theory* 28 (1982), 129–137. · Kingma, D. P., Welling, M. Auto-encoding variational Bayes. *ICLR* (2014).
- Jordan, M. I., Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6 (1994), 181–214.
- Azzalini, A., Bowman, A. W. A look at some data on the Old Faithful geyser. *Applied Statistics* 39 (1990), 357–365. · 자료: `datasets::faithful`.
- Stigler, S. M. *The Seven Pillars of Statistical Wisdom*. Harvard, 2016.

A

연습문제 풀이

각 풀이의 머리에 있는 문제 ○○쪽을 누르면 문제로 돌아간다. 계산 문제의 수치는 노트북에서 같은 계산으로 확인했다. 유도형 연습의 풀이는 본문에서 연습으로 돌린 정리의 완성된 증명이다.

1절

연습 1.1 문제 8쪽

$\log f = -\log\sigma - \frac{(x-\mu)^2}{2\sigma^2}$ + 상수에서 $s_\mu = \frac{x-\mu}{\sigma^2}$, $s_\sigma = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$ 이다. $Es_\mu^2 = \frac{1}{\sigma^2}$, $Es_\sigma^2 = \frac{1}{\sigma^2}E(Z^2-1)^2 = \frac{2}{\sigma^2}$, $Es_\mu s_\sigma = \frac{1}{\sigma^3}E[Z(Z^2-1)] = 0$ 이다. 변환 규칙으로는 $\eta = \sigma^2$ 의 정보 $\frac{1}{2\sigma^4}$ 에서 $I(\sigma) = I(\eta)\left(\frac{d\eta}{d\sigma}\right)^2 = \frac{1}{2\sigma^4}\cdot 4\sigma^2 = \frac{2}{\sigma^2}$ 로 같다(9회 연습 2.2의 식을 거꾸로 쓴 것).

연습 1.2 문제 8쪽

첫 블록 행: $I_{\psi\psi}S^{-1} - I_{\psi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi}S^{-1} = SS^{-1} = \mathbf{1}$. 둘째 블록 행: $I_{\lambda\psi}S^{-1} - I_{\lambda\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi}S^{-1} = \mathbf{0}$.

연습 1.3 문제 8쪽

(μ,σ) 의 $I^{-1} = \mathrm{diag}(\sigma^2, \sigma^2/2)$ 이고 $\nabla c = (-\sigma/\mu^2,\ 1/\mu)$ 이므로 $\nabla c' I^{-1}\nabla c = \frac{\sigma^4}{\mu^4} + \frac{\sigma^2}{2\mu^2} = c^2(c^2 + \frac{1}{2})$.

연습 1.4 문제 9쪽

$\ell = n[(\alpha-1)\overline{\log x} - \log\Gamma(\alpha) - \alpha\log\beta] - \sum x_i/\beta$ 에서 $\partial_\beta\ell = -n\alpha/\beta + \sum x_i/\beta^2 = 0$ 이므로 $\hat\beta_\alpha = \bar x/\alpha$. 넣으면 $\ell_p(\alpha) = n[(\alpha-1)\overline{\log x} - \log\Gamma(\alpha) - \alpha\log(\bar x/\alpha) - \alpha]$. 미분하면 $\overline{\log x} - \psi(\alpha) - \log\bar x + \log\alpha + 1 - 1 = 0$.

연습 1.5 문제 9쪽

$\mu_1 - \mu_2 = \psi$ 를 고정하면 $\mu_2 = \lambda$, $\mu_1 = \lambda + \psi$ 이고 σ^2 의 MLE는 잔차 제곱합$/(m+n)$ 이다. λ 에 대해 최소화한 잔차 제곱합은 $\mathrm{SSW} + \frac{mn}{m+n}(\bar x - \bar y - \psi)^2$ (SSW는 두 표본 안의 제곱합)이므로 $\Lambda = (m+n)\log\!\left(1 + \frac{mn}{m+n}\frac{(\bar x-\bar y-\psi_0)^2}{\mathrm{SSW}}\right) = (m+n)\log\!\left(1 + \frac{t^2}{m+n-2}\right)$ 이다. $t^2 = \frac{(\bar x-\bar y-\psi_0)^2}{s_p^2(1/m+1/n)}$, $s_p^2 = \mathrm{SSW}/(m+n-2)$.

연습 1.6 문제 9쪽

(a) 정칙조건(9회 (R2)–(R3)의 성분판)이 $\int f\,dx$를 적분 기호 아래에서 미분하게 해 준다. $0 = \partial_j \int f\,dx = \int \partial_j f\,dx = \int (\partial_j \log f) f\,dx = E s_j$. (b) $\int s_j f\,dx = 0$을 다시 θ_k로 미분하면 $0 = \int (\partial_k s_j) f\,dx + \int s_j (\partial_k f)\,dx = E\big[\frac{\partial^2 \log f}{\partial\theta_j \partial\theta_k}\big] + E[s_j s_k]$이다. $\partial_k f = s_k f$를 썼다. (c) $E\,\mathbf{s} = \mathbf{0}$이므로 $\operatorname{Cov}\mathbf{s} = E\,\mathbf{s}\mathbf{s}' - E\mathbf{s}\,E\mathbf{s}' = E\,\mathbf{s}\mathbf{s}' = I(\theta)$이고, (b)를 행렬로 쓰면 $I(\theta) = -E\nabla^2 \log f$이다. (d) 표본의 로그가능도는 $\sum_i \log f(X_i;\theta)$이므로 점수는 $\sum_i \mathbf{s}(X_i;\theta)$이다. 독립인 벡터의 합의 공분산은 공분산의 합이므로(3회) $nI(\theta)$이다.

연습 1.7 문제 9쪽

(a) $I^{1/2}$는 대칭 제곱근이다(5회 보조정리 3.2). 5회 정리 3.8로 $I^{1/2}\mathbf{W} \sim \mathcal{N}_p(\mathbf{0}, I^{1/2} I^{-1} I^{1/2}) = \mathcal{N}_p(\mathbf{0}, \mathbf{1})$이므로 성분이 독립인 표준정규이고, $\mathbf{W}' I \mathbf{W} = \|I^{1/2}\mathbf{W}\|^2 \sim \chi^2(p)$이다(5회 정리 3.17). (b) $I(\cdot)$이 θ_0에서 연속이고 $\hat\theta_n \xrightarrow{P} \theta_0$이므로 연속사상정리(7회 정리 1.3)로 $I(\hat\theta_n) \xrightarrow{P} I$이다. $\sqrt{n}\mathbf{d}_n = \sqrt{n}(\hat\theta_n - \theta_0)$로 쓰면 $\mathbf{d}_n' I(\hat\theta_n)\mathbf{d}_n \cdot n = \sqrt{n}\mathbf{d}_n' I \sqrt{n}\mathbf{d}_n + \sqrt{n}\mathbf{d}_n'\big(I(\hat\theta_n) - I\big)\sqrt{n}\mathbf{d}_n$이고, 둘째 항은 $O_P(1)\cdot o_P(1)\cdot O_P(1) = o_P(1)$이다(7회 정리 2.5). 첫째 항은 연속사상정리로 $\mathbf{W}' I \mathbf{W} \sim \chi^2(p)$로 분포수렴하고, Slutsky(7회 정리 4.2)로 합도 그렇다. (c) $P\big(n(\hat\theta_n - \theta_0)' I(\hat\theta_n)(\hat\theta_n - \theta_0) \le \chi^2_{p,\alpha}\big) \to 1-\alpha$이므로, 이 부등식을 만족하는 θ_0의 집합이 근사 $100(1-\alpha)\%$ 신뢰영역이다. $I(\hat\theta_n)$이 고정된 양의 정부호 행렬이므로 그 집합은 $\hat\theta_n$을 중심으로 한 타원체다. (d) 7회 정리 5.4로 $\sqrt{n}\big(g(\hat\theta_n) - g(\theta_0)\big) \xrightarrow{D} \nabla g' \mathbf{W} \sim \mathcal{N}(0, \nabla g' I^{-1} \nabla g)$이다. $\nabla g(\theta_0) \neq \mathbf{0}$이어야 이 분산이 양수이고 극한이 퇴화하지 않는다. $\nabla g = \mathbf{0}$이면 일차항이 사라져 이차 전개가 필요하다.

2절

연습 2.1 문제 12쪽

$\frac{125}{2+\theta} - \frac{38}{1-\theta} + \frac{34}{\theta} = 0$에 $\theta(2+\theta)(1-\theta)$를 곱하면 $125\theta(1-\theta) - 38\theta(2+\theta) + 34(2+\theta)(1-\theta) = 15\theta - 197\theta^2 + 68 = 0$. 근은 $\frac{15 \pm \sqrt{225 + 4\cdot 197 \cdot 68}}{394} = \frac{15 \pm 231.97}{394}$, 양의 근 0.6268.

연습 2.2 문제 12쪽

첫 범주의 개체는 확률 $\frac{1}{2} : \frac{\theta}{4}$의 비로 두 칸에 나뉘므로, 첫 범주 125마리가 주어지면 $x_2 \sim \mathrm{b}\big(125, \frac{\theta/4}{1/2 + \theta/4}\big) = \mathrm{b}\big(125, \frac{\theta}{2+\theta}\big)$. 완전 자료 가능도는 $(\theta/4)^{x_2+34}((1-\theta)/4)^{38} \times$상수이므로 MLE는 $\frac{x_2 + 34}{x_2 + 34 + 38}$.

연습 2.3 문제 12쪽

절단되지 않은 개체는 밀도 $\frac{1}{\theta} e^{-y_i/\theta}$, 절단된 개체는 $P(T > c) = e^{-c/\theta} = e^{-y_i/\theta}$를 준다. 곱

은 $\theta^{-r}e^{-\sum y_i/\theta}$ 이고, 미분하면 $\hat\theta = \sum y_i/r$. 무기억성: $P(T>c+s\mid T>c)=e^{-s/\theta}$ 이므로 $T-c\mid T>c\sim\mathrm{Exp}(\theta)$.

연습 2.4 문제 12쪽

$\ell=\sum_i \log\sum_j \pi_j\varphi(y_i;\mu_j,\sigma_j^2)$을 μ_k로 미분하면 $\sum_i \frac{\pi_k\varphi_k(y_i)}{\sum_j\pi_j\varphi_j(y_i)}\cdot\frac{y_i-\mu_k}{\sigma_k^2}=\sum_i r_{ik}\frac{y_i-\mu_k}{\sigma_k^2}$. 명제 2.5로는 완전 자료 점수 $\sum_i \mathbb{1}\{Z_i=k\}\frac{y_i-\mu_k}{\sigma_k^2}$ 의 조건부 기댓값이고, $E[\mathbb{1}\{Z_i=k\}\mid y_i]=r_{ik}$ 이다.

연습 2.5 문제 12쪽

$X=\mu+\sigma U$, $U\sim\mathcal{N}(0,1)$ 이면 $E[U\mid \alpha<U\le\beta]=\frac{\int_\alpha^\beta u\varphi(u)du}{\Phi(\beta)-\Phi(\alpha)}$ 이고 $u\varphi(u)=-\varphi'(u)$ 이므로 분자는 $\varphi(\alpha)-\varphi(\beta)$.

연습 2.6 문제 12쪽

(a) θ의 한 근방에서 $\|\nabla f(\mathbf{y},\mathbf{z};\theta)\|\le G(\mathbf{z})$, $\int G\,d\mathbf{z}<\infty$인 G가 있으면 지배수렴으로 미분과 적분의 순서를 바꿀 수 있다 [T6]. 그러면 $\nabla g=\int\nabla f\,d\mathbf{z}=\int(\nabla\log f)\,f\,d\mathbf{z}$이다. (b) $\nabla\ell=\frac{\nabla g}{g}=\int(\nabla\log f)\frac{f}{g}\,d\mathbf{z}=\int\nabla\log f(\mathbf{y},\mathbf{z};\theta)\,k(\mathbf{z}\mid\mathbf{y};\theta)\,d\mathbf{z}=E_\theta[\nabla\log f(\mathbf{y},\mathbf{Z};\theta)\mid\mathbf{y}]$. (c) $\nabla_\theta Q(\theta\mid\theta')=\int\nabla\log f(\mathbf{y},\mathbf{z};\theta)\,k(\mathbf{z}\mid\mathbf{y};\theta')\,d\mathbf{z}$이다. 가중치가 θ에 의존하지 않으므로 미분이 $\log f$에만 걸린다. $\theta'=\theta$로 두면 (b)의 식이다. (d) 연습 2.2로 $E[x_2\mid\mathbf{y}]=\frac{125\theta}{2+\theta}$이다. 완전 자료 점수는 x_2에 대해 일차식이므로 $E\big[\frac{x_2+34}{\theta}\mid\mathbf{y}\big]-\frac{38}{1-\theta}=\frac{125}{2+\theta}+\frac{34}{\theta}-\frac{38}{1-\theta}$이고, 관측 점수와 같다.

3절

연습 3.1 문제 18쪽

$H(\theta\mid\theta')-H(\theta'\mid\theta')=E_{\theta'}\log\frac{k(\mathbf{Z};\theta)}{k(\mathbf{Z};\theta')}\le\log E_{\theta'}\frac{k(\mathbf{Z};\theta)}{k(\mathbf{Z};\theta')}=\log\int k(\mathbf{z};\theta)\,d\mathbf{z}\le 0$ (Jensen, log는 오목).

연습 3.2 문제 18쪽

EM 사상 $T(\theta)=\frac{u(\theta)+34}{u(\theta)+72}$, $u=\frac{125\theta}{2+\theta}$에서 $T'=\frac{38u'}{(u+72)^2}$, $u'=\frac{250}{(2+\theta)^2}$. $\hat\theta=0.6268$에서 $u=29.83$, $u'=36.23$이므로 $T'=\frac{38\cdot 36.23}{101.83^2}=0.1328$. Louis: $q=\hat\theta/(2+\hat\theta)=0.2386$, $\mathrm{Var}(x_2\mid\mathbf{y})=125q(1-q)=22.71$, $I_m=22.71/0.6268^2=57.8$. $I_c=\frac{u+34}{\hat\theta^2}+\frac{38}{(1-\hat\theta)^2}=162.4+272.9=435.3$이고 $57.8/435.3=0.1328$.

연습 3.3 문제 18쪽

E 단계: $E[T_i \mid T_i > c] = c + \theta_t$ 이므로 완전 자료 합의 기댓값은 $\sum y_i + (n-r)\theta_t$. M 단계는 완전 자료 MLE(합/n)이다. 고정점 $\theta = \frac{1}{n}(\sum y_i + (n-r)\theta)$를 풀면 $\theta = \sum y_i / r$. 사상의 미분이 $(n-r)/n$ 이다. $\hat\theta$에서 $I_c = E[-\ell_c'' \mid y] = -\frac{n}{\theta^2} + \frac{2(\sum y_i + (n-r)\theta)}{\theta^3} = \frac{n}{\theta^2}$, $J = \frac{r}{\theta^2}$ 이므로 $I_m / I_c = (n-r)/n$.

연습 3.4 문제 18쪽

정리 3.2의 증명에서 M 단계가 쓰인 곳은 첫 괄호가 0 이상이라는 것뿐이다.

연습 3.5 문제 18쪽

(a) $f(\mathbf{y}, \mathbf{z}; \theta) = g(\mathbf{y}; \theta)\, k(\mathbf{z} \mid \mathbf{y}; \theta)$에 로그를 취한다. 평균하면 $E_q \log f(\mathbf{y}, \mathbf{Z}; \theta) = \ell(\theta) + E_q \log k(\mathbf{Z} \mid \mathbf{y}; \theta)$이다. ℓ은 $\mathbf{z}$에 의존하지 않으므로 평균해도 그대로다. (b) 양변에서 $E_q \log q$를 빼면 $\mathcal{F}(q, \theta) = \ell(\theta) + E_q \log \frac{k(\mathbf{Z}\mid \mathbf{y};\theta)}{q(\mathbf{Z})} = \ell(\theta) - D\big(q \,\|\, k(\cdot \mid \mathbf{y}; \theta)\big)$. (c) Gibbs 부등식으로 $D \ge 0$이고 등호는 $q = k(\cdot \mid \mathbf{y}; \theta)$일 때뿐이다. 따라서 $\ell \ge \mathcal{F}$이고 등호 조건도 같다. (d) θ_t를 고정하면 $\ell(\theta_t)$는 상수이므로 $\mathcal{F}$의 최대화는 $D(q \| k(\cdot \mid \mathbf{y}; \theta_t))$의 최소화이고, 답은 $q = k(\cdot \mid \mathbf{y}; \theta_t)$다. 숨은 자료의 조건부분포를 구하는 것이 E 단계다.

연습 3.6 문제 18쪽

(a) θ_{t+1}은 $Q(\cdot \mid \theta_t)$의 최대점이고 내점이며 Q가 미분가능하므로, 최대점에서 기울기는 $\mathbf{0}$이다. (b) $\nabla_1 Q$가 두 인수에 대해 연속이므로 $\mathbf{0} = \nabla_1 Q(\theta_{t+1} \mid \theta_t) \to \nabla_1 Q(\theta^* \mid \theta^*)$이다. (c) 명제 2.5의 둘째 등식에서 $\theta = \theta^*$로 두면 $\nabla \ell(\theta^*) = \nabla_1 Q(\theta^* \mid \theta^*) = \mathbf{0}$이다. (d) 예제 4.4에서 모든 성분이 같은 $(\bar y, \hat\sigma^2)$인 점은 EM이 떠나지 않는 고정점이고, 연습 4.2으로 정류점이다. 그러나 두 성분의 평균을 반대 방향으로 조금 벌리면 가능도가 늘어나는 자료가 많으므로(대개 안장점) 최대점이 아니다. 따름정리는 극한이 정류점이라는 것만 말한다.

4절

연습 4.1 문제 21쪽

Q에서 σ_k^2에 관한 부분은 $\sum_i r_{ik}\big[-\frac{1}{2}\log\sigma_k^2 - \frac{(y_i-\mu_k)^2}{2\sigma_k^2}\big]$이고, 미분하면 $-\frac{n_k}{2\sigma_k^2} + \frac{\sum r_{ik}(y_i-\mu_k)^2}{2\sigma_k^4} = 0$. 여러 변수이면 $\sum_i r_{ik}\big[-\frac{1}{2}\log\det\Sigma_k - \frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_k)\big]$을 Σ_k^{-1}로 미분해 $\frac{n_k}{2}\Sigma_k - \frac{1}{2}\sum r_{ik}(\mathbf{y}_i-\boldsymbol{\mu}_k)(\mathbf{y}_i-\boldsymbol{\mu}_k)' = 0$.

연습 4.2 문제 21쪽

모든 성분이 $(\bar{y}, \hat{\sigma}^2)$이면 $r_{ik} = \pi_k$. $\partial \ell / \partial \mu_k = \pi_k \sum (y_i - \bar{y})/\hat{\sigma}^2 = 0$, $\partial \ell / \partial \sigma_k^2 = \pi_k \left[-\frac{n}{2\hat{\sigma}^2} + \frac{\sum (y_i - \bar{y})^2}{2\hat{\sigma}^4} \right] = 0$. π_k 방향은 혼합 밀도가 $\varphi(y; \bar{y}, \hat{\sigma}^2)$로 π에 무관하므로 0이다.

연습 4.3 문제 21쪽

$r_i = \frac{\pi\, p(y_i;\lambda_1)}{\pi\, p(y_i;\lambda_1) + (1-\pi) p(y_i;\lambda_2)}$, $\pi = \frac{1}{n}\sum r_i$, $\lambda_1 = \frac{\sum r_i y_i}{\sum r_i}$, $\lambda_2 = \frac{\sum (1-r_i) y_i}{\sum (1-r_i)}$. 도수 자료이면 각 값에 도수를 무게로 곱한다.

연습 4.4 문제 21쪽

e^{-1000}은 배정밀도에서 0이므로 $\log(0+0) = -\infty$. logsumexp는 $-1000 + \log(1 + e^{-1}) = -999.687$. 두 성분 모두에서 먼 관측은 두 밀도가 모두 0으로 밑돌아 책임이 0/0이 된다. 로그에서 최댓값을 빼면 상대적인 크기가 보존된다.

5절

연습 5.1 문제 25쪽

$EX^k = \pi E[X^k \mid Z=1] + (1-\pi) E[X^k \mid Z=2]$(전확률). $m_1 = \mu$, $m_2 = \mu^2 + \sigma^2$, $m_3 = \mu^3 + 3\mu\sigma^2$, $m_4 = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$.

연습 5.2 문제 25쪽

6차 표준적률의 표준오차를 $\sqrt{\widehat{\mathrm{Var}}(Z^6)/n}$으로 어림하면 약 2.6인데, 두 해의 차이는 0.24다. 여섯째 적률은 다섯 적률보다 훨씬 흔들리는 통계량이어서 이 기준으로는 두 해를 거의 가를 수 없다. 프로파일 가능도(예제 5.4)도 같은 말을 한다.

연습 5.3 문제 25쪽

각 관측은 어느 칸에 들어갔는지만 알려 주므로 칸 j의 확률 $P(a_j < Y \le b_j)$가 그 관측의 가능도다. 폭 h의 칸이면 $P \approx h\, g(c_j)$ (c_j는 가운데 값)이므로 가능도는 $h^n \prod g(c_j)^{n_j}$이고, h^n은 모수에 의존하지 않는다.

연습 5.4 문제 25쪽

두 성분의 상관계수는 0.29(짧은 분출)와 0.38(긴 분출)이다. 전체 상관계수 0.90의 대부분은 무리 사이의 차이에서 온다. 무리 안의 상관은 약하다. 5회의 Galton 자료와 8회의 Berkeley에서 본

것처럼, 섞인 집단의 상관은 집단 안의 상관과 다를 수 있다.

6절

연습 6.1 문제 28쪽

$\sigma \to 0$이면 어떤 μ_k와도 같지 않은 관측 y_i의 혼합 밀도는 $\sum \pi_k \varphi(y_i;\mu_k,\sigma^2) \le \frac{1}{\sqrt{2\pi}\sigma} e^{-\delta^2/(2\sigma^2)} \to 0$ (δ는 가장 가까운 평균까지의 거리)이다. 서로 다른 값이 K개보다 많으면 그런 관측이 늘 있다. 그 관측의 로그밀도가 $-\infty$로 가는 속도($-\delta^2/(2\sigma^2)$)가 평균 위의 관측의 증가 속도($-\log\sigma$)보다 빠르므로 위로 유계다. (δ가 0으로 가는 경우도 μ_k가 많아야 K개의 값에만 붙을 수 있다는 것으로 같은 논증이 된다.)

연습 6.2 문제 28쪽

$\sigma_1 \to 0$이면 $\sigma_2 \le \sigma_1/c \to 0$이다. 그러면 연습 6.1의 논증이 그대로 쓰여, 어느 평균에도 붙지 않은 관측의 밀도가 지수적으로 0으로 간다. 두 분산이 모두 0에서 떨어져 있으면 밀도는 $\frac{1}{\sqrt{2\pi}\min\sigma_k}$ 이하로 유계다.

연습 6.3 문제 28쪽

로그가능도는 $K=1,2,3$에서 -421.4, -276.4, -267.9이고 모수의 수는 $3K-1$이다. AIC는 846.8, 562.7, 551.8로 $K=3$을, BIC는 854.1, 580.8, 580.6으로 $K=2$와 3을 거의 같게 본다. 셋째 성분(비율 0.15, 평균 3.73, 표준편차 0.54)은 긴 분출 무리의 왼쪽 꼬리를 맞춘다. 정보 기준도 성분 수의 문제를 완전히 풀지 못한다(6.2절).

연습 6.4 문제 28쪽

(a) 두 항이 모두 0 이상이므로 합은 각 항보다 크거나 같다. (b) 관측 1에는 첫 항, 나머지에는 둘째 항을 쓰면 $\ell(\theta) = \sum_i \log g(y_i) \ge \log\big(\pi\varphi(y_1;y_1,\sigma_1^2)\big) + \sum_{i\ge 2}\log\big((1-\pi)\varphi(y_i;\mu_2,\sigma_2^2)\big)$이고, $\varphi(y_1;y_1,\sigma_1^2) = \frac{1}{\sqrt{2\pi}\sigma_1}$이다. (c) 첫 항 $\log\pi - \frac{1}{2}\log 2\pi - \log\sigma_1$은 $\sigma_1\to 0$에서 $+\infty$로 가고, 둘째 합은 σ_1에 의존하지 않는 유한한 값이다. 둘째 합이 고정되어 있어야 첫 항의 발산이 다른 관측의 손실로 상쇄되지 않는다. (d) 분산이 같으면 $\sigma_1\to 0$일 때 $\sigma_2\to 0$이어서 둘째 합의 각 항이 $-\infty$로 간다. 어느 평균에도 놓이지 않은 관측의 로그밀도는 $-\delta^2/(2\sigma^2)$의 속도로 줄어 $-\log\sigma$의 증가를 이긴다(연습 6.1의 풀이).

7절

연습 7.1 문제 30쪽

$\operatorname{Var} S_i = \operatorname{Var} X_{i1} + \operatorname{Var} X_{i2} + 2\operatorname{Cov} = 2(\tau^2+\sigma^2) + 2\tau^2$. $\frac{1}{n}\sum (S_i - \bar S)^2 < 2\hat\sigma^2$이면 식의 $\hat\tau^2$이 음수다. τ^2이 작으면 이 확률이 크다. 그때 MLE는 경계 $\hat\tau^2 = 0$이고, σ^2은 D와 S를 함께 써서 다시 추정한다(모든 X_{ij}가 같은 평균의 iid가 된다).

연습 7.2 문제 30쪽

$\log p(\mu \mid \mathbf{x}) = -\frac{(\mu-m)^2}{2\tau^2} - \frac{(x_1-\mu)^2+(x_2-\mu)^2}{2\sigma^2} +$상수는 μ의 이차식이고, μ^2의 계수 $-\frac{1}{2}(\frac{1}{\tau^2} + \frac{2}{\sigma^2})$와 μ의 계수 $\frac{m}{\tau^2} + \frac{x_1+x_2}{\sigma^2}$에서 분산 v와 평균 $v(\frac{m}{\tau^2} + \frac{x_1+x_2}{\sigma^2})$가 나온다.

연습 7.3 문제 30쪽

$v = \frac{\tau^2\sigma^2/2}{\tau^2+\sigma^2/2}$이므로 $\tilde\mu_i = \frac{\sigma^2/2}{\tau^2+\sigma^2/2} m + \frac{\tau^2}{\tau^2+\sigma^2/2}\bar X_i$. $\tau^2 \to 0$이면 m(모두 같은 사람), $\tau^2 \to \infty$이면 $\bar X_i$(각자의 평균).

연습 7.4 문제 30쪽

(a) $\mu_i \sim \mathcal{N}(m, \tau^2), \varepsilon_{i1}, \varepsilon_{i2} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$이 독립이면 (X_{i1}, X_{i2})는 독립 정규 벡터 $(\mu_i, \varepsilon_{i1}, \varepsilon_{i2})$의 선형변환이므로 이변량 정규다(5회 정리 3.8). $EX_{ij} = m$, $\operatorname{Var} X_{ij} = \tau^2+\sigma^2$, $\operatorname{Cov}(X_{i1}, X_{i2}) = \operatorname{Var}\mu_i = \tau^2$. (b) (D_i, S_i)도 선형변환이므로 이변량 정규다. $\operatorname{Cov}(D_i, S_i) = \operatorname{Var} X_{i1} - \operatorname{Var} X_{i2} = 0$이므로 5회 정리 3.11로 독립이다. $\operatorname{Var} D_i = 2(\tau^2+\sigma^2) - 2\tau^2 = 2\sigma^2$, $\operatorname{Var} S_i = 2(\tau^2+\sigma^2) + 2\tau^2 = 4\tau^2 + 2\sigma^2$이고 $ED_i = 0$, $ES_i = 2m$. (c) 쌍들은 독립이고 각 쌍에서 $D_i \perp S_i$이므로 결합밀도는 $\prod_i \varphi(D_i; 0, 2\sigma^2) \cdot \prod_i \varphi(S_i; 2m, 4\tau^2+2\sigma^2)$이다. $(X_{i1}, X_{i2}) \mapsto (D_i, S_i)$의 Jacobian은 상수이므로 가능도는 이 곱과 상수배만 다르다. 평균이 0인 정규 표본의 분산 MLE는 $\frac{1}{n}\sum D_i^2$, 평균이 미지인 정규 표본의 분산 MLE는 $\frac{1}{n}\sum (S_i - \bar S)^2$이고, 평균의 MLE는 $\bar S$이다. (d) $a = 2\sigma^2$, $b = 4\tau^2 + 2\sigma^2$에서 $\sigma^2 = a/2$, $\tau^2 = (b-a)/4$로 되돌릴 수 있으므로 일대일이다. 불변성으로 $\hat\sigma^2 = \frac{1}{2n}\sum D_i^2$, $\hat m = \bar S/2$, $\hat\tau^2 = \frac{1}{4}\big(\frac{1}{n}\sum (S_i - \bar S)^2 - 2\hat\sigma^2\big)$이다. 이 변환이 모수공간 $\{\sigma^2 > 0, \tau^2 \ge 0\}$ 안에 머물려면 $b \ge a$여야 하므로, 추정값이 그 조건을 만족할 때만 이 식이 MLE다(연습 7.1). (e) D_i^2는 평균 $2\sigma^2$인 iid이므로 Khinchin으로 $\frac{1}{n}\sum D_i^2 \xrightarrow{P} 2\sigma^2$, 따라서 $\hat\sigma^2 \xrightarrow{P} \sigma^2$이다. 쌍의 수와 함께 늘어나는 모수가 없다.

8절

연습 8.1 문제 33쪽

$E_q \log f - E_q \log q = E_q \log p_\theta(\mathbf{x} \mid \mathbf{Z}) + E_q \log \frac{p(\mathbf{Z})}{q(\mathbf{Z})}$ 이고 둘째 항이 $-D(q\|p)$. 정규끼리이면 $E_q \log q = -\frac{1}{2}\log(2\pi s^2) - \frac{1}{2}$, $E_q \log p = -\frac{1}{2}\log 2\pi - \frac{1}{2}(\mu^2 + s^2)$이므로 $D = \frac{1}{2}(\mu^2 + s^2 - 1 - \log s^2)$.

연습 8.2 문제 33쪽

배정 단계는 각 관측의 항을 가장 가까운 평균으로 바꾸므로 줄이거나 그대로다. 평균 단계는 각 무리 안에서 $\sum \|\mathbf{y}_i - \mathbf{c}\|^2$를 최소화하는 $\mathbf{c}$가 평균이므로 줄이거나 그대로다. 배정의 가짓수가 유한(K^n)이고 W가 엄밀히 줄지 않으면 배정이 바뀌지 않으므로 유한 번 안에 멈춘다.

연습 8.3 문제 33쪽

$r_{i1} = \big(1 + \exp(-\frac{(y_i-\mu_2)^2 - (y_i-\mu_1)^2}{2\sigma^2})\big)^{-1}$ 이고 $(y-\mu_2)^2 - (y-\mu_1)^2 = 2(\mu_1-\mu_2)\big(y - \frac{\mu_1+\mu_2}{2}\big)$.

연습 8.4 문제 33쪽

(a) 비율이 같으므로 π_k는 약분된다. 분자와 분모를 $\exp(-d_{ik}^2/(2\sigma^2))$로 나누면 분자는 1, 분모는 $\sum_j \exp(-(d_{ij}^2 - d_{ik}^2)/(2\sigma^2))$이다. (b) $k = k^*$이면 $j \neq k^*$에서 $d_{ij}^2 - d_{ik^*}^2 > 0$이므로 그 항은 $\sigma \to 0$에서 0으로 간다. $j = k^*$인 항은 1이므로 $r_{ik^*} \to 1$. (c) $k \neq k^*$이면 $j = k^*$인 항의 지수 $-(d_{ik^*}^2 - d_{ik}^2)/(2\sigma^2)$이 $+\infty$로 가므로 분모가 무한대로 가고 $r_{ik} \to 0$. (d) 책임은 무리 배정의 지시함수로 가고 n_k는 무리 k의 관측 수로 간다. 무리가 비어 있지 않으면 $\frac{\sum_i r_{ik}\mathbf{y}_i}{\sum_i r_{ik}}$는 무리 평균으로 간다. 무리가 비면 $n_k \to 0$이어서 평균이 정해지지 않는다. k-평균에서는 그 평균을 그대로 두거나 다시 고른다. (e) $r_{i1} = r_{i2} \to \frac{1}{2}$이고 나머지는 0으로 간다. 거리가 같은 두 평균이 관측을 반씩 나눠 갖는다.