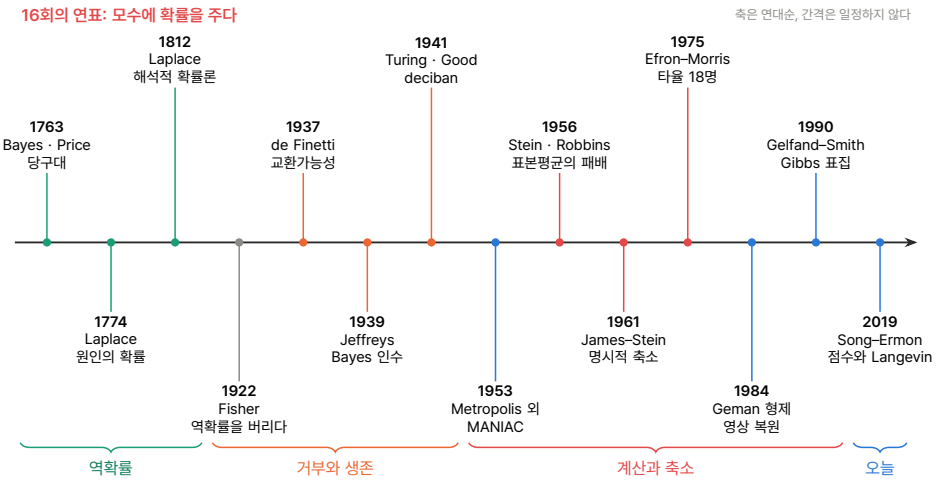
수리통계학 × 통계학의 역사

16회 강의노트

베이즈의 귀환
MCMC와 Stein의 역설

1763 → 1990

이 과정은 Bayes의 당구대에서 시작했다. 마지막 회차는 그리로 돌아간다. 모수를 고정된 미지수가 아니라 확률변수로 보는 순간, 추론은 사후분포를 계산하는 일이 된다. Laplace는 그 계산으로 남아의 출생 확률을 재었고, Fisher는 그것을 거부했으며, 베이즈 추론은 암호 해독과 몬테카를로 속에서 살아남아 1990년 Gibbs 표집과 함께 돌아왔다. 같은 시기 Stein은 정규 평균 세 개 이상을 함께 추정할 때 표본평균이 최선이 아님을 보였다. 그 답은 Galton의 회귀였다.



다루는 것	사후분포와 켤레 사전분포 · Laplace 근사 · 사후평균과 Bayes 추정 · Jeffreys 사전분포의 불변성 · Bayes 인수와 증거의 무게 · 마르코프 연쇄와 상세 균형 · Metropolis–Hastings와 Gibbs · $\hat{R}$ · Stein의 보조정리와 James–Stein · 경험적·계층 Bayes · ridge·lasso와 사전분포 · Langevin 표집
원자료	Laplace의 파리 출생 기록(1745–1770) · Arbuthnot의 런던 세례 기록(1629–1710) · Efron–Morris의 1970년 타율 18명
노트북	`sessions/16_bayes_mcmc/16_bayes_mcmc.ipynb` — 절·정리 번호가 같다
워크북	`16_bayes_mcmc_workbook.pdf` (손으로 푸는 34문항) · 풀이판 `_workbook_sol.pdf`
선수 지식	1–15회 (특히 1회 Bayes 정리, 4회 베타분포와 당구대, 5회 축소, 7회 몬테카를로, 9회 Fisher 정보, 12회 Bayes 규칙, 13회 ridge)

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 ①②③으로 쓰고, 단계마다 오른쪽 끝에 근거를 단다. 읽을 때는 근거를 가리고 "왜 이 단계가 성립하는가"를 먼저 말해 본다. 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 [T2](귀류), [T6](극한과 적분의 교환), [T7](변수변환)을 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 1–15회에서 가져오는 것 2

0.2 표기 3

1 Laplace의 역확률 (1763–1812) 4

1.1 사후분포와 켤레 사전분포 4

1.2 Laplace 근사 6

2 Bayes 추정과 Jeffreys 사전분포 8

2.1 손실함수와 Bayes 추정량 8

2.2 Jeffreys 사전분포 11

3 Fisher의 거부와 증거의 무게 14

3.1 Bayes 인수 14

3.2 Lindley의 역설 15

4 MCMC — 사후분포에서 뽑기 17

4.1 마르코프 연쇄와 정상분포 17

4.2 Metropolis–Hastings와 Gibbs 18

4.3 수렴과 진단 19

5 Stein의 역설 23

5.1 Stein의 보조정리와 James–Stein 정리 23

5.2 경험적 Bayes — 왜 줄이는가 25

6 Efron–Morris의 타율 28

6.1 자료와 James–Stein 28

6.2 계층 Bayes와 Gibbs 표집 29

7 베이즈가 실패할 때 31

7.1 봉우리가 여럿일 때 . . . 31
7.2 제안 폭과 유효 표본 크기 . . . 32
7.3 계산 코드의 점검 . . . 32
7.4 축소가 개인을 해칠 때 . . . 33

8 오늘 — 벌점은 사전분포다 35

8.1 ridge와 lasso의 사전분포 . . . 35
8.2 Langevin 표집 . . . 36
8.3 Tweedie의 공식과 확산모형 . . . 37

9 정리와 역사 노트 40

9.1 한 장 요약 . . . 40
9.2 Stigler의 명명 법칙 . . . 41
9.3 7 Pillars 지도 — 완성 . . . 41
9.4 과정을 마치며 . . . 42

A 연습문제 풀이 44

0

들어가며

> **여는 질문.** 1745년부터 1770년까지 파리에서 남아 251,527명, 여아 241,945명이 태어났다. Laplace는 물었다. 남아가 태어날 확률이 1/2보다 크다는 것을 우리는 얼마나 확신할 수 있는가? 그리고 이 질문에 답하려면 확률이 무엇에 대한 것이어야 하는가?

이 과정은 1회에서 확률의 공리로, 4회에서 Bayes의 당구대로 시작했다. 당구공이 멈춘 자리를 모르는 채 공을 몇 번 더 굴려 그 자리를 추론하는 문제였다. 그때 모르는 자리 p는 **확률변수**였다. 공이 떨어진 자리는 균등분포를 따랐기 때문이다. 그 뒤의 회차들에서 모수는 고정된 미지수였다. 7회의 대수의 법칙, 9회의 최대가능도, 12회의 검정, 13회의 최소제곱은 모두 "참값 θ는 하나이고, 무작위인 것은 자료"라는 관점 위에 섰다.

이 마지막 회차는 그 관점을 뒤집는다. 모수에도 확률을 준다. 그러면 추론은 한 가지 계산이 된다. 자료가 주어졌을 때 모수의 조건부분포, 곧 **사후분포**를 구하는 것이다. Laplace는 1774년 이 계산을 일반화했고, 1781–1786년 파리와 런던의 출생 기록에 썼다. Fisher는 사전분포가 주관적이라며 이것을 거부했고, 20세기 전반의 통계학은 그를 따랐다. 베이즈 추론은 Jeffreys의 책, Bletchley Park의 암호 해독, 그리고 1953년 Los Alamos의 몬테카를로 속에서 살아남았다. 1990년 Gelfand와 Smith가 Gibbs 표집을 통계학에 소개하면서 계산의 벽이 무너졌다. 같은 시기에 다른 방향에서 놀라운 결과가 나왔다. Stein(1956)은 정규 평균 세 개 이상을 함께 추정할 때 표본평균이 최선이 아님을 보였다. 그 더 나은 추정량은 사전분포를 쓰지 않고도 베이즈 추정량처럼 행동했다.

> **색이 바뀐다** 이 과정의 그림과 수식에서 모수는 보라(고정된 미지수), 자료와 확률변수는 청록이었다. 이번 회차부터 모수 θ도 청록이다. 모수가 확률변수가 되었기 때문이다.

1절은 사후분포와 켤레 사전분포로 Laplace의 계산을 재현한다. 2절은 사후평균이 제곱손실의 Bayes 추정량임을 보이고 Jeffreys 사전분포를 다룬다. 3절은 Bayes 인수와 Turing의 "증거의 무게", 4절은 MCMC의 원리(상세 균형, Metropolis–Hastings, Gibbs)를 증명한다. 5절은 James–Stein 정리와 그것의 경험적 Bayes 해석, 6절은 Efron–Morris의 타율 자료로 그것을 확인한다. 7절은 사전분포와 MCMC가 실패하는 곳, 8절은 ridge · lasso · weight decay와 Langevin 표집 속의 베이즈다. 9절은 16회 전체를 정리한다.

학습 목표

1. 켤레 사전분포로 사후분포를 구하고, 파리 출생 자료의 $P(p \le \frac{1}{2} \mid \text{자료}) = 1.15 \times 10^{-42}$를 정확한 계산과 Laplace 근사로 구한다.
2. 손실함수에 따른 Bayes 추정량(사후평균 · 사후중앙값)을 유도하고, Jeffreys 사전분포의 불변성을 증명한다.
3. 상세 균형에서 Metropolis–Hastings와 Gibbs 표집의 정상분포를 증명하고, 구현해 $\hat{R}$로 수렴을 진단한다.
4. Stein의 보조정리로 James–Stein 정리를 증명하고, 그것을 경험적 Bayes와 Galton의 회귀로 해석한다.
5. 타율 자료로 James–Stein · 계층 Bayes를 비교하고, ridge · lasso · weight decay · Langevin 표집을 사전분포의 언어로 읽는다.

0.1 1–15회에서 가져오는 것

앞 회차	내용
1회 정리 5.5 · 4회 정리 4.5	Bayes 정리, Bayes의 당구대(베타 사후분포)
4회 5절 · 5회 따름정리 2.2	Laplace 분포(1774), 정규를 정규로 섞으면 축소
3회 정리 2.5 · 2.6	반복기댓값, 전분산 공식
7회 정리 6.1 · 9회 정의 2.1	몬테카를로의 오차, Fisher 정보
12회 정리 5.2 · 정의 4.1	Bayes 규칙(사후 기대손실 최소), 축차확률비 검정
13회 정의 8.4 · 정리 8.6	ridge, Hoerl–Kennard

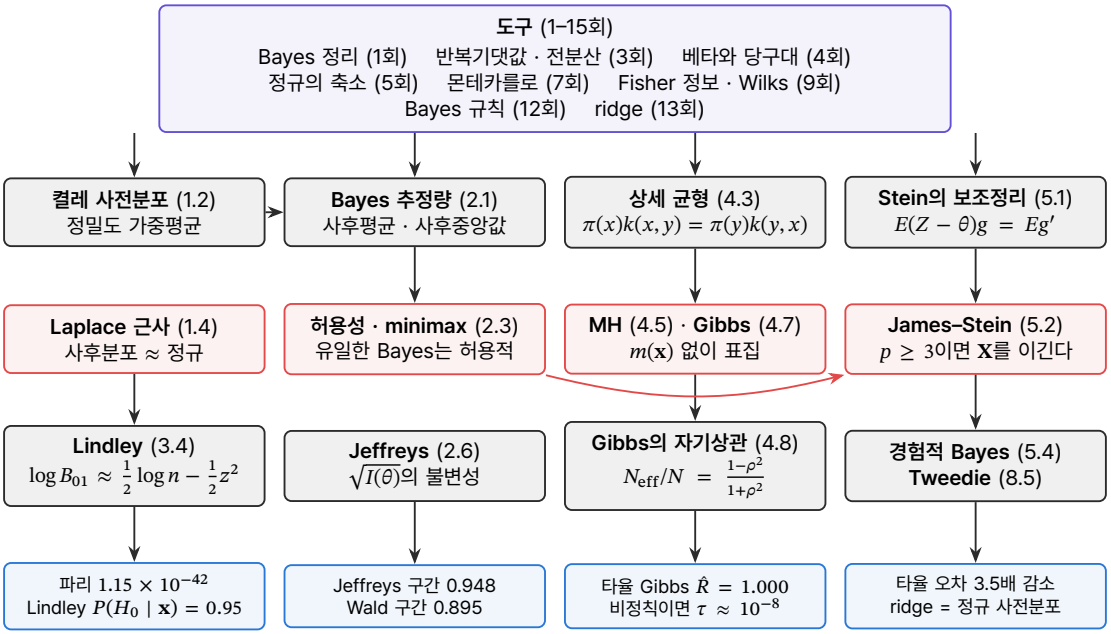


그림 0.1 이번 회차 정리들의 의존 관계. 화살표는 "증명에 쓰인다"는 뜻이다.

0.2 표기

사전분포의 밀도를 $\pi(\theta)$, 가능도를 $f(\mathbf{x} \mid \theta)$로 쓴다. 사후분포는 $\pi(\theta \mid \mathbf{x}) \propto \pi(\theta) f(\mathbf{x} \mid \theta)$이고, 비례상수 $m(\mathbf{x}) = \int \pi(\theta) f(\mathbf{x} \mid \theta)\, d\theta$를 **주변 가능도**(증거)라 한다. 사후분포의 평균, 중앙값, 최빈값을 각각 사후평균, 사후중앙값, MAP라 한다. 적분이 유한하지 않은 "사전분포"(예컨대 실수 전체의 균등)를 **비정칙** 사전분포라 한다.

1

Laplace의 역확률 (1763–1812)

> **역사** Thomas Bayes의 논문은 그가 죽은 뒤 1763년 Richard Price가 발표했다. 당구대의 공이 멈춘 자리처럼 모르는 확률 p에 균등 사전분포를 주고, 관측 뒤의 분포를 구했다. 11년 뒤 25세의 Laplace는 Bayes를 모른 채 같은 원리를 "사건으로부터 원인의 확률"이라 부르며 일반화했다. 그는 이 계산을 출생 기록에 썼다. 파리의 1745–1770년 기록에서 남아가 여아보다 많은 것이 우연인지, 런던의 남아 비율이 파리보다 큰 것이 우연인지를 물었다. 적분이 손으로 되지 않자, 가능도의 꼭대기 근처를 정규곡선으로 근사하는 방법을 만들었다. 오늘 "Laplace 근사"라 부르는 것이다. 1812년의 **확률의 해석적 이론**은 이 모든 것을 담았다.

1.1 사후분포와 켤레 사전분포

정의 1.1 (사후분포) 모수 θ가 사전분포 π를 따르고, $\theta=\theta$가 주어졌을 때 자료가 $f(\mathbf{x}\mid\theta)$를 따른다고 하자. 자료가 주어졌을 때 θ의 조건부분포

$$\pi(\theta\mid\mathbf{x})=\frac{\pi(\theta)\,f(\mathbf{x}\mid\theta)}{m(\mathbf{x})},\qquad m(\mathbf{x})=\int \pi(\theta')f(\mathbf{x}\mid\theta')\,d\theta'$$

를 **사후분포**라 한다. 1회 정리 5.5(Bayes 정리)의 밀도판이다.

사후분포는 가능도에 사전분포를 곱한 것에 비례한다. 9회의 MLE가 가능도의 꼭대기 하나를 고른다면, 베이즈 추론은 가능도의 모양 전체를 사전분포로 가중해 분포로 남긴다. 사전분포가 평평하면 사후분포의 꼴은 가능도의 꼴과 같다.

정리 1.2 (켤레 사전분포)

(a) (베타–이항) $p\sim\mathrm{Beta}(a,b)$, $S\mid p\sim\mathrm{b}(n,p)$이면 $p\mid S=s\sim\mathrm{Beta}(a+s,\ b+n-s)$.

(b) (감마–Poisson) $\lambda\sim\Gamma(\alpha,\ \text{율}\ \beta)$, $X_1,\dots,X_n\mid\lambda\overset{\text{iid}}{\sim}\mathrm{Poi}(\lambda)$이면 $\lambda\mid\mathbf{x}\sim\Gamma(\alpha+\sum x_i,\ \text{율}\ \beta+n)$.

(c) (정규–정규) $\mu\sim\mathcal{N}(m,\tau^2)$, $X_1,\dots,X_n\mid\mu\overset{\text{iid}}{\sim}\mathcal{N}(\mu,\sigma^2)$ (σ^2은 앎)이면

$$\mu\mid\mathbf{x}\sim\mathcal{N}\!\left(\frac{(n/\sigma^2)\bar{x}+(1/\tau^2)m}{n/\sigma^2+1/\tau^2},\ \frac{1}{n/\sigma^2+1/\tau^2}\right).$$

증명. **전략** 사후밀도는 $\pi(\theta)f(\mathbf{x}\mid\theta)$에 비례한다. θ에 의존하는 인수만 모아 이미 아는 분포의 밀도의 핵(상수배를 뺀 부분)과 맞춘다. 밀도는 적분이 1이므로 상수는 저절로 정해진다.

(a)

① $\pi(p \mid s) \propto p^{a-1}(1-p)^{b-1} \cdot \binom{n}{s} p^s (1-p)^{n-s} \propto p^{a+s-1}(1-p)^{b+n-s-1}$ 이다. 정의 1.1

② 이것은 $\mathrm{Beta}(a+s,\ b+n-s)$ 의 핵이다. 4회 정의 4.1, 정리 4.5

(b)

① 관측이 독립이므로 $f(\mathbf{x} \mid \lambda) = \prod_i \lambda^{x_i} e^{-\lambda}/x_i!$ 이다. 3회 정의 3.1

② λ 에 의존하는 인수만 남기면

$$\pi(\lambda \mid \mathbf{x}) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \cdot \lambda^{\sum x_i} e^{-n\lambda} = \lambda^{\alpha+\sum x_i - 1} e^{-(\beta+n)\lambda}.$$ 지수법칙

③ 이것은 모양 $\alpha + \sum x_i$, 율 $\beta + n$ 인 감마분포의 핵이다. 4회 정의 3.6

(c)

① 가능도의 μ 에 대한 부분은 $\exp\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\}$ 이다. $\sum_i (x_i-\mu)^2 = \sum_i (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$

② 사전밀도를 곱하면 지수는 $-\frac{n}{2\sigma^2}(\mu-\bar{x})^2 - \frac{1}{2\tau^2}(\mu-m)^2$ 이다. 정의 1.1

③ $P = n/\sigma^2 + 1/\tau^2$, $M = \big((n/\sigma^2)\bar{x} + m/\tau^2\big)/P$ 로 두고 μ 에 대해 제곱을 완성하면

$$-\frac{n}{2\sigma^2}(\mu-\bar{x})^2 - \frac{1}{2\tau^2}(\mu-m)^2 = -\frac{P}{2}(\mu-M)^2 + (\mu\text{에 무관한 상수}).$$ μ^2, μ 의 계수 비교

④ 이것은 $\mathcal{N}(M, 1/P)$ 의 핵이다. 5회 연습 1.6의 계산과 같다. 5회 정의 1.7

■

(c)의 사후평균은 자료의 평균 $\bar{x}$ 를 사전평균 m 쪽으로 **축소**한 것이다. 축소의 비율은 두 정보의 비, 곧 n/σ^2 대 $1/\tau^2$ 다. 자료가 많아지면 사전분포의 몫은 사라진다. 5회 따름정리 2.2(Galton의 2단 quincunx)와 같은 식이고, 5절의 James–Stein이 이 식으로 돌아온다.

> **예제 1.3** (파리의 남아) 균등 사전분포(Bayes의 당구대)로 남아 251,527명, 여아 241,945명을 넣으면 $p \mid$ 자료 $\sim \mathrm{Beta}(251528,\ 241946)$ 이다. 사후평균은 0.50971, 95% 신용구간은 $[0.50831,\ 0.51110]$ 이고
>
> $$P\big(p \le \tfrac{1}{2} \mid \text{자료}\big) = 1.15 \times 10^{-42}$$
>
> 이다. Laplace의 결론대로, 남아가 더 많이 태어난다는 것은 사실상 확실하다. 같은 계산을 Arbuthnot(1710)의 런던 세례 기록(1629–1710년, 남아 484,382명, 여아 453,841명)에 하면 런던의 남아 비율은 0.5163이고, 사후분포에서 런던과 파리의 차이는 0.0066, 95% 구간 $[0.0048,\ 0.0083]$ 이다. 런던의 비율이 더 크다는 것도 우연이 아니다(그림 1.1). Laplace가 비교에 쓴 런던 자료는 다른 기간의 것이지만 결론은 같았다.

1.2 Laplace 근사

명제 1.4 (Laplace 근사) h가 꼭대기 $\hat\theta$에서 두 번 미분가능하고 $h''(\hat\theta)<0$이며, 꼭대기 밖에서는 h가 $h(\hat\theta)$보다 일정량 이상 작다고 하자. $n\to\infty$일 때

$$\int e^{nh(\theta)}\,d\theta \sim e^{nh(\hat\theta)}\sqrt{\frac{2\pi}{n\,|h''(\hat\theta)|}}.$$

따라서 사후밀도 $\propto e^{nh(\theta)}$는 $\mathcal{N}\!\left(\hat\theta,\ 1/(n|h''(\hat\theta)|)\right)$로 근사된다.

증명. $\hat\theta$ 근처에서 $nh(\theta)\approx nh(\hat\theta)-\frac{n}{2}|h''(\hat\theta)|(\theta-\hat\theta)^2$이고, 오른쪽의 지수를 적분하면 정규 적분(3회)이 된다. 꼭대기에서 δ 이상 떨어진 곳의 적분은 $e^{n(h(\hat\theta)-c)}$보다 작아 상대적으로 사라진다. 나머지 항을 다루는 세부는 Stirling 공식(5회 보조정리 1.1)의 증명과 같다. ∎

이항 가능도에서 $h(p)=\hat p\log p+(1-\hat p)\log(1-p)$이므로 $\hat\theta=\hat p$, $1/(n|h''|)=\hat p(1-\hat p)/n$이다. 파리의 자료에 쓰면 $(0.5-\hat p)/\sqrt{\hat p(1-\hat p)/n}=-13.64$이고 $\Phi(-13.64)=1.11\times10^{-42}$로, 정확한 값 1.15×10^{-42}와 3% 차이다. 정규분포의 13.6 표준편차 밖의 꼬리를 이 정도로 맞춘다는 것이 놀랍다. 사후분포가 MLE를 중심으로, 9회의 Fisher 정보의 역수를 분산으로 하는 정규분포에 가까워진다는 것이 일반적인 사실이다.

> **증명 없이 쓰는 사실** (Bernstein–von Mises 정리) 정칙 모수모형에서, 참값에서 양의 밀도를 갖는 사전분포를 쓰면 사후분포와 $\mathcal{N}\!\left(\hat\theta_{\mathrm{MLE}},\ I_n(\theta_0)^{-1}\right)$의 총변동 거리는 확률적으로 0으로 간다. 표본이 크면 베이즈 신용구간과 9회의 Wald 신뢰구간이 같아진다. 증명에는 9회의 MLE 점근론과 사후분포의 꼬리를 묶는 조건이 필요하다.

연습문제

연습 1.1 (사후평균은 가중평균 ★) 베타–이항에서 사후평균 $\frac{a+s}{a+b+n}$이 사전평균 $\frac{a}{a+b}$와 표본비율 s/n의 가중평균이고, 무게가 $\frac{a+b}{a+b+n}$과 $\frac{n}{a+b+n}$임을 보여라. $a+b$를 "사전 표본 크기"라 부르는 이유는? 풀이 44쪽

연습 1.2 (감마–Poisson ★) 4회의 기병대 자료(200군단·년에 사망 122건)에 $\Gamma(1,\ 율\ 1)$ 사전분포를 주고 λ의 사후평균과 95% 신용구간을 구하라. MLE 0.61과 비교하라. 풀이 44쪽

연습 1.3 (Laplace 근사의 오차 ★★) 파리의 자료에서 정규근사의 꼬리확률 1.11×10^{-42}와 정확한 값 1.15×10^{-42}의 상대오차가 3% 정도인 이유를, 베타분포가 꼬리에서 정규분포와 어떻게 다른지(치우침)로 설명하라. 풀이 44쪽

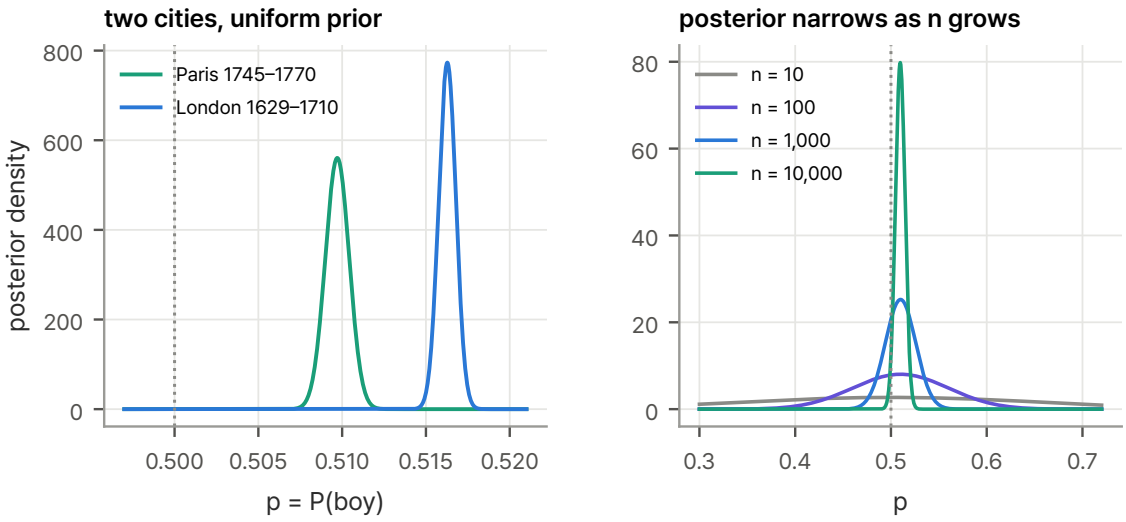


그림 1.1 왼쪽: 파리(1745–1770)와 런던(1629–1710)의 남아 출생 확률의 사후분포. 회색 점선은 1/2이다. 두 분포 모두 1/2에서 수십 표준편차 떨어져 있고 서로도 겹치지 않는다. 오른쪽: 균등 사전분포에서 자료가 늘 때 사후분포가 좁아지는 모습(파리의 처음 n명을 같은 비율로 축소한 가상의 자료).

> **연습 1.4** (예측분포 ★★) 베타–이항에서 다음 하나가 성공일 사후예측확률이 사후평균 $\frac{a+s}{a+b+n}$ 임을 보여라. $a = b = 1$이면 4회 따름정리 4.6(계승규칙)이다. 풀이 44쪽

2

Bayes 추정과 Jeffreys 사전분포

> **역사** Laplace(1774)는 사후분포에서 값 하나를 고를 때 "오차의 절댓값의 기댓값"을 줄이는 값, 곧 사후중앙값을 골랐다. Gauss(1809)는 평평한 사전분포의 사후 최빈값으로 최소제곱을 정당화했다(5회). Wald(1950)는 이것들을 손실함수의 언어로 묶었다(12회). 남은 문제는 사전분포였다. Bayes와 Laplace의 균등분포는 모수를 어떻게 쓰느냐에 따라 달라진다. p에 균등이면 p^2에는 균등이 아니다. Fisher가 1922년 역확률을 버린 이유 중 하나가 이것이었다. Harold Jeffreys는 1946년 이 반론에 답했다. 사전분포를 Fisher 자신의 정보량으로 만들면 모수를 바꿔 써도 결론이 같다.

2.1 손실함수와 Bayes 추정량

12회 정의 5.1의 언어를 다시 쓴다. 추정량 δ의 위험은 $R(\theta,\delta)=E_\theta L(\theta,\delta(\mathbf{X}))$, 사전분포 π에 대한 Bayes 위험은 $r(\pi,\delta)=\int R(\theta,\delta)\pi(\theta)d\theta$이다. 12회 정리 5.2에 따라 Bayes 추정량은 $\mathbf{x}$마다 사후 기대손실을 최소화하는 값이다.

정리 2.1 (Bayes 추정량) 사후분포 $\pi(\theta\mid\mathbf{x})$에 대해

(a) 제곱손실 $L=(\theta-a)^2$의 Bayes 추정량은 사후평균 $E(\theta\mid\mathbf{x})$이고, 최소 사후 기대손실은 사후분산이다.

(b) 절대손실 $L=|\theta-a|$의 Bayes 추정량은 사후중앙값이다.

(c) 가중 제곱손실 $L=w(\theta)(\theta-a)^2$의 Bayes 추정량은 $E(w(\theta)\theta\mid\mathbf{x})/E(w(\theta)\mid\mathbf{x})$이다.

증명. **전략** $\mathbf{x}$를 고정하면 사후 기대손실은 a 하나의 함수다. 그 함수를 펼쳐 최소점을 찾는다. (a)는 제곱을 평균과 편차로 나누고, (b)는 미분하고, (c)는 a의 이차식으로 쓴다.

(a)

① $m=E(\theta\mid\mathbf{x})$로 두고 $\theta-a=(\theta-m)+(m-a)$를 제곱해 사후 기대값을 취한다.

$$E[(\theta-a)^2\mid\mathbf{x}]=E[(\theta-m)^2\mid\mathbf{x}]+2(m-a)E[\theta-m\mid\mathbf{x}]+(m-a)^2. \qquad \text{선형성}$$

② 가운데 항은 $E[\theta-m\mid\mathbf{x}]=0$이라 사라진다. 3회 정리 2.8과 같은 계산이다. m의 정의

③ 남은 것은 $\mathrm{Var}(\theta\mid\mathbf{x})+(m-a)^2$이고, $a=m$에서 최소이며 최솟값은 사후분산이다. $(m-a)^2\ge 0$

(b)

① $g(a) = E[|\theta - a| \mid \mathbf{x}] = \int_{-\infty}^{a}(a-\theta)\pi(\theta \mid \mathbf{x})\,d\theta + \int_{a}^{\infty}(\theta - a)\pi(\theta \mid \mathbf{x})\,d\theta$이다. 절댓값을 부호로 나눈다

② a로 미분하면 적분 끝점에서 나오는 항은 $(a-a)\pi(a \mid \mathbf{x}) = 0$이라 사라지고

$$g'(a) = \int_{-\infty}^{a}\pi(\theta \mid \mathbf{x})\,d\theta - \int_{a}^{\infty}\pi(\theta \mid \mathbf{x})\,d\theta = P(\theta < a \mid \mathbf{x}) - P(\theta > a \mid \mathbf{x}). \quad \text{[T6]}$$

③ g'은 a에 대해 증가하고, 사후중앙값 $\tilde{m}$에서 $g'(\tilde{m}) = \frac{1}{2} - \frac{1}{2} = 0$이다. 분포함수의 단조성

④ $a < \tilde{m}$이면 $g' \le 0$, $a > \tilde{m}$이면 $g' \ge 0$이므로 g는 $\tilde{m}$에서 최소다. 평균값 정리

(c)

① $h(a) = E[w(\theta)(\theta - a)^2 \mid \mathbf{x}] = E[w\theta^2 \mid \mathbf{x}] - 2aE[w\theta \mid \mathbf{x}] + a^2 E[w \mid \mathbf{x}]$이다. 전개, 선형성

② $E[w \mid \mathbf{x}] > 0$이면 h는 위로 볼록한 이차식이고, 꼭짓점 $a = E[w\theta \mid \mathbf{x}]/E[w \mid \mathbf{x}]$에서 최소다. $h'(a) = 0$

조건 확인 (a)는 $E(\theta^2 \mid \mathbf{x}) < \infty$, (b)는 $E(|\theta| \mid \mathbf{x}) < \infty$, (c)는 $E(w\theta^2 \mid \mathbf{x}) < \infty$를 쓴다. 그렇지 않으면 사후 기대손실이 모든 a에서 무한하다. (b)의 미분은 적분 구간의 끝이 a에 달린 경우의 미분이다. 사후밀도가 있으면 미적분의 기본정리로 끝점의 항을, 피적분함수 $a-\theta$의 a-미분 1로 나머지를 얻는다 [T6]. 사후분포에 밀도가 없어도 g는 볼록이고 같은 결론이 오른쪽·왼쪽 미분으로 나온다. ∎

0–1 손실 $L = \mathbf{1}\{|\theta - a| > \epsilon\}$에서 $\epsilon \to 0$이면 Bayes 추정량은 사후 최빈값(MAP)으로 간다. 평평한 사전분포의 MAP는 MLE다. 손실함수가 바뀌면 같은 사후분포에서 다른 점추정량이 나온다는 것이 핵심이다. 사후분포가 대칭이면 셋은 같다.

예제 2.2 (이항의 Bayes 추정량) $\mathrm{Beta}(a,b)$ 사전분포에서 사후평균 $\delta_{a,b} = \frac{S+a}{n+a+b}$의 위험은 편향과 분산으로 나누어

$$R(p, \delta_{a,b}) = \frac{np(1-p) + \big(a - (a+b)p\big)^2}{(n+a+b)^2}$$

이다. $a = b = \sqrt{n}/2$로 고르면 분자의 p^2항과 p항이 지워져 $R = \frac{n}{4(n+\sqrt{n})^2}$, p에 무관한 상수가 된다. $n = 20$이면 이 위험은 0.00835이고, MLE S/n의 위험 $p(1-p)/20$의 최댓값 0.0125보다 작다. 대신 $p < 0.212$나 $p > 0.788$에서는 MLE가 낫다.

정리 2.3 (Bayes 추정량의 두 성질)

(a) (허용성) 유한한 Bayes 위험을 갖는 Bayes 추정량이 유일하면, 그것은 **허용적**이다. 곧 모든 θ에서 위험이 같거나 작고 어떤 θ에서 더 작은 추정량은 없다.

(b) (minimax) Bayes 추정량의 위험이 θ에 무관한 상수이면 그것은 minimax다.

증명. **전략** (a)는 귀류법이다 [T2]. 지배하는 추정량이 있으면 그것도 Bayes 추정량이 되어 유일성과 부딪친다. (b)는 "최댓값은 평균보다 작지 않다"를 한 번 쓴다.

(a)

① δ'이 δ_π를 지배한다고 하자. 모든 θ에서 $R(\theta,\delta') \le R(\theta,\delta_\pi)$이고 어떤 θ_0에서 부등호가 엄격하다. — 귀류 가정 [T2]

② 양변을 π로 적분하면 $r(\pi,\delta') \le r(\pi,\delta_\pi)$이다. — 적분의 단조성

③ δ_π는 Bayes 위험을 최소로 하므로 $r(\pi,\delta') = r(\pi,\delta_\pi)$, 곧 δ'도 Bayes 추정량이다. — 12회 정리 5.2

④ 유일성으로 $\delta' = \delta_\pi$이고, 두 위험은 모든 θ에서 같다. ①의 θ_0과 모순이다. — 가정: 유일

(b)

① 임의의 δ에 대해 $\sup_\theta R(\theta,\delta) \ge \int R(\theta,\delta)\pi(\theta)\,d\theta = r(\pi,\delta)$이다. — 평균은 최댓값을 넘지 않는다

② $r(\pi,\delta) \ge r(\pi,\delta_\pi)$이다. — δ_π는 Bayes

③ $R(\theta,\delta_\pi) \equiv c$이므로 $r(\pi,\delta_\pi) = c = \sup_\theta R(\theta,\delta_\pi)$이다. 따라서 $\sup_\theta R(\theta,\delta) \ge \sup_\theta R(\theta,\delta_\pi)$이다. — 위험이 상수

조건 확인 (a)의 ③에서 $r(\pi,\delta_\pi) < \infty$를 쓴다. Bayes 위험이 무한하면 모든 추정량의 Bayes 위험이 ∞라 ②와 ③이 아무것도 말하지 않는다. 유일성 대신 다음 조건을 써도 된다. 모든 추정량의 위험이 θ에 연속이고 π가 모든 열린구간에 양의 확률을 주면, ①의 엄격한 부등호가 θ_0의 한 근방에서 성립하므로 ②의 부등호가 엄격해지고, δ_π가 Bayes라는 것과 모순이다. ∎

따라서 예제 2.2의 $\delta_{\sqrt{n}/2,\sqrt{n}/2}$는 minimax다. 허용성은 이 회차의 중심 개념이다. 허용적이지 않은 추정량은 다른 추정량에 "모든 곳에서 진다". 5절은 표본평균이 세 차원 이상에서 바로 그런 추정량임을 보인다.

명제 2.4 (불편추정량은 Bayes가 아니다) 제곱손실에서 $\delta(\mathbf{X})$가 불편이면서($E(\delta \mid \theta) = \theta$) Bayes 추정량이고($\delta = E(\theta \mid \mathbf{X})$) 둘째 적률이 유한하면, Bayes 위험은 $E(\delta - \theta)^2 = 0$이다.

증명. **전략** $(\theta, \mathbf{X})$의 결합분포에서 $E(\delta\theta)$를 두 가지 조건화로 계산한다(3회 정리 2.5). 한 번은 불편성, 한 번은 Bayes 추정량의 정의를 쓴다. 단계별 증명은 연습 2.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 45쪽에 있다.

11회의 UMVUE는 사실상 어떤 정칙 사전분포의 Bayes 추정량도 아니다. 예컨대 $\bar{X}$는 정규 사전분포 $\mathcal{N}(m, \tau^2)$의 사후평균(정리 1.2(c))의 $\tau \to \infty$ 극한일 뿐이다. 이런 극한을 **일반화 Bayes** 추정량이라 한다. 극한은 허용성을 물려받지 않을 수 있다.

2.2 Jeffreys 사전분포

균등 사전분포는 모수화에 달려 있다. $p \sim \mathrm{U}(0,1)$이면 로그오즈 $\psi = \log \frac{p}{1-p}$의 밀도는 $e^{\psi}/(1+e^{\psi})^2$(로지스틱)이고 균등이 아니다. 어느 모수화에서 "모름"을 균등으로 나타내야 하는가? Jeffreys의 답은 모수화를 고르지 않는 규칙이었다.

정의 2.5 (Jeffreys 사전분포) 한 관측의 Fisher 정보 $I(\theta)$(9회 정의 2.1)에 대해 $\pi_J(\theta) \propto \sqrt{I(\theta)}$를 **Jeffreys 사전분포**라 한다. 다차원이면 $\pi_J(\theta) \propto \sqrt{\det I(\theta)}$이다.

정리 2.6 (Jeffreys 사전분포의 불변성) $\phi = h(\theta)$가 미분가능한 단조함수이면, θ에서 Jeffreys 사전분포를 만들어 ϕ로 변수변환한 밀도는 ϕ에서 직접 만든 Jeffreys 사전분포와 같다.

증명. **전략** 두 밀도를 각각 계산해 같은 식이 나옴을 보인다. 하나는 Fisher 정보가 모수를 바꿀 때 어떻게 변하는지에서, 다른 하나는 밀도의 변수변환 공식에서 나온다.

① $\theta = h^{-1}(\phi)$로 쓰면 연쇄법칙으로 $\dfrac{\partial}{\partial\phi}\log f(x \mid \phi) = \dfrac{\partial}{\partial\theta}\log f(x \mid \theta)\cdot\dfrac{d\theta}{d\phi}$이다. — 연쇄법칙

② 제곱해 기댓값을 취하면 $\dfrac{d\theta}{d\phi}$는 관측에 무관한 상수이므로

$$I_\phi(\phi) = E\Big[\Big(\frac{\partial}{\partial\theta}\log f\Big)^2\Big]\Big(\frac{d\theta}{d\phi}\Big)^2 = I_\theta(\theta)\Big(\frac{d\theta}{d\phi}\Big)^2.$$ — 9회 정의 2.1, 연습 2.2

③ 제곱근을 취하면 $\sqrt{I_\phi(\phi)} = \sqrt{I_\theta(\theta)}\left|\dfrac{d\theta}{d\phi}\right|$이다. — $\sqrt{u^2} = |u|$

④ 밀도 $\pi_J(\theta) \propto \sqrt{I_\theta(\theta)}$를 단조변환 $\phi = h(\theta)$로 옮긴 밀도는 $\pi_J(h^{-1}(\phi))\,|d\theta/d\phi|$이고, 이것이 ③의 오른쪽이다. — 2회 정리 4.3 [T7]

조건 확인 h는 미분가능하고 도함수가 0이 아니어야 한다(그래야 $d\theta/d\phi$가 존재한다). ②는 9회의 정칙 조건 아래에서 I_θ가 유한할 때의 식이다. Jeffreys 사전분포가 비정칙이면 두 밀도는

상수배를 빼고 같다는 뜻이다. 변수변환 공식은 상수배와 함께 성립하므로 결론이 그대로다. ■

예제 2.7 (Jeffreys 사전분포의 예)

(a) 베르누이: $I(p)=\frac{1}{p(1-p)}$ 이므로 $\pi_J(p)\propto p^{-1/2}(1-p)^{-1/2}$, 곧 $\mathrm{Beta}(\frac12,\frac12)$(arcsine 분포). $\phi=\arcsin\sqrt{p}$로 쓰면 I_ϕ가 상수이므로, Jeffreys 사전분포는 7회의 분산안정화 척도 위의 균등분포다.

(b) Poisson: $I(\lambda)=1/\lambda$이므로 $\pi_J\propto\lambda^{-1/2}$. 비정칙이지만 사후분포 $\Gamma(\sum x_i+\frac12,\ \text{율 } n)$은 $\sum x_i=0$이어도 정칙이다.

(c) 정규 위치 $\mathcal{N}(\mu,1)$: I가 상수이므로 평평한 사전분포. 정규 척도 $\mathcal{N}(0,\sigma^2)$: $I(\sigma)=2/\sigma^2$이므로 $\pi_J\propto 1/\sigma$, 곧 $\log\sigma$에 평평하다.

Jeffreys 사전분포의 신용구간은 빈도주의의 기준으로도 좋다. $\mathrm{b}(20,p)$에서 $\mathrm{Beta}(s+\frac12,\ 20-s+\frac12)$의 2.5%·97.5% 분위수를 **Jeffreys 구간**이라 하자. $p\in[0.05,0.95]$의 91개 점에서 정확한 포함확률을 계산하면 평균 0.948, 최솟값 0.896이다. 8회의 Wald 구간 $\hat p\pm1.96\sqrt{\hat p(1-\hat p)/n}$은 평균 0.895, 최솟값 0.639다(그림 2.1). 끝 근처에서 Wald 구간은 $\hat p=0$일 때 길이가 0이 되어 실패한다. 사전분포의 $\frac12$이 그 구멍을 막는다.

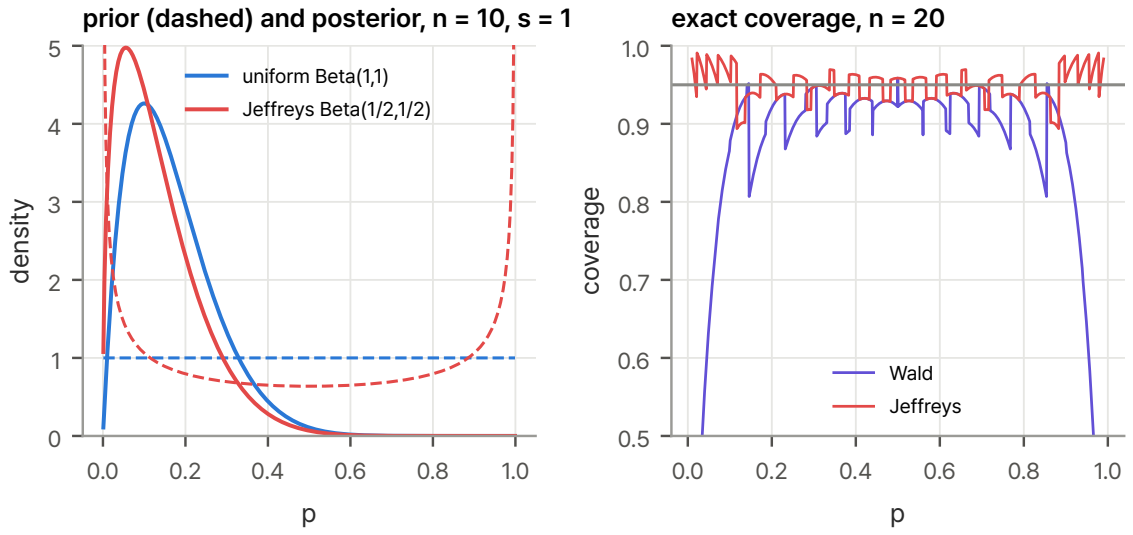


그림 2.1 왼쪽: 같은 자료에 균등($\mathrm{Beta}(1,1)$)과 Jeffreys($\mathrm{Beta}(\frac12,\frac12)$) 사전분포를 쓴 사후분포. 오른쪽: $n=20$에서 Jeffreys 구간과 Wald 구간의 정확한 포함확률. 회색 선은 0.95다. Wald 구간은 p가 끝에 가까울수록 무너진다.

증명 없이 쓰는 사실 (다차원의 어려움과 기준 사전분포) 다차원에서 $\sqrt{\det I}$는 문제를 일으킨다. $\mathcal{N}(\mu,\sigma^2)$에서 $\sqrt{\det I}\propto 1/\sigma^2$를 쓰면 σ^2의 사후분포의 자유도가 $n-1$이 아니라 n이 된다. Jeffreys 자신도 위치와 척도를 따로 다뤄 $1/\sigma$를 권했다. Bernardo(1979)의 **기준 사전분포**는 관심 모수와 장애 모수를 나누어 이 문제를 체계적으로 푼다.

연습문제

연습 2.1 (minimax 추정량 ★) 예제 2.2의 위험 공식을 유도하고, $a=b=\sqrt{n}/2$에서 위험이 상수임을 확인하라. $n=20$에서 MLE가 minimax 추정량보다 나은 p의 범위를 구하라. 풀이 45쪽

연습 2.2 ($\bar X$는 Bayes가 아니다 ★) $X_1,\dots,X_n\mid\mu\overset{\text{iid}}{\sim}\mathcal{N}(\mu,1)$, $\mu\sim\mathcal{N}(0,\tau^2)$에서 명제 2.4를 써서 $\bar X$가 Bayes 추정량이 아님을 보여라. 사후평균의 Bayes 위험을 구하고 $\tau\to\infty$일 때의 극한을 말하라. 풀이 45쪽

연습 2.3 (정규 척도와 카이제곱 ★★) $X_1,\dots,X_n\mid\sigma^2\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma^2)$, $\pi(\sigma)\propto 1/\sigma$이면 $\sum X_i^2/\sigma^2\mid\mathbf{x}\sim\chi^2_n$임을 보여라. 따라서 95% 신용구간은 6회 연습 2.5의 카이제곱 신뢰구간(평균을 알아 자유도가 n인 판)과 같다. 풀이 45쪽

연습 2.4 (로그오즈의 균등 ★★) 로그오즈 ψ에 평평한(비정칙) 사전분포를 주면 p의 사전분포는 $\propto p^{-1}(1-p)^{-1}$(Haldane)이다. 이것을 보이고, $s=0$일 때 사후분포가 비정칙임을 보여라. 풀이 45쪽

연습 2.5 (절대손실의 Laplace ★) 파리의 사후분포 $\mathrm{Beta}(251528,241946)$에서 사후평균과 사후중앙값의 차이는 10^{-6}보다 작다. 그 이유를 사후분포의 치우침으로 설명하라. 풀이 45쪽

연습 2.6 (불편추정량은 Bayes가 아니다: 증명 ★★ 유도형) 명제 2.4를 다음 순서로 증명하라. $(\theta,\mathbf{X})$의 결합분포에서 기댓값을 잡는다.

(a) θ로 조건화하고 불편성 $E(\delta\mid\theta)=\theta$를 써서 $E(\delta\theta)=E\theta^2$을 보여라.

(b) $\mathbf{X}$로 조건화하고 $\delta=E(\theta\mid\mathbf{X})$를 써서 $E(\delta\theta)=E\delta^2$을 보여라.

(c) (a)와 (b)로 $E(\delta-\theta)^2=0$을 얻어라. 둘째 적률이 유한하다는 가정은 어디에 쓰이는가?

(d) 결론을 대우로 바꿔 말하라. Bayes 위험이 양수인 모형에서는 어느 추정량도 불편이면서 Bayes일 수 없다.

풀이 45쪽

3

Fisher의 거부와 증거의 무게

역사 Fisher는 1922년 논문에서 역확률을 "완전히 버려야 할 것"이라 썼다. 균등 사전분포가 모수화에 달려 있다는 것(2절)과, 사전분포가 관측되지 않는다는 것이 이유였다. 그는 그 자리에 가능도와 유의확률을, Neyman과 Pearson은 오류율을 놓았다(8·12회). 케임브리지의 지구물리학자 Harold Jeffreys는 홀로 다른 길을 갔다. 그의 *Theory of Probability*(1939)는 가설 검정을 사후확률로 다시 세웠다. 귀무가설에 사전확률 덩어리를 주고, 두 가설의 주변 가능도의 비, 곧 **Bayes 인수**로 판정했다. 1940–41년 Bletchley Park에서 Alan Turing은 Enigma의 설정을 좁히기 위해 같은 양을 썼다. 그는 가능도비의 상용로그를 "ban", 그 10분의 1을 "deciban"이라 부르고 단서마다 더했다. 조수였던 I. J. Good은 전쟁 뒤 이것을 "증거의 무게"로 정리했다(1950). 미국에서 Wald가 같은 시기에 축차확률비 검정(12회)을 만든 것은 우연이 아니다. 두 사람 모두 로그 가능도비를 증거가 들어오는 대로 더했다.

3.1 Bayes 인수

정의 3.1 (Bayes 인수) 두 모형 $H_0 : \theta \sim \pi_0$, $H_1 : \theta \sim \pi_1$의 주변 가능도 $m_j(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\pi_j(\theta)d\theta$의 비

$$B_{10}(\mathbf{x}) = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})}$$

를 H_1의 H_0에 대한 **Bayes 인수**라 한다. 단순가설(π_j가 한 점)이면 Bayes 인수는 가능도비다. $10\log_{10} B_{10}$을 **증거의 무게**(단위 deciban)라 한다.

명제 3.2 (사후 오즈와 증거의 덧셈)

(a) 사전확률이 $P(H_0), P(H_1)$이면 $\dfrac{P(H_1 \mid \mathbf{x})}{P(H_0 \mid \mathbf{x})} = \dfrac{P(H_1)}{P(H_0)} \cdot B_{10}(\mathbf{x})$.

(b) 자료가 차례로 들어오면 $\log B_{10}(\mathbf{x}_{1:n}) = \sum_{k=1}^{n} \log \dfrac{m_1(\mathbf{x}_k \mid \mathbf{x}_{1:k-1})}{m_0(\mathbf{x}_k \mid \mathbf{x}_{1:k-1})}$이다. 단순가설이고 관측이 독립이면 각 항은 그 관측의 로그 가능도비다.

증명. (a) $P(H_j \mid \mathbf{x}) \propto P(H_j)m_j(\mathbf{x})$(1회 정리 5.5). (b) $m_j(\mathbf{x}_{1:n}) = \prod_k m_j(\mathbf{x}_k \mid \mathbf{x}_{1:k-1})$(곱셈법칙)의 로그를 취한다. 단순가설이면 $m_j(\mathbf{x}_k \mid \mathbf{x}_{1:k-1}) = f(\mathbf{x}_k \mid \theta_j)$. ∎

(b)는 Turing의 방법이자 12회 정의 4.1의 축차확률비 검정이다. 증거는 단서마다 더해지고, 누적 무게가 미리 정한 문턱을 넘으면 결정한다. Good에 따르면 Turing은 1 deciban을 사람이 직관으로 느낄 수 있는 가장 작은 증거의 변화로 보았다. 10 deciban은 가능도비 10, 20 deciban은 100이다.

예제 3.3 (Arbuthnot의 증거의 무게) 8회에서 Arbuthnot의 82년은 부호검정의 유의확률 $2^{-82} \approx 2 \times 10^{-25}$로 읽었다. 이번에는 해마다의 출생 수를 다 써서 $H_1 : p = 0.516$(런던 전체의 비율)과 $H_0 : p = \tfrac{1}{2}$의 증거의 무게를 잰다. 남아 b_t명, 여아 g_t명인 해의 무게는 $10\big[b_t \log_{10} \frac{0.516}{0.5} + g_t \log_{10} \frac{0.484}{0.5}\big]$이다. 82년의 합은 2159 deciban이고 한 해에 평균 26.3 deciban이다. 한 해의 기록만으로도 가능도비가 약 400이다(그림 3.1). H_1을 하나의 값이 아니라 균등 사전분포 $p \sim \mathrm{U}(0,1)$로 두면 $\log_{10} B_{10} = 213.0$, 곧 2130 deciban이다. 29 deciban 차이는 H_1이 p를 미리 알지 못해 치르는 값이다.

3.2 Lindley의 역설

명제 3.4 (점 귀무가설의 Bayes 인수) $H_0 : \theta = \theta_0$, $H_1 : \theta \sim \pi_1$(θ_0에서 연속이고 양수)라 하고, 정칙 모형에서 $z_n = \sqrt{n I(\theta_0)}(\hat\theta_n - \theta_0)$라 하자. n이 크면

$$\log B_{01} = -\frac{z_n^2}{2} + \frac{1}{2}\log\frac{n I(\theta_0)}{2\pi} - \log \pi_1(\theta_0) + o_P(1).$$

따라서 z_n을 고정하면 $B_{01} \to \infty$이다. 유의확률이 0.05로 같아도 n이 클수록 자료는 H_0을 지지한다.

증명. 명제 1.4로 $m_1(\mathbf{x}) \approx f(\mathbf{x} \mid \hat\theta)\pi_1(\hat\theta)\sqrt{2\pi/(n I(\hat\theta))}$이다. $m_0(\mathbf{x}) = f(\mathbf{x} \mid \theta_0)$이고, 9회 정리 5.2(Wilks)의 증명에서 $2\log \frac{f(\mathbf{x}\mid\hat\theta)}{f(\mathbf{x}\mid\theta_0)} = z_n^2 + o_P(1)$이다. 비를 취하고 $\hat\theta \to \theta_0$을 쓴다. ∎

상수를 버리면 $\log B_{01} \approx \frac{1}{2}\log n - \frac{1}{2} z_n^2$이다. 모수가 d개 더 많은 모형과 비교하면 $\frac{d}{2}\log n$이 되고, 이것이 Schwarz(1978)의 **BIC**다. AIC(10회 연습 6.3)의 벌점 d와 달리 n과 함께 커진다.

예제 3.5 (Lindley의 출생 자료) 어느 도시에서 남아 49,581명, 여아 48,870명이 태어났다 ($n = 98{,}451$). $H_0 : p = \tfrac{1}{2}$의 정규근사 검정통계량은 $z = 2.27$, 양측 유의확률은 0.023이다. 5% 수준에서 기각한다. 그러나 $H_1 : p \sim \mathrm{U}(0,1)$의 Bayes 인수는 $B_{01} = \binom{n}{x} 2^{-n}(n+1) = 19.2$이고, 두 가설에 반씩 사전확률을 주면 $P(H_0 \mid \mathbf{x}) = 0.95$다. 명제 3.4의 근사도 $\log B_{01} = 2.9555$로 정확한 값 2.9555와 소수 넷째 자리까지 맞는다.

두 답이 다른 이유는 질문이 다르기 때문이다. 유의확률은 H_0 아래에서 이만큼 극단적인 자료의 확률을 묻는다. Bayes 인수는 H_0과 **구체적인 대립가설**이 각각 이 자료를 얼마나 잘 예측했는지 묻는다. 균등 대립가설은 $p = 0.3$이나 $p = 0.9$에도 무게를 주므로, 0.5036이라는 관측은 H_1에게도 놀랍다. 그렇다고 Bayes 인수가 옳은 것도 아니다. π_1을 넓히면 B_{01}은 한없이 커진다(연습 3.2). 검정의 답은 대립가설을 무엇으로 두느냐에 달려 있다.

> **증명 없이 쓰는 사실** (de Finetti의 표현정리) 0–1 수열 $X_1, X_2, \dots$ 의 결합분포가 순서를 바꿔도 같으면(**교환가능**), 어떤 분포 μ가 있어 $P(X_1 = x_1, \dots, X_n = x_n) = \int_0^1 p^s (1-p)^{n-s} \mu(dp)$이다 ($s = \sum x_i$). 곧 "모르는 p가 있고 그 아래에서 독립"이라는 Bayes 모형은, 관측의 교환가능성이라는 믿음에서 저절로 나온다. de Finetti(1937)와 Savage(1954)는 이것으로 사전분포를 주관적 믿음의 표현으로 정당화했다.

연습문제

연습 3.1 (deciban의 덧셈 ★) 독립인 두 단서의 가능도비가 3과 5다. 각각의 무게와 합친 무게를 deciban으로 구하고, 사전 오즈가 $1:20$일 때 사후 오즈를 구하라. 풀이 46쪽

연습 3.2 (정규 평균의 Bayes 인수 ★★) $X_1, \dots, X_n \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, $H_0 : \mu = 0$, $H_1 : \mu \sim \mathcal{N}(0, \tau^2)$이면

$$B_{01} = \sqrt{1 + n\tau^2}\, \exp\!\left(-\frac{z^2}{2} \cdot \frac{n\tau^2}{1 + n\tau^2}\right), \qquad z = \sqrt{n}\bar{x}$$

임을 보여라. z와 n을 고정하고 $\tau \to \infty$로 보내면 어떻게 되는가? 풀이 46쪽

연습 3.3 (파리의 Bayes 인수 ★) 파리의 자료에서 $H_0 : p = \frac{1}{2}$ 대 $H_1 : p \sim \mathrm{U}(0,1)$의 $\log_{10} B_{10}$을 명제 3.4의 근사로 구하라($z = 13.64$). 노트북의 정확한 값은 37.66이다. 풀이 46쪽

연습 3.4 (Pólya 항아리 ★★) 흰 공 a개, 검은 공 b개의 항아리에서 공을 꺼내 같은 색 공 하나를 더해 돌려놓는다. n번 중 흰 공이 s번인 특정 순서의 확률이 순서에 무관하고 $\int_0^1 p^s(1-p)^{n-s}\mathrm{Beta}(a,b)(dp)$임을 보여라. 풀이 46쪽

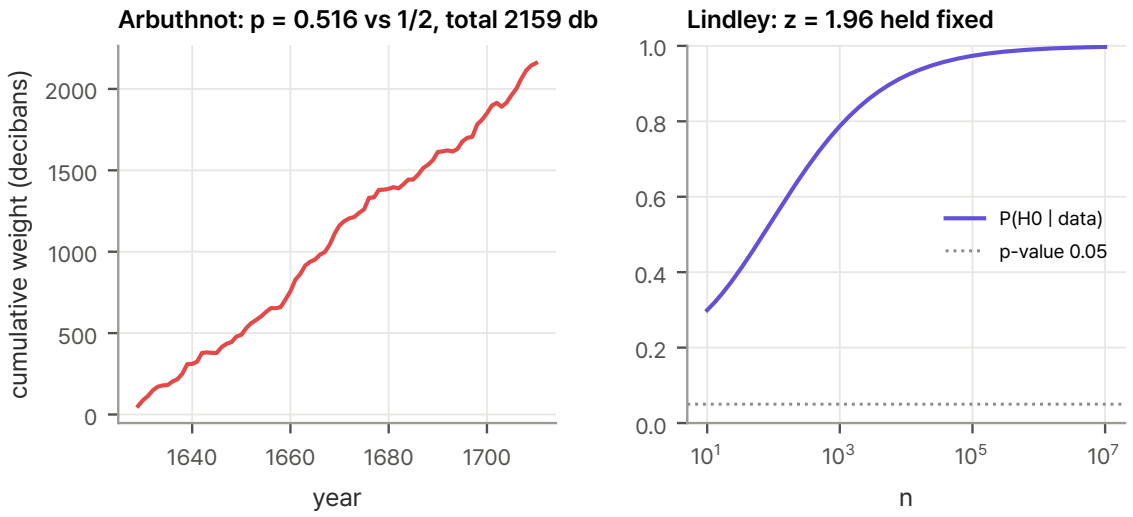


그림 3.1 왼쪽: Arbuthnot 자료에서 $H_1 : p = 0.516$ 대 $H_0 : p = \frac{1}{2}$의 누적 증거의 무게(deciban). 해마다 약 26 deciban씩 쌓인다. 오른쪽: 검정통계량을 $z = 1.96$(유의확률 0.05)으로 고정했을 때 n에 따른 $P(H_0 \mid \mathbf{x})$(H_1은 균등, 사전확률 반반). 유의확률은 그대로인데 사후확률은 1로 간다.

4

MCMC — 사후분포에서 뽑기

> **역사** 1953년 Los Alamos의 Nicholas Metropolis, Arianna와 Marshall Rosenbluth, Augusta와 Edward Teller는 MANIAC 컴퓨터로 단단한 공 224개로 된 액체의 상태방정식을 계산했다. 224×2 차원의 적분을 7회의 단순 몬테카를로로는 할 수 없었다. 대부분의 배치에서 공이 겹치기 때문이다. 그들은 공 하나를 조금 옮겨 보고, 에너지가 낮아지면 받아들이고 높아지면 확률적으로만 받아들였다. 이렇게 만든 마르코프 연쇄는 볼츠만 분포를 정상분포로 가졌다. 물리학의 도구였던 이 방법은 Hastings(1970)가 일반화하고, Geman 형제(1984)가 영상 복원에 쓴 Gibbs 표집을 거쳐, Gelfand와 Smith(1990)의 논문으로 통계학에 들어왔다. 정규화 상수 $m(\mathbf{x})$를 몰라도 사후분포에서 표본을 뽑을 수 있다는 것이 열쇠였다. 1990년대 BUGS 소프트웨어와 함께 베이즈 추론은 손 계산의 한계에서 풀려났다.

4.1 마르코프 연쇄와 정상분포

정의 4.1 (전이핵과 정상분포) 상태공간 위의 수열 $X_0, X_1, \dots$ 에서 X_{t+1}의 조건부분포가 X_t에만 달려 있고 그 밀도가 $k(x, y)$(**전이핵**)이면 **마르코프 연쇄**라 한다. 밀도 π가

$$\int \pi(x) k(x, y)\, dx = \pi(y) \qquad \text{모든 } y$$

를 만족하면 π를 **정상분포**라 한다. $X_t \sim \pi$이면 $X_{t+1} \sim \pi$다. 전이핵이 받아들이지 않는 확률만큼 제자리에 머무는 덩어리를 가질 수 있으며, 그때 적분은 그 덩어리를 포함한다.

정의 4.2 (상세 균형) $\pi(x)k(x,y) = \pi(y)k(y,x)$가 모든 x, y에서 성립하면 연쇄가 π에 대해 **상세 균형**(가역성)을 만족한다고 한다.

정리 4.3 (상세 균형이면 정상분포) 상세 균형을 만족하는 π는 정상분포다.

증명. 양변을 x로 적분하면 $\int \pi(x)k(x,y)dx = \pi(y)\int k(y,x)dx = \pi(y)$. ∎

상세 균형은 "x에서 y로 가는 확률의 흐름과 y에서 x로 오는 흐름이 같다"는 뜻이다. 정상성은 들어오는 흐름의 합과 나가는 흐름의 합만 같으면 되므로, 상세 균형은 충분조건일 뿐 필요조건은 아니다(연습 4.2). 그러나 국소적인 조건이라 확인하기 쉽고, 다음의 알고리즘은 그것을 만족하도록 설계된다.

4.2 Metropolis–Hastings와 Gibbs

알고리즘 4.4 (Metropolis–Hastings) 목표 밀도 π(상수배까지만 알면 된다)와 제안 밀도 $q(x,\cdot)$가 주어진다. 현재 상태 x에서

1. $y \sim q(x,\cdot)$를 제안한다.
2. 확률 $\alpha(x,y)=\min\left\{1,\ \dfrac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\}$로 y로 옮기고, 아니면 x에 머문다.

q가 대칭($q(x,y)=q(y,x)$, 예컨대 $y=x+\sigma\epsilon$, $\epsilon\sim\mathcal{N}(0,1)$)이면 $\alpha=\min\{1,\pi(y)/\pi(x)\}$이고 이것이 1953년의 Metropolis 알고리즘이다.

정리 4.5 (Metropolis–Hastings의 정상분포) 알고리즘 4.4의 전이핵은 π에 대해 상세 균형을 만족한다. 따라서 π는 정상분포다.

증명. **전략** 상세 균형을 직접 확인한다. $\pi(x)$를 받아들임 확률의 min 안으로 넣으면 x와 y에 대칭인 식이 된다.

① $x\neq y$이면 y로 옮기려면 y를 제안하고 받아들여야 하므로 $k(x,y)=q(x,y)\alpha(x,y)$이다. 알고리즘 4.4

② 양수를 min 안으로 곱하면

$$\pi(x)k(x,y)=\pi(x)q(x,y)\min\left\{1,\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right\}=\min\{\pi(x)q(x,y),\ \pi(y)q(y,x)\}.$$

$c\min\{a,b\}=\min\{ca,cb\},\ c>0$

③ 오른쪽은 x와 y를 바꿔도 같으므로 $\pi(x)k(x,y)=\pi(y)k(y,x)$이다. 대칭

④ $x=y$의 제자리 덩어리는 양변에 같은 항으로 나타나므로 상세 균형을 어기지 않는다. 정의 4.2

⑤ 따라서 π는 정상분포다. 정리 4.3

조건 확인 ②는 $\pi(x)q(x,y)>0$일 때의 계산이다. 연쇄가 $\pi(x)>0$인 점에서 출발하면 그 뒤로도 $\pi>0$인 점에만 머문다($\pi(y)=0$인 제안은 $\alpha=0$으로 거절된다). $\pi(x)q(x,y)=0$이고 $\pi(y)q(y,x)>0$인 쌍에서는 $\alpha(y,x)=0$이라 양변이 모두 0이다. π를 상수배까지만 알아도 α는 비만 쓰므로 계산이 바뀌지 않는다. ■

α에는 $\pi(y)/\pi(x)$만 들어 있다. 사후분포 $\pi(\theta\mid\mathbf{x})=\pi(\theta)f(\mathbf{x}\mid\theta)/m(\mathbf{x})$에서 $m(\mathbf{x})$는 비에서 지워진다. 1절에서 Laplace를 막았던 적분이 필요 없다.

알고리즘 4.6 (Gibbs 표집) $\theta = (\theta_1, \dots, \theta_d)$이고 각 **완전 조건부분포** $\pi(\theta_j \mid \theta_{-j})$에서 뽑을 수 있다고 하자. 한 바퀴에서 $j = 1, \dots, d$ 차례로 $\theta_j \sim \pi(\theta_j \mid \theta_{-j})$로 바꾼다($\theta_{-j}$는 가장 최근 값).

정리 4.7 (Gibbs는 받아들임 확률 1의 MH) 좌표 j를 완전 조건부분포로 바꾸는 단계는 제안 $q(\theta, \theta') = \pi(\theta'_j \mid \theta_{-j})\mathbb{1}(\theta'_{-j} = \theta_{-j})$의 MH이고 받아들임 확률은 1이다. 따라서 각 단계와 그 합성(한 바퀴)은 π를 정상분포로 갖는다.

증명. **전략** 결합밀도를 $\pi(\theta) = \pi(\theta_{-j})\pi(\theta_j \mid \theta_{-j})$로 쪼개 MH 비에 넣으면 모든 인수가 약분된다. 합성은 "π를 보존하는 핵을 차례로 쓰면 π가 보존된다"로 처리한다. 단계별 증명은 연습 4.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 47쪽에 있다.

한 바퀴의 합성은 일반적으로 가역이 아니다. 그래도 정상분포는 지켜진다. 켤레 구조가 있는 계층모형에서는 완전 조건부분포가 정리 1.2의 꼴이 되어 Gibbs 표집이 자연스럽다(6절).

4.3 수렴과 진단

증명 없이 쓰는 사실 (마르코프 연쇄의 에르고딕 정리) 연쇄가 π를 정상분포로 갖고, π-기약(양의 확률의 어느 집합에나 언젠가 들어간다)이고 비주기적이면, 거의 모든 출발점에서 $\frac{1}{N}\sum_{t=1}^{N} g(X_t) \to E_\pi g$ (확률 1)이다. 추가 조건 아래 중심극한정리도 성립하며

$$\operatorname{Var}\Big(\frac{1}{N}\sum_{t=1}^{N} g(X_t)\Big) \approx \frac{\sigma_g^2}{N}\Big(1 + 2\sum_{k\ge 1}\rho_k\Big),$$

ρ_k는 정상 상태에서 $g(X_t)$의 시차 k 자기상관이다. 7회 정리 6.1의 σ^2/N에서 N을 $N_{\text{eff}} = N/(1 + 2\sum\rho_k)$로 바꾼 것이며, N_{eff}를 **유효 표본 크기**(ESS)라 한다. 3회 예제 6.6의 AR(1)에서 본 $n_{\text{eff}} \approx n(1-\rho)/(1+\rho)$가 그 특수한 경우다.

명제 4.8 (상관된 정규에서의 Gibbs) (X, Y)가 평균 0, 분산 1, 상관 ρ의 이변량 정규이면 Gibbs 표집 $X_t \sim \mathcal{N}(\rho Y_{t-1}, 1-\rho^2)$, $Y_t \sim \mathcal{N}(\rho X_t, 1-\rho^2)$의 X 열은 계수 ρ^2의 AR(1)

$$X_t = \rho^2 X_{t-1} + \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1-\rho^4)$$

이다. 따라서 $\rho_k = \rho^{2k}$이고 $N_{\text{eff}}/N = (1-\rho^2)/(1+\rho^2)$이다.

증명. **전략** 두 조건부 정규 걸음을 잡음 항으로 쓰고 하나를 다른 하나에 대입한다. 그러면 X 열만의 재귀식이 나오고, AR(1)의 자기상관을 더한다(3회 예제 6.6). 단계별 증명은 연습 4.7(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 48쪽에 있다.

$\rho = 0.5,\ 0.9,\ 0.99$이면 $N_{\text{eff}}/N = 0.60,\ 0.105,\ 0.010$이다. 좌표축 방향으로만 움직이는 Gibbs는 대각선으로 길쭉한 분포에서 계단을 잘게 밟는다(그림 4.1). 모수의 상관을 줄이는 재모수화(중심화)가 MCMC의 효율을 좌우하는 이유다.

연쇄가 정상분포에 도달했는지는 한 연쇄만 보아서는 알 수 없다. Gelman과 Rubin(1992)은 흩어진 출발점에서 여러 연쇄를 돌려 연쇄 사이의 분산과 연쇄 안의 분산을 비교했다. 14회 분산분석의 논리 그대로다.

정의 4.9 ($\hat{R}$ (잠재적 축소 인자)) 길이 n의 연쇄 m개에서 연쇄 j의 평균을 $\bar{\theta}_j$, 분산을 s_j^2이라 하고 $W = \frac{1}{m}\sum s_j^2$, $B = \frac{n}{m-1}\sum(\bar{\theta}_j - \bar{\theta})^2$이라 하자.

$$\hat{V} = \frac{n-1}{n}W + \frac{1}{n}B, \qquad \hat{R} = \sqrt{\hat{V}/W}.$$

모든 연쇄가 같은 분포에서 뽑히고 있으면 $\hat{V}$와 W는 모두 목표분포의 분산을 추정하므로 $\hat{R} \approx 1$이다. 출발점의 영향이 남아 연쇄들이 서로 다른 곳을 돌고 있으면 B가 커서 $\hat{R} > 1$이다. 실무에서는 $\hat{R} < 1.01$을 기준으로 쓴다. $\hat{R} \approx 1$은 필요조건일 뿐이다. 모든 연쇄가 같은 봉우리에 갇혀 있으면 $\hat{R}$도 속는다(7절).

연습문제

연습 4.1 (두 상태 연쇄 ★) 전이행렬 $\left(\begin{smallmatrix} 1-a & a \\ b & 1-b \end{smallmatrix}\right)$$(0 < a, b < 1)$의 정상분포를 구하고 상세 균형을 확인하라. P^n이 정상분포로 가는 속도는 $|1-a-b|^n$임을 보여라. 풀이 47쪽

연습 4.2 (가역이 아닌 정상분포 ★) 상태 $\{1,2,3\}$에서 $1 \to 2 \to 3 \to 1$을 확률 0.9로, 반대 방향을 0.1로 움직이는 연쇄의 정상분포가 균등임을 보이고, 상세 균형을 만족하지 않음을 보여라. 풀이 47쪽

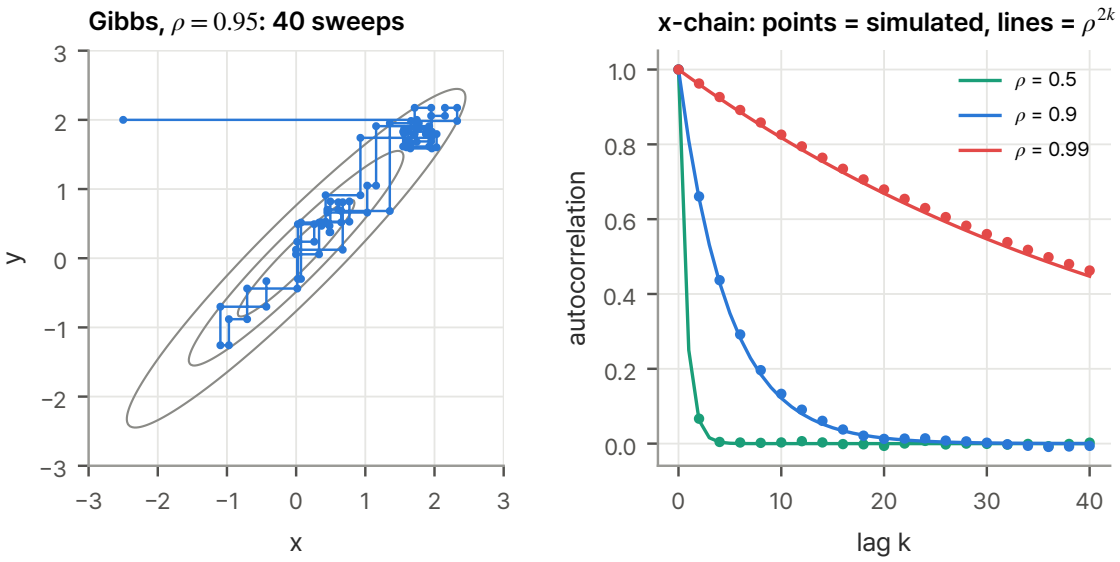


그림 4.1 왼쪽: $\rho = 0.95$인 이변량 정규에서 Gibbs 표집의 처음 40걸음. 좌표축 방향으로만 움직인다. 오른쪽: X열의 자기상관(점)과 명제 4.8의 ρ^{2k}(선). ρ가 1에 가까울수록 연쇄가 천천히 잊는다.

연습 4.3 (독립 제안 ★★) $q(x, y) = g(y)$(현재 상태와 무관)이면 MH 받아들임 확률이 $\min\{1, w(y)/w(x)\}$, $w = \pi/g$임을 보여라. 7회의 중요도 표집의 가중치와 비교하라. g의 꼬리가 π보다 가벼우면 어떤 일이 생기는가? 풀이 47쪽

연습 4.4 ($\hat{R}$의 하한 ★) 연쇄 평균이 모두 같으면 $\hat{R} = \sqrt{(n-1)/n} < 1$임을 보여라. $\hat{R}$이 1보다 큰 것은 무엇 때문인가? 풀이 47쪽

연습 4.5 (Gibbs의 무작위 순서 ★★) 매 단계 좌표 j를 균등하게 무작위로 골라 완전 조건부분포로 바꾸는 "무작위 순서" Gibbs의 전이핵은 π에 대해 상세 균형을 만족함을 보여라. (체계적 순서의 한 바퀴는 일반적으로 가역이 아니다.) 풀이 47쪽

연습 4.6 (Gibbs는 MH: 증명 ★ 유도형) 정리 4.7를 다음 순서로 증명하라.

(a) $\theta'_{-j} = \theta_{-j}$이면 $\pi(\theta) = \pi(\theta_{-j})\pi(\theta_j \mid \theta_{-j})$, $\pi(\theta') = \pi(\theta_{-j})\pi(\theta'_j \mid \theta_{-j})$임을 조건부밀도의 정의로 보여라.

(b) 제안 $q(\theta, \theta') = \pi(\theta'_j \mid \theta_{-j})$와 $q(\theta', \theta) = \pi(\theta_j \mid \theta_{-j})$를 MH 비에 넣어 비가 1임을 보여라.

(c) 정리 4.5로 각 좌표 단계가 π를 보존함을 결론짓고, 핵 K_1, K_2가 π를 보존하면 합성 K_1K_2도 π를 보존함을 적분의 순서를 바꿔 보여라.

풀이 47쪽

연습 4.7 (상관된 정규의 Gibbs: 증명 ★★ 유도형) 명제 4.8를 다음 순서로 증명하라.

(a) 5회의 조건부 정규에서 $Y_{t-1} = \rho X_{t-1} + \sqrt{1-\rho^2}\,\epsilon_{t-1}$, $X_t = \rho Y_{t-1} + \sqrt{1-\rho^2}\,\epsilon'_t$로 쓸 수 있음을 설명하라. 잡음 $\epsilon_{t-1}, \epsilon'_t$는 서로, 그리고 과거와 독립인 $\mathcal{N}(0,1)$이다.

(b) 대입해 $X_t = \rho^2 X_{t-1} + \eta_t$를 얻고 $\operatorname{Var}\eta_t = 1-\rho^4$을 보여라.

(c) AR(1)의 자기상관이 ρ^{2k}임을 쓰고 $1 + 2\sum_{k\ge 1}\rho^{2k} = \frac{1+\rho^2}{1-\rho^2}$을 기하급수로 계산하라.

풀이 48쪽

5

Stein의 역설

> **역사** 1955년 Berkeley의 제3회 수리통계 심포지엄에서 Charles Stein은 정규 평균 p개를 동시에 추정할 때 $p \geq 3$이면 표본평균이 허용적이지 않다는 것을 보였다. 모든 곳에서 표본평균보다 위험이 작은 추정량이 있다는 것이다. 표본평균은 불편이고, 최대가능도이고, 최소분산이고, minimax였다. 아무도 믿으려 하지 않았다. 6년 뒤 Stein은 대학원생 Willard James와 함께 그런 추정량을 명시적으로 적었다(1961). 같은 1955년 심포지엄에서 Herbert Robbins는 "경험적 Bayes"를 발표했다. 많은 비슷한 문제를 함께 풀면 사전분포를 자료에서 추정할 수 있다는 생각이었다. 두 생각이 같은 것임을 보인 것은 1970년대 Bradley Efron과 Carl Morris였다. Stigler(1990)는 여기에 한 줄을 더했다. James–Stein 추정량은 Galton의 회귀다.

이 절에서 $\mathbf{X} \sim \mathcal{N}_p(\theta, I_p)$이고 손실은 $\|\delta - \theta\|^2 = \sum_i (\delta_i - \theta_i)^2$이다. $\mathbf{X}$ 자신의 위험은 $E\|\mathbf{X} - \theta\|^2 = p$로 θ에 무관하다.

5.1 Stein의 보조정리와 James–Stein 정리

보조정리 5.1 (Stein의 보조정리) $Z \sim \mathcal{N}(\theta, 1)$이고 g가 절대연속이며 $E|g'(Z)| < \infty$이면 $E[(Z-\theta)g(Z)] = E[g'(Z)]$이다. 다차원에서 $\mathbf{X} \sim \mathcal{N}_p(\theta, I)$이고 $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^p$가 적당히 매끄러우면 $E[(\mathbf{X}-\theta)'\mathbf{g}(\mathbf{X})] = E[\nabla \cdot \mathbf{g}(\mathbf{X})]$이다.

증명. **전략** 부분적분을 Fubini로 한다. 정규밀도를 $\varphi(z) = \int_z^\infty u\varphi(u)\,du$로 쓰면 $E[g'(Z)]$는 이중적분이 되고, 적분 순서를 바꾸면 g'의 적분이 g로 돌아온다.

① $Z - \theta \sim \mathcal{N}(0,1)$이고 g 대신 $g(\cdot + \theta)$를 쓰면 되므로 $\theta = 0$으로 둔다. — 평행이동

② 표준정규밀도는 $\varphi'(u) = -u\varphi(u)$이므로 $z > 0$에서 $\varphi(z) = \int_z^\infty u\varphi(u)\,du$, $z < 0$에서 $\varphi(z) = -\int_{-\infty}^z u\varphi(u)\,du$이다. — 미적분의 기본정리, $\varphi(\pm\infty) = 0$

③ 양의 쪽에서 적분 순서를 바꾼다.

$$\int_0^\infty g'(z)\varphi(z)\,dz = \int_0^\infty g'(z)\int_z^\infty u\varphi(u)\,du\,dz = \int_0^\infty u\varphi(u)\int_0^u g'(z)\,dz\,du. \quad \text{Fubini [T6]}$$

④ g가 절대연속이므로 $\int_0^u g'(z)\,dz = g(u) - g(0)$이고, 양의 쪽은 $\int_0^\infty u\varphi(u)\big(g(u)-g(0)\big)du$이다. — 절대연속의 정의

⑤ 음의 쪽도 같은 계산으로 $\int_{-\infty}^0 g'\varphi = \int_{-\infty}^0 u\varphi(u)\big(g(u)-g(0)\big)du$이다. — ②의 둘째 식

⑥ 더하면 $E[g'(Z)] = E[Z(g(Z) - g(0))] = E[Zg(Z)]$이다. $EZ = 0$

⑦ 다차원에서는 좌표 i마다 나머지 좌표를 고정하고 ⑥을 X_i에 쓴 뒤, 나머지 좌표에 대해 기댓값을 취해 i에 대해 더한다. 독립, 3회 정리 2.5

조건 확인 ③에서 적분 순서를 바꾸려면 피적분함수의 절댓값을 이중으로 적분한 값이 유한해야 한다. 그 값은 $\int_0^\infty |g'(z)| \int_z^\infty u\varphi(u)\,du\,dz = \int_0^\infty |g'(z)|\varphi(z)\,dz$이므로 가정 $E|g'(Z)| < \infty$가 바로 그 조건이다 [T6]. ④의 절대연속성이 깨지면 등식이 틀린다(연습 5.2(b)의 sign). ⑥에서 $E|Zg(Z)| < \infty$는 ③–⑤의 적분이 유한하다는 것에서 나온다. ∎

> **정리 5.2** (James–Stein) $p \ge 3$이면 $\delta_{\mathrm{JS}}(\mathbf{X}) = \left(1 - \dfrac{p-2}{\|\mathbf{X}\|^2}\right)\mathbf{X}$의 위험은
>
> $$R(\theta, \delta_{\mathrm{JS}}) = p - (p-2)^2\, E_\theta\!\left[\frac{1}{\|\mathbf{X}\|^2}\right] < p \qquad \text{모든 } \theta.$$

증명. **전략** $\delta_{\mathrm{JS}} = \mathbf{X} - \mathbf{g}(\mathbf{X})$로 쓰고 손실을 전개한다. 모르는 θ가 들어 있는 교차항을 Stein의 보조정리로 θ가 없는 기댓값으로 바꾼다.

① $\mathbf{g}(\mathbf{x}) = (p-2)\mathbf{x}/\|\mathbf{x}\|^2$로 두면 $\delta_{\mathrm{JS}} = \mathbf{X} - \mathbf{g}(\mathbf{X})$이다. 정의

② 손실을 전개한다.

$$\|\delta_{\mathrm{JS}} - \theta\|^2 = \|\mathbf{X} - \theta\|^2 - 2(\mathbf{X} - \theta)'\mathbf{g}(\mathbf{X}) + \|\mathbf{g}(\mathbf{X})\|^2. \qquad \|u - v\|^2 \text{의 전개}$$

③ $\partial_i\big(x_i/\|\mathbf{x}\|^2\big) = (\|\mathbf{x}\|^2 - 2x_i^2)/\|\mathbf{x}\|^4$을 i에 대해 더하면 $\nabla\cdot\mathbf{g}(\mathbf{x}) = (p-2)^2/\|\mathbf{x}\|^2$이다. 몫의 미분

④ 교차항은 $E[(\mathbf{X} - \theta)'\mathbf{g}(\mathbf{X})] = E[\nabla\cdot\mathbf{g}(\mathbf{X})] = (p-2)^2 E[1/\|\mathbf{X}\|^2]$이다. 보조정리 5.1, ③

⑤ 마지막 항은 $\|\mathbf{g}(\mathbf{X})\|^2 = (p-2)^2/\|\mathbf{X}\|^2$이다. ①

⑥ 기댓값을 더하면

$$R(\theta, \delta_{\mathrm{JS}}) = p - 2(p-2)^2 E\!\left[\frac{1}{\|\mathbf{X}\|^2}\right] + (p-2)^2 E\!\left[\frac{1}{\|\mathbf{X}\|^2}\right] = p - (p-2)^2 E\!\left[\frac{1}{\|\mathbf{X}\|^2}\right]. \quad E\|\mathbf{X} - \theta\|^2 = p$$

⑦ $p \ge 3$이면 $0 < E[1/\|\mathbf{X}\|^2] < \infty$이므로 위험은 p보다 작다. 연습 5.1, 5.3

조건 확인 보조정리 5.1를 쓰려면 $\mathbf{g}$의 각 좌표가 거의 모든 좌표 직선 위에서 절대연속이고 $E|\partial_i g_i(\mathbf{X})| < \infty$여야 한다. $\mathbf{g}$는 원점에서만 특이하고 $|\partial_i g_i(\mathbf{x})| \le 3(p-2)/\|\mathbf{x}\|^2$이다. $\|\mathbf{x}\|^{-2}$을 원점 근처의 p차원 공에서 적분하면 극좌표로 $\int_0^1 r^{-2} r^{p-1} dr = \int_0^1 r^{p-3} dr < \infty$이므로 $p \ge 3$에서 적분가능하다 [T7]. ⑦의 상계는 $\|\mathbf{X}\|^2$이 $K \sim \mathrm{Poi}(\|\theta\|^2/2)$로 섞인 χ^2_{p+2K}라는 것에서 $E[1/\|\mathbf{X}\|^2] = E[1/(p-2+2K)] \le 1/(p-2)$로 나온다. ∎

따름정리 5.3 (Stein (1956)) $p \geq 3$이면 제곱손실에서 $\mathbf{X}$는 허용적이지 않다.

$\theta = \mathbf{0}$이면 $\|\mathbf{X}\|^2 \sim \chi^2_p$이고 $E[1/\chi^2_p] = 1/(p-2)$(연습 5.1)이므로 위험은 $p - (p-2) = 2$다. 차원이 1000이어도 위험이 2다. $\|\theta\|$가 커지면 위험은 p로 다시 올라가지만 넘지 않는다. $\|\mathbf{X}\|^2$은 비중심 카이제곱이고 $E[1/\|\mathbf{X}\|^2]$은 Poisson 가중합으로 정확히 계산된다(연습 5.3). $p = 10$에서 $\|\theta\|^2 = 0,\ 10,\ 100$이면 위험은 2, 6.21, 9.40이다(그림 5.1). $p = 1, 2$에서는 $\mathbf{X}$가 허용적이다(Blyth의 방법, 증명은 생략).

어디로 줄이는지는 중요하지 않다. 임의의 점 $\boldsymbol{\nu}$를 향해 $\boldsymbol{\nu} + \left(1 - \frac{p-2}{\|\mathbf{X}-\boldsymbol{\nu}\|^2}\right)(\mathbf{X} - \boldsymbol{\nu})$로 줄여도 같은 정리가 성립한다. 줄일 곳을 자료의 평균 $\bar{X}\mathbf{1}$로 고르면 자유도 하나를 써서 $p-2$가 $p-3$이 되고,

$$\delta_{\mathrm{EM}}(\mathbf{X}) = \bar{X}\mathbf{1} + \left(1 - \frac{p-3}{\sum_i (X_i - \bar{X})^2}\right)(\mathbf{X} - \bar{X}\mathbf{1})$$

는 $p \geq 4$에서 $\mathbf{X}$를 지배한다. 6절의 Efron–Morris 추정량이다. 축소 인자가 음수가 되는 것을 막은 $\delta^+_{\mathrm{JS}} = \max\!\left(0, 1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)\mathbf{X}$는 δ_{JS}를 다시 지배한다(연습 5.4).

5.2 경험적 Bayes — 왜 줄이는가

정리 5.4 (James–Stein의 경험적 Bayes 해석) $\theta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$, $X_i \mid \theta \sim \mathcal{N}(\theta_i, 1)$ (독립)이라 하자.

(a) Bayes 추정량은 $E(\theta_i \mid \mathbf{X}) = (1 - B)X_i$, $B = \dfrac{1}{1+\tau^2}$이고 Bayes 위험은 $p(1-B)$이다.

(b) 주변분포에서 $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1+\tau^2)$이고 $E\left[\dfrac{p-2}{\|\mathbf{X}\|^2}\right] = B$이다. 곧 δ_{JS}는 모르는 B를 불편추정량으로 바꿔 넣은 Bayes 추정량이다.

(c) δ_{JS}의 Bayes 위험은 $p(1-B) + 2B$이다. 사전분포를 모르는 값은 τ와 p에 상관없이 $2B \leq 2$다.

증명. **전략** (a)는 켤레 사전분포(정리 1.2(c)), (b)는 독립 정규의 합과 역카이제곱의 평균, (c)는 정리 5.2의 위험을 사전분포로 적분하는 것이다. 단계별 증명은 연습 5.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 49쪽에 있다.

Bayes 추정량은 τ^2을 알아야 한다. 그러나 p개의 X_i가 모두 같은 τ^2의 정보를 담고 있다. 하나하나는 자기 θ_i를 알려 주지만, 함께 모이면 θ_i들이 얼마나 흩어져 있는지를 알려 준다. 이것이 Robbins의 경험적 Bayes이고, 3회의 전분산 공식 $\mathrm{Var}X_i = E\mathrm{Var}(X_i \mid \theta_i) + \mathrm{Var}\,\theta_i = 1 + \tau^2$에서 τ^2을 거꾸로 읽는 것이다.

Galton의 회귀로 읽기 (Stigler 1990) (θ_i, X_i)의 산점도를 생각하자. X를 θ에 회귀하면 기울기는 1이다(X는 불편). 우리가 원하는 것은 거꾸로 θ를 X에 회귀한 직선이고, 그 기울기는 $\mathrm{Cov}(\theta,X)/\mathrm{Var}X = \tau^2/(1+\tau^2) = 1 - B < 1$이다. Galton이 키에서 본 평균으로의 회귀(5회 따름정리 2.2)다. θ_i를 관측하지 못하므로 이 회귀를 직접 맞출 수는 없지만, 기울기는 $\mathrm{Var}X$만으로 정해지고 그것은 X_i들의 흩어짐으로 추정된다. 표본평균은 "잘못된" 회귀선을 쓴 추정량이다(그림 5.1 오른쪽).

연습문제

연습 5.1 (역카이제곱의 평균 ★) $V \sim \chi^2_p$이면 $E[1/V] = 1/(p-2)$ $(p \geq 3)$임을 감마함수로 보여라. $p \leq 2$이면 왜 무한한가? 풀이 48쪽

연습 5.2 (Stein 보조정리의 쓰임 ★★) (a) $g(z) = z$와 $g(z) = z^2$에 보조정리 5.1를 적용해 $\mathcal{N}(\theta,1)$의 둘째·셋째 적률을 구하라. (b) $g(z) = \mathrm{sign}(z)$이면 $\theta = 0$에서 등식이 성립하지 않음을 보여라. 어느 조건이 깨졌는가? James–Stein의 $\mathbf{g}$는 원점에서 미분가능하지 않은데 왜 괜찮은가? 풀이 48쪽

연습 5.3 (비중심 카이제곱 ★★) $\|\mathbf{X}\|^2 (\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\theta}, I))$의 분포는 $K \sim \mathrm{Poi}(\|\boldsymbol{\theta}\|^2/2)$로 섞은 χ^2_{p+2K}임을 받아들이고, $E[1/\|\mathbf{X}\|^2] = E\big[1/(p-2+2K)\big]$을 보여라. Jensen 부등식으로 $R(\boldsymbol{\theta}, \delta_{\mathrm{JS}}) \leq p - \frac{(p-2)^2}{p+\|\boldsymbol{\theta}\|^2}$임을 보이고, 이것과 선형 축소의 하한 $\frac{p\|\boldsymbol{\theta}\|^2}{p+\|\boldsymbol{\theta}\|^2}$의 차이가 4를 넘지 않음을 보여라. 풀이 48쪽

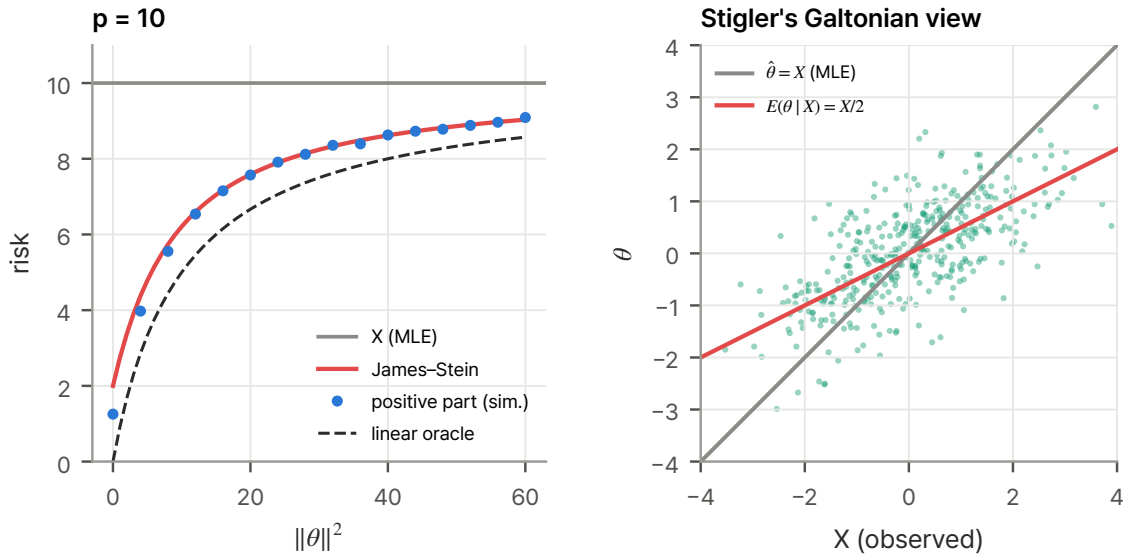


그림 5.1 왼쪽: $p = 10$에서 $\|\boldsymbol{\theta}\|^2$에 따른 위험. $\mathbf{X}$(회색, 10), James–Stein(정확한 값), 양의 부분 James–Stein(모의실험), 선형 축소의 이론적 하한 $p\|\boldsymbol{\theta}\|^2/(p+\|\boldsymbol{\theta}\|^2)$(점선). 오른쪽: $\theta_i \sim \mathcal{N}(0,1)$, $X_i \sim \mathcal{N}(\theta_i, 1)$의 산점도. X를 θ에 회귀한 직선(기울기 1)과 θ를 X에 회귀한 직선(기울기 $\frac{1}{2}$). James–Stein은 뒤의 직선을 자료로 추정한다.

연습 5.4 (양의 부분 ★★) δ_{JS}^{+}가 δ_{JS}를 지배함을 보여라. (힌트: $\|\mathbf{X}\|^2 < p-2$인 사건에서 $\delta_{\mathrm{JS}} = -c\mathbf{X}$ $(c>0)$이고 $\delta_{\mathrm{JS}}^{+} = \mathbf{0}$이다. 이 사건에서 $\|-c\mathbf{X}-\theta\|^2 - \|\theta\|^2 = c^2\|\mathbf{X}\|^2 + 2c\mathbf{X}'\theta$이고, $\mathbf{X}'\theta$의 조건부 기댓값의 부호를 따진다.) 풀이 49쪽

연습 5.5 (분산을 모를 때 ★★) $X_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$이고 σ^2을 모르지만 σ^2의 추정량 S^2이 있어 $\nu S^2/\sigma^2 \sim \chi^2_\nu$이고 $\mathbf{X}$와 독립이라 하자. $\left(1 - \frac{(p-2)\nu}{\nu+2}\cdot\frac{S^2}{\|\mathbf{X}\|^2}\right)\mathbf{X}$의 위험이 $p\sigma^2$보다 작음을 보여라. (힌트: $E[S^2]$, $E[S^4]$을 쓴다.) 풀이 49쪽

연습 5.6 (경험적 Bayes 해석: 증명 ★★ 유도형) 정리 5.4를 다음 순서로 증명하라.

(a) 정리 1.2(c)에 $n=1$, $\sigma^2=1$, 사전평균 0을 넣어 θ_i의 사후평균과 사후분산을 구하고, 정리 2.1(a)로 Bayes 위험이 $p(1-B)$임을 보여라. 좌표들이 독립이라는 것은 어디에 쓰이는가?

(b) $X_i = \theta_i + \epsilon_i$의 주변분포가 $\mathcal{N}(0, 1+\tau^2)$임을 보이고, $\|\mathbf{X}\|^2 = (1+\tau^2)V$, $V \sim \chi^2_p$에서 $E\left[\frac{p-2}{\|\mathbf{X}\|^2}\right] = B$를 얻어라(연습 5.1).

(c) 정리 5.2의 위험을 θ의 사전분포로 적분하면 $E[1/\|\mathbf{X}\|^2]$이 주변분포의 기댓값이 됨을 설명하고(3회 정리 2.5), Bayes 위험 $p(1-B)+2B$를 얻어라.

풀이 49쪽

6

Efron–Morris의 타율

> **역사** Stein의 정리는 20년 가까이 수학적 호기심으로 남았다. 실제 자료에서 쓰면 어떻게 되는가? Efron과 Morris(1975)는 야구를 골랐다. 1970년 시즌에 타석에 45번 선 메이저리그 타자 18명의 타율로 그 시즌 나머지의 타율을 예측했다. 나머지 시즌의 수백 타석이 "참값" 노릇을 하므로 추정량을 채점할 수 있었다. 1977년 그들은 *Scientific American*에 "통계학의 Stein 역설"을 써서 이 결과를 일반 독자에게 알렸다. 표 속 이름들은 야구팬에게 낯익다. 그 해 Roberto Clemente는 첫 45타석에서 18안타를 쳤다.

6.1 자료와 James–Stein

선수 i의 첫 45타석 타율을 y_i, 그 시즌 나머지의 타율을 p_i라 하자. $45y_i \sim \mathrm{b}(45, \theta_i)$이고 정규근사로 $y_i \approx \mathcal{N}(\theta_i, \sigma^2)$, $\sigma^2 = \bar{y}(1-\bar{y})/45$로 둔다. 18명 모두에 같은 σ^2을 쓰는 단순화다.

예제 6.1 (18명의 타율) $\bar{y} = 0.2654$, $\sigma^2 = 0.004332$ ($\sigma = 0.066$), $\sum (y_i - \bar{y})^2 = 0.0825$이다. 5절의 평균을 향한 James–Stein($p = 18$)의 축소 인자는

$$c = 1 - \frac{(18-3)\sigma^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{15 \times 0.004332}{0.0825} = 0.212$$

이다. 각 선수의 편차를 평균에서 21%만 남기고 79%를 줄인다. 결과는 표와 같다.

선수	y_i	JS	p_i	선수	y_i	JS	p_i
Clemente	.400	.294	.346	Swoboda	.244	.261	.230
F. Robinson	.378	.289	.298	Unser	.222	.256	.264
Howard	.356	.285	.276	Williams	.222	.256	.256
Johnstone	.333	.280	.222	Scott	.222	.256	.303
Berry	.311	.275	.273	Petrocelli	.222	.256	.264
Spencer	.311	.275	.270	Rodriguez	.222	.256	.226
Kessinger	.289	.270	.263	Campaneris	.200	.252	.285
Alvarado	.267	.266	.210	Munson	.178	.247	.316
Santo	.244	.261	.269	Alvis	.156	.242	.200

제곱오차의 합 $\sum (\hat{\theta}_i - p_i)^2$은 y_i가 0.0754, James–Stein이 0.0213으로 3.5배 작다. 18명 중 16명에서 James–Stein이 더 가깝다. 나빠진 둘은 Swoboda와 Rodriguez로, 평균보다 조금

> 낮게 출발해 실제로도 낮게 끝난 선수들이다. Clemente의 .400은 .294로 깎였고 실제는 .346이었다. Munson은 .178에서 .247로 올려졌고 실제는 .316이었다.

이항의 분산이 θ_i에 달려 있다는 것이 마음에 걸리면 7회의 분산안정화 변환 $x_i = \sqrt{45}\arcsin(2y_i - 1)$을 쓸 수 있다. 그러면 $x_i \approx \mathcal{N}(\cdot, 1)$이고, Efron과 Morris가 실제로 쓴 것이 이것이다. 이 척도에서 축소 인자는 0.209, 제곱오차의 비는 3.52로 거의 같다.

6.2 계층 Bayes와 Gibbs 표집

경험적 Bayes는 τ^2을 추정한 뒤 그것이 참값인 양 썼다. 완전한 Bayes는 μ와 τ에도 사전분포를 주어 그 불확실성까지 사후분포에 넣는다.

$$y_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma^2), \qquad \theta_i \mid \mu, \tau \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2), \qquad \pi(\mu, \tau) \propto 1 \ \ (\tau > 0).$$

명제 6.2 (계층 정규모형의 완전 조건부분포) k명의 선수에 대해 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$라 하면

$$\begin{aligned} \theta_i \mid \mu, \tau, \mathbf{y} &\sim \mathcal{N}\big(v(y_i/\sigma^2 + \mu/\tau^2),\ v\big), \quad v = (1/\sigma^2 + 1/\tau^2)^{-1} \quad \text{(서로 독립)}, \\ \mu \mid \boldsymbol{\theta}, \tau, \mathbf{y} &\sim \mathcal{N}(\bar\theta,\ \tau^2/k), \\ \tau^2 \mid \boldsymbol{\theta}, \mu, \mathbf{y} &\sim \text{역감마}\Big(\frac{k-1}{2},\ \frac{1}{2}\sum_i (\theta_i - \mu)^2\Big). \end{aligned}$$

또 $\boldsymbol{\theta}$를 적분하면 $y_i \mid \mu, \tau \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2 + \tau^2)$이다.

증명. 결합밀도는 $\prod_i \varphi_\sigma(y_i - \theta_i)\varphi_\tau(\theta_i - \mu)$이다. θ_i에 대해서는 정리 1.2(c)($n = 1$, 사전평균 μ)이다. μ에 대해서는 $\prod \varphi_\tau(\theta_i - \mu) \propto \exp\{-\frac{k}{2\tau^2}(\mu - \bar\theta)^2\}$. τ에 대해서는 $\tau^{-k}\exp\{-\text{SS}/(2\tau^2)\}$, $\text{SS} = \sum(\theta_i - \mu)^2$이고, τ의 평평한 사전분포를 τ^2으로 바꾸면 야코비안 $\frac{d\tau}{d\tau^2} \propto (\tau^2)^{-1/2}$이 붙어 $(\tau^2)^{-(k+1)/2}e^{-\text{SS}/(2\tau^2)}$, 곧 모양 $(k-1)/2$의 역감마다. 마지막은 독립 정규의 합(5회). ∎

명제의 세 줄을 차례로 뽑으면 Gibbs 표집이다(알고리즘 4.6). 흩어진 출발점에서 연쇄 4개를 각 6000번 돌리고 처음 1000번을 버렸다. $\hat R$은 μ와 τ 모두 1.000이다. μ의 사후평균은 0.2656, τ의 사후중앙값은 0.0298이다. 선수들의 실제 능력은 표준편차 3푼 정도로만 흩어져 있고, 45타석의 표본 표준편차 $\sigma = 0.066$이 그 두 배가 넘는다. 관측된 흩어짐의 대부분은 운이다. 그래서 강하게 줄인다.

θ_i의 사후평균을 추정값으로 쓰면 제곱오차의 합은 0.0213으로 James–Stein과 같고, 두 추정값의 차이는 가장 큰 것도 0.0006이다. 같은 사후분포를 다른 알고리즘으로 확인할 수 있다. 명제의 마지막 줄의 주변 가능도로 $(\mu, \log\tau)$에 대한 무작위 걸음 Metropolis를 돌리면(야코비안 때문에 로그 사후밀도에 $\log\tau$가 더해진다) 받아들임 비율 0.69, $\hat R$ 1.000과 1.003, μ의 사후평균 0.2659, τ의 사후중앙값 0.0286이다. 두 알고리즘은 몬테카를로 오차 안에서 같은 답을 낸다(그림 6.1).

세 추정량의 차이 James–Stein은 공식 하나, 경험적 Bayes는 τ^2의 점추정, 계층 Bayes는 τ의 사후분포 전체를 쓴다. 18명처럼 k가 크면 셋은 거의 같다. k가 작거나 $\hat\tau^2$이 0 근처이면 차이가 난다. 경험적 Bayes는 $\hat\tau=0$으로 모두 평균으로 합쳐 버리지만, 계층 Bayes는 τ가 0이 아닐 가능성을 남긴다. 이 차이는 사후 신용구간의 폭에서 가장 크게 나타난다.

연습문제

연습 6.1 (Clemente의 손 계산 ★) 예제 6.1의 $\bar y$, σ^2, $\sum(y_i-\bar y)^2$으로 c를 구하고 Clemente와 Munson의 James–Stein 추정값을 계산하라. 풀이 50쪽

연습 6.2 (주변 가능도 ★) 명제 6.2의 마지막 줄을 증명하라. 이것을 써서 τ^2의 적률추정량 $\hat\tau^2=\max\!\left(0,\ \frac{1}{k-1}\sum(y_i-\bar y)^2-\sigma^2\right)$을 구하고 타율 자료에서 계산하라. 풀이 50쪽

연습 6.3 (왜 $p-3$인가 ★★) $\mathbf{X}\sim\mathcal{N}_p(\theta, I)$에서 평균을 향한 축소 $\bar X\mathbf{1}+\left(1-\frac{p-3}{\|\mathbf{X}-\bar X\mathbf{1}\|^2}\right)(\mathbf{X}-\bar X\mathbf{1})$의 위험을 구하라. (힌트: 직교변환으로 $\mathbf{X}-\bar X\mathbf{1}$을 $p-1$차원의 정규벡터로 보고 정리 5.2을 $p-1$차원에 쓴다. $\bar X\mathbf{1}$ 방향의 위험은 1이다.) 풀이 50쪽

연습 6.4 ($1/\tau$ 사전분포의 실패 ★★) 계층 모형에서 $\pi(\mu,\tau)\propto 1/\tau$(척도 모수의 Jeffreys)를 쓰면 사후분포가 비정칙임을 보여라. (힌트: $\tau\to 0$일 때 주변 가능도 $\prod_i \varphi_{\sqrt{\sigma^2+\tau^2}}(y_i-\mu)$는 양의 극한을 갖는다. $\int_0^\epsilon d\tau/\tau$는?) 풀이 50쪽

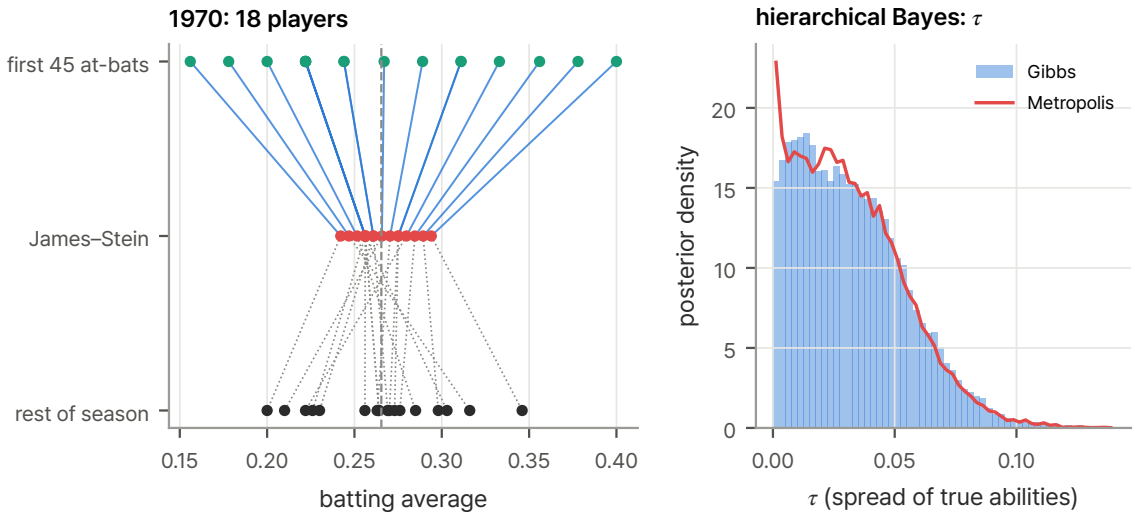


그림 6.1 왼쪽: 18명의 첫 45타석 타율(위), James–Stein 추정값(가운데), 시즌 나머지 타율(아래). 화살표는 평균 0.265를 향해 줄어든다. 오른쪽: 계층 모형의 τ의 사후분포. Gibbs 표집(막대)과 주변 가능도의 Metropolis(선)가 겹친다.

7

베이즈가 실패할 때

역사 Gibbs 표집이 퍼지자 새로운 종류의 오류가 나타났다. Hobert와 Casella(1996)는 계층 선형모형에서 널리 쓰이던 비정칙 사전분포가 비정칙 사후분포를 낳는데도, 모든 완전 조건부분포는 정칙이어서 Gibbs 표집이 아무 경고 없이 돌아간다는 것을 보였다. 출력은 그럴듯한 숫자였지만 그 숫자가 근사하는 분포는 존재하지 않았다. Gelman(2006)은 분산 모수의 사전분포가 집단이 적을 때 결론을 좌우한다는 것을 보였다. Stein의 축소도 오해를 낳았다. Efron과 Morris는 처음부터 James–Stein이 **합계**의 위험을 줄일 뿐 개별 선수를 보호하지는 않는다고 경고했다.

가정 주의 주의 1 — 사후분포가 존재하는가비정칙 사전분포를 쓰면 사후분포가 정칙인지 따로 증명해야 한다. 연습 2.4의 Haldane 사전분포는 성공이 하나도 없으면 비정칙 사후분포를 준다. 계층 모형에서 $\pi(\tau^2) \propto 1/\tau^2$을 쓰면 τ^2의 완전 조건부분포는 역감마$\left(\frac{k}{2}, \frac{1}{2}\mathrm{SS}\right)$로 정칙이지만 결합 사후분포는 비정칙이다(연습 6.4과 같은 논리). 타율 자료에서 이 Gibbs 표집을 2만 번 돌리면 연쇄는 멈추지 않고 돈다. 그러나 τ의 중앙값은 10^{-8}이고, 85%의 시간을 $\tau < 10^{-4}$에서 보내며, 한때 10^{-15}까지 내려간다. 정칙 사전분포 $\pi(\tau) \propto 1$로 같은 일을 하면 중앙값은 $10^{-1.54} = 0.029$다(그림 7.1 왼쪽). 연쇄가 0 근처의 무한한 질량에 빨려 들어가는 것이다.

7.1 봉우리가 여럿일 때

가정 주의 주의 2 — $\hat{R}$은 갇힌 연쇄를 보지 못한다목표가 $\frac{1}{2}\mathcal{N}(-d,1) + \frac{1}{2}\mathcal{N}(d,1)$이고 걸음 폭이 0.5인 무작위 걸음 Metropolis를 생각하자. $d = 3$에서 두 연쇄는 -3, 두 연쇄는 $+3$에서 출발해 5000번 돌리면 $\hat{R} = 1.21$로 수렴하지 않았음을 알린다. 걸음 폭을 4로 넓히면 $\hat{R} = 1.0006$이고 평균은 -0.015로 참값 0에 가깝다. 그러나 $d = 6$에서 네 연쇄를 모두 -6에서 출발시키면 $\hat{R} = 1.001$인데 평균은 -6.04다. 연쇄들은 왼쪽 봉우리에 갇혀 서로 완벽하게 일치한다. 같은 $d = 6$에서 출발점을 흩어 놓으면 $\hat{R} = 6.9$다. $\hat{R}$이 쓸모 있으려면 출발점이 목표분포보다 넓게 흩어져 있어야 한다.

봉우리가 여럿인 사후분포는 드물지 않다. 10회의 혼합모형은 성분의 이름표를 바꿔도 가능도가 같아 $K!$개의 봉우리를 갖는다. 신경망의 가중치도 그렇다. 봉우리 사이를 건너는 방법으로 온도를 바꿔 가며 여러 연쇄를 돌리는 병렬 담금질(parallel tempering)과, 다음 절의 기울기 정보를 쓰는

방법들이 있다.

7.2 제안 폭과 유효 표본 크기

무작위 걸음 Metropolis의 걸음 폭은 받아들임 비율과 걸음의 크기를 맞바꾼다. $\mathcal{N}(0,1)$을 목표로 2만 번 돌리면 결과는 이렇다.

걸음 폭	0.1	2.4	20
받아들임 비율	0.97	0.44	0.07
유효 표본 크기 N_{eff}	45	5011	941

걸음이 너무 작으면 거의 다 받아들이지만 제자리걸음이고, 너무 크면 거의 다 거절해 같은 값을 반복한다. 받아들임 비율 97%는 좋은 소식이 아니다.

> **증명 없이 쓰는 사실** (최적 척도 (Roberts–Gelman–Gilks 1997)) 독립 성분의 d차원 목표에서 정규 제안 $\mathcal{N}(x, \ell^2/d \cdot I)$의 효율은 $d \to \infty$일 때 $\ell \approx 2.38$에서 최대이고, 그때 받아들임 비율은 0.234다. 1차원에서는 약 0.44가 최적이다. 차원이 커지면 걸음 폭을 $1/\sqrt{d}$로 줄여야 하므로 독립 표본 하나를 얻는 데 $O(d)$번의 걸음이 든다. 기울기를 쓰는 방법(Langevin, Hamiltonian)은 이 의존성을 줄인다(8절).

7.3 계산 코드의 점검

사후분포를 계산하는 코드가 맞는지는 8회 8절의 생각으로 점검할 수 있다(Talts 외 2018). 사전분포에서 θ를 뽑고, 그것으로 자료를 만들고, 코드로 사후표본 L개를 뽑는다. 코드가 맞으면 $(\theta, \mathbf{x})$의 결합분포에서 θ와 사후표본들은 $\mathbf{x}$가 주어졌을 때 같은 분포에서 나온 $L+1$개의 교환가능한 값이므로, 사후표본들 사이에서 θ의 순위는 $\{0, \dots, L\}$ 위의 균등분포를 따른다. 사전분포 Beta(2, 2), $n = 20$, $L = 99$로 5000번 되풀이해 순위를 열 칸에 나누었다.

- 올바른 코드 Beta$(2+s,\ 2+n-s)$: 카이제곱 적합도 검정의 유의확률 0.15. 균등하다.
- 자료를 두 번 센 코드 Beta$(2+2s,\ 2+2(n-s))$: 양 끝 칸이 839와 883(기댓값 500)인 U자이고 유의확률은 10^{-146}이다. 사후분포가 너무 좁으면 참값이 사후표본의 바깥으로 자주 떨어진다.
- 사전분포를 Beta(1, 1)로 잘못 넣은 코드: 유의확률 0.56. 이 정도의 사전분포 오류는 5000번으로도 잡지 못한다.

점검은 계산의 오류를 잡지만, 작은 모형의 오류까지 잡는 것은 아니다.

7.4 축소가 개인을 해칠 때

> **가정 주의** 주의 3 — 합계의 위험과 개별 위험James–Stein 정리는 $\sum_i E(\delta_i - \theta_i)^2$을 줄인다. 개별 성분의 위험은 오히려 커질 수 있다. $p = 50$이고 $\theta = (5, 0, \dots, 0)$이면, James–Stein의 합계 위험은 18.1로 $\mathbf{X}$의 50보다 훨씬 작다. 그러나 첫 성분의 위험은 11.2로, X_1의 위험 1의 열 배가 넘는다. 나머지 49개가 0이라 $\|\mathbf{X}\|^2$이 작고, 그래서 튀는 하나까지 0 쪽으로 세게 끌어당기기 때문이다. 타율 표의 Clemente가 이런 선수다. 그가 정말로 다른 선수들과 다르다면, 평균으로 끌어당기는 것은 그에게 불리하다. Efron과 Morris는 한 성분이 움직일 수 있는 양에 상한을 두는 "제한된 이동" 추정량을 제안했다.

축소의 정당성은 성분들이 **교환가능**하다는 판단, 곧 사전에 어느 선수가 더 잘 칠지 모른다는 판단에서 온다(3절의 de Finetti). Clemente가 이미 내셔널리그 타격왕에 네 번 오른 선수라는 것을 안다면 그 정보를 사전분포에 넣어야 한다. 서로 관계없는 양들(예컨대 타율, 수입차 비율, 강수량)을 한 묶음으로 줄여도 합계 위험은 줄지만, 그 합계는 누구에게도 의미 없는 손실이다.

> **가정 주의** 주의 4 — 사전분포의 민감도표본이 크면 사전분포는 잊힌다(Bernstein–von Mises). 그러나 (1) 점 귀무가설의 Bayes 인수는 대립가설의 사전분포의 폭에 끝까지 의존하고(연습 3.2), (2) 계층 모형에서 집단의 수 k가 작으면 τ의 사후분포는 τ의 사전분포에 크게 의존한다. 보고서에는 사전분포를 바꿨을 때 결론이 얼마나 움직이는지 함께 적어야 한다.

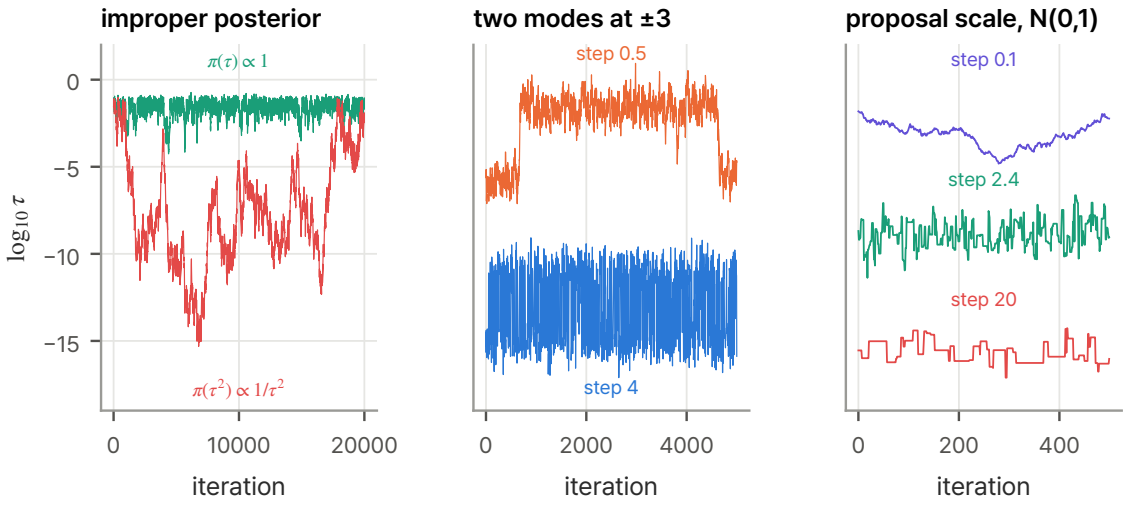


그림 7.1 왼쪽: 타율 자료의 계층 모형 Gibbs 표집에서 $\log_{10}\tau$의 경로. 정칙 사전분포 $\pi(\tau) \propto 1$(청록)과 비정칙 사후분포를 주는 $\pi(\tau^2) \propto 1/\tau^2$(빨강). 가운데: 봉우리 둘($d = 3$)에서 걸음 폭 0.5와 4의 연쇄. 오른쪽: $\mathcal{N}(0, 1)$에서 걸음 폭 0.1, 2.4, 20의 처음 500걸음.

연습문제

연습 7.1 (비정칙 사후분포 판별 ★★) Haldane 사전분포 $\pi(p) \propto p^{-1}(1-p)^{-1}$에서 $\mathrm{b}(n,p)$를 관측할 때 사후분포가 정칙일 필요충분조건은 $0 < s < n$임을 보여라. 이때 사후평균은 MLE s/n과 같음을 보여라. 풀이 50쪽

연습 7.2 (성분 하나의 위험 ★★) $\theta = (a, 0, \dots, 0)$이고 $p \to \infty$이면 James–Stein의 첫 성분 $\left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)X_1$의 위험이 a^2에 가까워짐을 대수의 법칙으로 설명하라. 합계 위험은 p에 비해 어떻게 되는가? 풀이 50쪽

연습 7.3 (받아들임 비율의 함정 ★) 걸음 폭 $\sigma \to 0$이면 무작위 걸음 Metropolis의 받아들임 비율은 1로 가지만 연쇄의 시차 1 자기상관도 1로 감을 설명하라. $\sigma \to \infty$이면 어떻게 되는가? 풀이 51쪽

8

오늘 — 벌점은 사전분포다

역사 Hoerl과 Kennard(1970)는 ridge를 공선성의 처방으로 내놓았고(13회), Tibshirani(1996)는 계수의 절댓값에 벌점을 주는 lasso로 변수 선택과 추정을 한 번에 했다. 두 벌점은 각각 정규 사전분포와 Laplace 사전분포의 사후 최빈값이다. 신경망의 weight decay도 같다. 계산 쪽에서는 물리학이 다시 들어왔다. Paul Langevin(1908)의 브라운 운동 방정식을 이산화하면 기울기를 따라 내려가며 잡음을 더하는 표집기가 된다. Roberts와 Tweedie(1996)는 여기에 MH 보정을 붙였고(MALA), Welling과 Teh(2011)는 미니배치 기울기로 그것을 큰 자료에 썼다(SGLD). 2019–2021년의 확산 생성모형은 자료의 로그밀도의 기울기, 곧 점수를 신경망으로 배워 Langevin 표집을 한다. 그 학습의 핵심 등식은 1956년 Robbins가 경험적 Bayes 논문에서 Tweedie의 것으로 소개한 공식이다.

8.1 ridge와 lasso의 사전분포

명제 8.1 (ridge는 정규 사전분포의 사후평균) $\mathbf{y} \mid \boldsymbol{\beta} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I)$, $\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \tau^2 I)$이면

$$\boldsymbol{\beta} \mid \mathbf{y} \sim \mathcal{N}_p\big((X'X + \lambda I)^{-1} X'\mathbf{y},\ \sigma^2 (X'X + \lambda I)^{-1}\big), \qquad \lambda = \sigma^2/\tau^2.$$

사후평균과 MAP는 모두 13회 정의 8.4의 ridge 추정량 $\hat{\boldsymbol{\beta}}_\lambda$이다.

증명. **전략** 로그 사후밀도를 $\boldsymbol{\beta}$의 이차형식으로 쓰고 행렬의 제곱 완성으로 정규 밀도의 핵과 맞춘다. 정리 1.2(c)의 다차원판이다. 단계별 증명은 연습 8.5(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 52쪽에 있다.

13회에서 λ는 공선성을 다스리는 손잡이였다. 여기서는 잡음과 신호의 분산비 σ^2/τ^2이다. 계수가 0 근처에 몰려 있다고 믿을수록(τ가 작을수록) 세게 줄인다. 5절의 James–Stein은 이 τ^2을 자료로 추정하고, 13회의 교차검증도 결국 같은 일을 다른 방법으로 한다.

명제 8.2 (lasso는 Laplace 사전분포의 MAP) β_j가 독립이고 밀도가 $\frac{1}{2b}e^{-|\beta_j|/b}$(4회 예제 5.4의 Laplace 분포)이면 MAP는

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \ \tfrac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_j |\beta_j|, \qquad \lambda = \sigma^2/b$$

이다. $X'X = I$이면 $z_j = (X'\mathbf{y})_j$에 대해 $\hat{\beta}_j = \operatorname{sign}(z_j)\max(|z_j| - \lambda, 0)$(**연성 임계값**)이다.

증명. 로그 사후밀도에 $-\sigma^2$을 곱하면 위의 목적함수다. $X'X = I$이면 $\|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \|\mathbf{y}\|^2 - 2\boldsymbol{\beta}'\mathbf{z} + \|\boldsymbol{\beta}\|^2$이므로 목적함수가 좌표별 $\frac{1}{2}(z_j - \beta_j)^2 + \lambda|\beta_j|$의 합으로 갈라진다. 이것의 최소점은 15회 명제 8.3(b)의 계산 그대로 연성 임계값이다. ∎

ridge는 모든 계수를 같은 비율로 줄이고 0으로 만들지 않는다. lasso는 작은 계수를 정확히 0으로 만든다. 15회에서 Huber 손실의 "이상치 몫"이 이 연성 임계값이었다. 다만 정확한 0은 MAP의 성질이지 사후분포의 성질이 아니다. Laplace 사전분포의 사후평균은 매끄러운 축소 함수이고 결코 0이 아니다(그림 8.1 가운데). Park과 Casella(2008)는 Laplace 분포가 정규의 척도 혼합이라는 사실(4회 예제 5.4)로 이 사후분포의 Gibbs 표집기를 만들었다(**Bayes lasso**).

증명 없이 쓰는 사실 (weight decay와 MAP) 신경망 학습에서 매 걸음 $\mathbf{w} \leftarrow (1 - \eta\lambda)\mathbf{w} - \eta\nabla L(\mathbf{w})$로 가중치를 조금씩 줄이는 것을 weight decay라 한다. 이것은 $L(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$의 경사하강, 곧 정규 사전분포의 MAP다. 9회 8절의 EWC는 사전분포의 중심과 정밀도를 앞 과제의 Laplace 근사(1절)로 바꾼 것이다. 7회 8절의 MC 드롭아웃은 가중치의 사후분포를 근사하는 한 방법으로 해석된다.

8.2 Langevin 표집

MH는 목표밀도의 값만 쓴다. 기울기 $\nabla \log \pi$를 쓰면 높은 밀도 쪽으로 제안을 기울일 수 있다.

정의 8.3 (비보정 Langevin 알고리즘) 걸음 크기 $\eta > 0$에서 $\theta_{t+1} = \theta_t + \eta\nabla \log \pi(\theta_t) + \sqrt{2\eta}\,\boldsymbol{\xi}_t$, $\boldsymbol{\xi}_t \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I)$를 **ULA**라 한다. $\nabla \log \pi$에는 정규화 상수가 들어가지 않는다.

$\eta \to 0$의 연속시간 극한인 Langevin 확산은 π를 정확히 정상분포로 갖는다. 그러나 $\eta > 0$에서 ULA의 정상분포는 π가 아니다.

명제 8.4 (정규 목표에서 ULA의 치우침) $\pi = \mathcal{N}(0, \sigma^2)$이고 $0 < \eta < 2\sigma^2$이면 ULA의 정상분포는 $\mathcal{N}\big(0,\ \sigma^2/(1 - \eta/(2\sigma^2))\big)$이다.

증명. **전략** 정규 목표에서 ULA는 계수 $1 - \eta/\sigma^2$의 AR(1)이다. 정상분산을 $v = a^2 v + 2\eta$에서 푼다. 단계별 증명은 연습 8.6(유도형)에서 직접 쓴다. 완성된 증명은 부록 A의 풀이 52쪽에 있다.

$\sigma^2 = 1$에서 $\eta = 0.1$이면 정상분산이 1.053, $\eta = 0.5$이면 1.333이다. 20만 걸음의 모의실험은 1.045와 1.328이다. 치우침을 없애는 방법은 둘이다. 걸음 크기를 줄이거나, Langevin 걸음을 MH의 제안 $q(\theta, \cdot) = \mathcal{N}(\theta + \eta\nabla \log \pi(\theta), 2\eta I)$로 쓰고 정리 4.5의 받아들임 단계를 붙인다(**MALA**). SGLD는 $\nabla \log \pi$를 미니배치로 추정한 값으로 바꾸고 $\eta_t \to 0$으로 줄여 간다. 잡음을 더한 SGD다. 기울기를 쓰는 표집기의 다른 갈래인 Hamiltonian Monte Carlo는 Stan의 기본 알고리즘이다.

8.3 Tweedie의 공식과 확산모형

정리 8.5 (Tweedie의 공식) $\mathbf{X} \mid \theta \sim \mathcal{N}_p(\theta, \sigma^2 I)$이고 $\theta \sim \pi$(임의의 사전분포)이며 $m(\mathbf{x}) = \int \varphi_\sigma(\mathbf{x} - \theta)\pi(\theta)d\theta$가 $\mathbf{X}$의 주변밀도이면

$$E(\theta \mid \mathbf{X} = \mathbf{x}) = \mathbf{x} + \sigma^2 \nabla \log m(\mathbf{x}).$$

증명. **전략** 주변밀도를 미분하면 정규 잡음의 밀도에서 $(\theta - \mathbf{x})/\sigma^2$이 내려온다. 그것을 $m(\mathbf{x})$로 나누면 사후평균에서 $\mathbf{x}$를 뺀 것이다.

① $\varphi_\sigma(\mathbf{u}) \propto \exp\{-\|\mathbf{u}\|^2/(2\sigma^2)\}$이므로 $\nabla_{\mathbf{x}} \varphi_\sigma(\mathbf{x} - \theta) = -\dfrac{\mathbf{x} - \theta}{\sigma^2}\varphi_\sigma(\mathbf{x} - \theta)$이다. 연쇄법칙

② 미분과 적분의 순서를 바꾼다.

$$\nabla m(\mathbf{x}) = \int \nabla_{\mathbf{x}} \varphi_\sigma(\mathbf{x} - \theta)\, \pi(\theta)\, d\theta = \frac{1}{\sigma^2} \int (\theta - \mathbf{x})\, \varphi_\sigma(\mathbf{x} - \theta)\, \pi(\theta)\, d\theta. \qquad \text{[T6], ①}$$

③ $m(\mathbf{x}) > 0$으로 나누면 $\varphi_\sigma(\mathbf{x} - \theta)\pi(\theta)/m(\mathbf{x})$는 사후밀도이므로 $\sigma^2 \nabla \log m(\mathbf{x}) = E(\theta \mid \mathbf{X} = \mathbf{x}) - \mathbf{x}$이다. 정의 1.1

조건 확인 ②에는 적분 기호 아래의 미분이 필요하다. $\|\mathbf{u}\|\varphi_\sigma(\mathbf{u})$는 $\mathbf{u}$에 대해 유계이므로 $\|\nabla_{\mathbf{x}} \varphi_\sigma(\mathbf{x} - \theta)\| \le C_\sigma$(상수)이고, 상수는 확률분포 π에 대해 적분가능하다. 이 지배함수는 $\mathbf{x}$에도 θ에도 의존하지 않으므로 사전분포에 적률이 없어도 공식이 성립한다 [T6]. $m(\mathbf{x}) > 0$은 $\varphi_\sigma > 0$에서 나온다. ■

사후평균을 구하는 데 사전분포가 필요 없다. 주변밀도 m만 알면 되고, m은 관측된 X들로 추정할 수 있다. 이것이 Robbins의 경험적 Bayes의 핵심이다. $\pi = \mathcal{N}(\mathbf{0}, \tau^2 I)$이면 $m = \mathcal{N}(\mathbf{0}, (\sigma^2 + \tau^2)I)$, $\nabla \log m(\mathbf{x}) = -\mathbf{x}/(\sigma^2 + \tau^2)$이므로 $E(\theta \mid \mathbf{x}) = (1 - B)\mathbf{x}$, $B = \sigma^2/(\sigma^2 + \tau^2)$로 정리 5.4가 된다.

확산모형은 깨끗한 영상 θ에 잡음을 더한 $\mathbf{X}$에서 θ를 맞히는 신경망(잡음 제거기)을 학습한다. 제곱손실로 학습한 잡음 제거기는 사후평균을 근사하므로(정리 2.1), Tweedie의 공식으로 거꾸로 읽으면 점수 $\nabla \log m$을 근사한다. 여러 잡음 수준의 점수로 Langevin 표집을 하면 새 영상이 나온다. Stein의 1956년과 Robbins의 1956년이 오늘의 생성모형 안에서 만난다.

연습문제

연습 8.1 (Robbins의 Poisson 공식 ★★) $X \mid \theta \sim \mathrm{Poi}(\theta)$, $\theta \sim \pi$이면 $E(\theta \mid X = x) = (x+1)\dfrac{m(x+1)}{m(x)}$임을 보여라($m$은 X의 주변 확률질량함수). 사고 기록에서 작년에 사고를 x번 낸 운전자의 올해 기대 사고 수를 사전분포 없이 추정하는 방법을 설명하라. 풀이 51쪽

연습 8.2 (ULA의 안정성 ★) 명제 8.4에서 $\eta \geq 2\sigma^2$이면 연쇄가 발산함을 보여라. 정규 목표 $\mathcal{N}(\mathbf{0}, \Sigma)$에서 ULA가 안정하려면 걸음 크기가 Σ의 어느 고윳값으로 제한되는가? 13회의 경사하강의 조건수와 비교하라. 풀이 51쪽

연습 8.3 (ridge의 사후 신용구간 ★★) 명제 8.1에서 $X'X = I$이면 β_j의 95% 신용구간이 $\dfrac{z_j}{1+\lambda} \pm 1.96\dfrac{\sigma}{\sqrt{1+\lambda}}$임을 보여라. 같은 β_j가 0일 때와 클 때 이 구간의 빈도주의 포함확률을 비교하라. 풀이 51쪽

연습 8.4 (Laplace 사후평균은 0이 아니다 ★★) $z \mid \beta \sim \mathcal{N}(\beta, 1)$, $\beta \sim \mathrm{Laplace}(b)$이면 $z \neq 0$에서 $E(\beta \mid z) \neq 0$이고 z의 홀함수이며 증가함을 보여라. (힌트: Tweedie의 공식을 한 번 더 미분하면 $\dfrac{d}{dz}E(\beta \mid z) = \mathrm{Var}(\beta \mid z)$이고, m은 짝함수다.) 풀이 51쪽

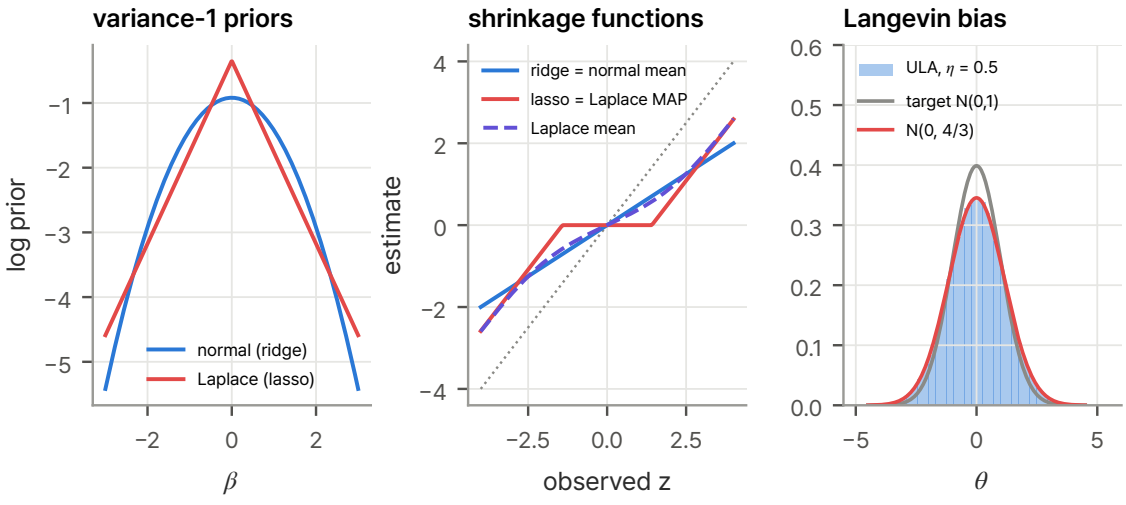


그림 8.1 왼쪽: 정규 사전분포와 같은 분산의 Laplace 사전분포의 로그밀도. 가운데: $z \sim \mathcal{N}(\beta, 1)$을 관측했을 때의 축소 함수. ridge(정규 사후평균, 직선), lasso(Laplace MAP, 연성 임계값), Laplace 사후평균(매끄러운 곡선). 오른쪽: $\mathcal{N}(0,1)$을 목표로 한 ULA의 히스토그램($\eta = 0.5$)과 목표밀도(회색), 명제 8.4의 정상밀도(빨강).

연습 8.5 (ridge의 사후분포: 증명 ★★ 유도형) 명제 8.1를 다음 순서로 증명하라. $A = X'X + \lambda I$로 둔다.

(a) 로그 사후밀도가 $-\frac{1}{2\sigma^2}\big(\|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2\big)$에 상수를 더한 것임을 보여라. $\lambda = \sigma^2/\tau^2$은 어디서 나오는가?

(b) 괄호 안이 $\boldsymbol{\beta}'A\boldsymbol{\beta} - 2\boldsymbol{\beta}'X'\mathbf{y} + \mathbf{y}'\mathbf{y} = (\boldsymbol{\beta} - A^{-1}X'\mathbf{y})'A(\boldsymbol{\beta} - A^{-1}X'\mathbf{y}) + (\boldsymbol{\beta}$에 무관한 항$)$임을 전개해 보여라. A가 대칭이고 가역인 이유를 밝혀라.

(c) (b)가 $\mathcal{N}_p\big(A^{-1}X'\mathbf{y},\ \sigma^2A^{-1}\big)$의 밀도의 핵임을 확인하고(5회 정리 3.15), 정규분포의 평균과 최빈값이 같음을 써서 사후평균과 MAP가 ridge 추정량임을 결론지어라.

풀이 52쪽

연습 8.6 (ULA의 치우침: 증명 ★ 유도형) 명제 8.4를 다음 순서로 증명하라.

(a) $\nabla \log \pi(\theta) = -\theta/\sigma^2$을 넣어 $\theta_{t+1} = a\theta_t + \sqrt{2\eta}\,\xi_t$, $a = 1 - \eta/\sigma^2$임을 보여라.

(b) $0 < \eta < 2\sigma^2$이면 $|a| < 1$임을 보여라.

(c) $\theta_t \sim \mathcal{N}(0, v)$가 정상이면 $\theta_{t+1} \sim \mathcal{N}(0, a^2v + 2\eta)$이므로 $v = a^2v + 2\eta$이다. 이것을 풀어 $v = \sigma^2/(1 - \eta/(2\sigma^2))$을 얻어라.

풀이 52쪽

9

정리와 역사 노트

9.1 한 장 요약

정리	내용	조건
켤레 사전분포 (1.2)	베타–이항, 감마–Poisson, 정규–정규: 사후평균은 정밀도 가중평균	—
Laplace 근사 (1.4)	사후분포 $\approx \mathcal{N}(\hat{\theta}, 1/(n\lvert h''\rvert))$	정칙, 꼭대기 하나
Bayes 추정량 (2.1)	제곱손실 → 사후평균, 절대손실 → 사후중앙값	—
허용성·minimax (2.3)	유일한 Bayes는 허용적, 위험이 상수인 Bayes는 minimax	Bayes 위험 유한
Jeffreys (2.6)	$\sqrt{I(\theta)}$는 모수화에 불변	단조 변환
Bayes 인수 (3.2, 3.4)	사후 오즈 = 사전 오즈 $\times B_{10}$, $\log B_{01} \approx \frac{1}{2}\log n - \frac{1}{2}z^2$	점 귀무가설
MH·Gibbs (4.5, 4.7)	상세 균형 → 정상분포, Gibbs는 받아들임 1의 MH	$m(\mathbf{x})$ 불필요
Gibbs의 자기상관 (4.8)	상관 ρ이면 $N_{\text{eff}}/N = (1-\rho^2)/(1+\rho^2)$	이변량 정규
James–Stein (5.2)	위험 $p - (p-2)^2 E\lVert\mathbf{X}\rVert^{-2} < p$	$p \geq 3$
경험적 Bayes (5.4)	$E\frac{p-2}{\lVert\mathbf{X}\rVert^2} = B$, 대가는 $2B$	정규 사전분포
ridge·lasso (8.1, 8.2)	정규 사전분포의 사후평균, Laplace 사전분포의 MAP	—
ULA의 치우침 (8.4)	정상분산 $\sigma^2/(1-\eta/(2\sigma^2))$	정규 목표
Tweedie (8.5)	$E(\theta \mid \mathbf{x}) = \mathbf{x} + \sigma^2 \nabla \log m(\mathbf{x})$	정규 잡음

모수에 분포를 주면 추론은 조건부분포 하나를 계산하는 일이 된다. 그 대가는 사전분포라는 선택이고, 보상은 하나의 원리로 추정(2절), 검정(3절), 예측(연습 1.4)을 다룰 수 있다는 것이다. 계산은 MCMC가 해결했다. 그러나 이 회차의 가장 놀라운 결과는 사전분포를 믿지 않는 사람에게도 해당한다. Stein의 정리는 빈도주의의 위험으로 표현되고 증명되었다. 그런데 그 처방은 Bayes였다. 같은 종류의 양을 여럿 추정할 때는 서로에게서 빌려 와야 한다.

9.2 Stigler의 명명 법칙

이름과 발견

- **Bayes 정리**: Bayes의 논문에는 오늘의 공식이 없다. 일반형은 Laplace(1774)의 것이다. Stigler(1983)는 반쯤 진지하게 맹인 수학자 Nicholas Saunderson이 먼저 알았을 가능성을 제기했다.
- **Metropolis 알고리즘**: Marshall Rosenbluth의 회고(2003)에 따르면 방법을 만들고 계산한 것은 주로 Rosenbluth 부부와 Teller 부부였고, Metropolis는 컴퓨터를 맡았다. Hastings(1970)의 일반화는 오랫동안 통계학에서 주목받지 못했다.
- **Gibbs 표집**: Geman 형제(1984)가 목표분포인 Gibbs 분포의 이름을 붙였다. 물리학자 J. W. Gibbs(1839–1903)는 표집과 관계가 없다.
- **James–Stein**: 허용적이지 않음은 Stein(1956) 혼자 보였다. James는 명시적인 추정량을 쓴 1961년 논문의 공저자다.
- **Tweedie의 공식**: 처음 출판한 것은 Robbins(1956)였다. Efron(2011)이 이 이름으로 다시 유행시켰다.

9.3 7 Pillars 지도 — 완성

Stigler(2016)는 회귀의 장에 다변량 분석, 베이즈 추론, 인과 추론을 함께 넣었다. Galton의 2단 quincunx(5회)는 사전분포와 가능도로 사후분포를 만드는 장치였고, Stein의 축소는 Galton의 회귀였다. 그래서 이번 회차의 기둥은 **Regression**이다. 사후분포가 가능도에 비례한다는 점에서 **Likelihood**, Jeffreys 사전분포가 Fisher 정보에서 나온다는 점에서 **Information**, 18명의 타율을 함께 써서 각자를 더 잘 추정한다는 점에서 **Aggregation**도 함께 칠한다. 그림 9.1는 16회 전체의 지도다. 일곱 기둥이 모두 칠해졌다.

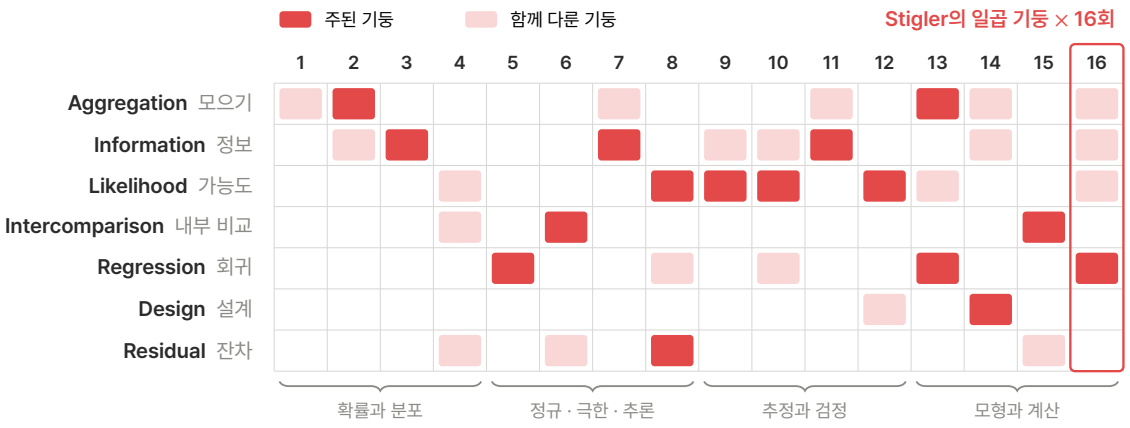


그림 9.1 16회 과정의 일곱 기둥 지도. 진한 칸은 그 회차의 주된 기둥, 옅은 칸은 함께 다룬 기둥이다. Aggregation과 Likelihood는 과정 전체를 가로지르고, Design은 12·14회에, Residual은 4·6·8·15회에 모여 있다.

9.4 과정을 마치며

이 과정은 1654년 Pascal과 Fermat가 나눈 판돈에서 시작했다. 그 뒤 370년의 흐름을 정리하면 세 가지 질문이 남는다.

확률은 무엇에 대한 것인가. 1회의 공리는 답하지 않았다. 4회의 당구대와 이번 회차에서 확률은 모르는 것에 대한 믿음이었고, 8회와 12회에서는 반복되는 절차의 빈도였다. 두 해석은 대부분의 큰 표본에서 같은 숫자를 낸다(Bernstein–von Mises, 2절의 Jeffreys 구간). 갈라지는 곳은 점 귀무가설(3절)과 모수가 많은 문제(5절)다.

자료를 어떻게 줄이는가. 평균(3회), 충분통계량(11회), 최소제곱(13회), 순위(15회), 사후분포(16회)는 모두 자료를 줄이는 방법이다. 각각이 어떤 가정 아래에서 정보를 잃지 않는지가 이 과정의 정리들이 말한 것이다.

모형이 틀리면 무엇이 남는가. 6회의 t 검정의 로버스트성, 10회의 EM의 국소해, 15회의 붕괴점, 이번 회차의 비정칙 사후분포는 모두 정리의 조건이 깨질 때의 이야기였다. 정리를 아는 것은 그 조건을 아는 것이다.

마지막 회차의 마지막 절은 신경망의 생성모형이 1956년의 두 논문 위에 서 있다는 것으로 끝났다. 오늘의 방법은 대부분 이 과정에서 본 오래된 아이디어의 새 조합이다. 이름과 맥락은 바뀌어도 정리는 남는다.

참고문헌

- Bayes, T. An essay towards solving a problem in the doctrine of chances. *Phil. Trans.* 53 (1763), 370–418. · Laplace, P. S. Mémoire sur la probabilité des causes par les événements. (1774). · *Théorie analytique des probabilités*. Paris, 1812.
- Fisher, R. A. On the mathematical foundations of theoretical statistics. *Phil. Trans. A* 222 (1922), 309–368. · Jeffreys, H. *Theory of Probability*. Oxford, 1939. · An invariant form for the prior probability in estimation problems. *Proc. R. Soc. A* 186 (1946), 453–461.
- Good, I. J. *Probability and the Weighing of Evidence*. Griffin, 1950. · A. M. Turing's statistical work in World War II. *Biometrika* 66 (1979), 393–396. · Lindley, D. V. A statistical paradox. *Biometrika* 44 (1957), 187–192.
- de Finetti, B. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* 7 (1937), 1–68. · Savage, L. J. *The Foundations of Statistics*. Wiley, 1954.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21 (1953), 1087–1092. · Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 (1970), 97–109.
- Geman, S., Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of im-

ages. *IEEE PAMI* 6 (1984), 721–741. · Gelfand, A. E., Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *JASA* 85 (1990), 398–409. · Gelman, A., Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statist. Sci.* 7 (1992), 457–472.

- Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp.* 1 (1956), 197–206. · James, W., Stein, C. Estimation with quadratic loss. *Proc. Fourth Berkeley Symp.* 1 (1961), 361–379. · Robbins, H. An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp.* 1 (1956), 157–163.
- Efron, B., Morris, C. Data analysis using Stein's estimator and its generalizations. *JASA* 70 (1975), 311–319. · Stein's paradox in statistics. *Scientific American* 236 (1977), 119–127. · Stigler, S. M. A Galtonian perspective on shrinkage estimators. *Statist. Sci.* 5 (1990), 147–155.
- Hobert, J. P., Casella, G. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *JASA* 91 (1996), 1461–1473. · Gelman, A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1 (2006), 515–534. · Roberts, G. O., Gelman, A., Gilks, W. R. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* 7 (1997), 110–120.
- Tibshirani, R. Regression shrinkage and selection via the lasso. *JRSS B* 58 (1996), 267–288. · Park, T., Casella, G. The Bayesian lasso. *JASA* 103 (2008), 681–686. · Roberts, G. O., Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* 2 (1996), 341–363. · Welling, M., Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. *ICML* (2011).
- Efron, B. Tweedie's formula and selection bias. *JASA* 106 (2011), 1602–1614. · Song, Y., Ermon, S. Generative modeling by estimating gradients of the data distribution. *NeurIPS* (2019). · Stigler, S. M. Who discovered Bayes's theorem? *Amer. Statist.* 37 (1983), 290–296. · *The Seven Pillars of Statistical Wisdom*. Harvard, 2016.
- 자료: HistData `Arbuthnot`; pscl `EfronMorris`; Laplace의 파리 출생 수(1745–1770)는 Laplace(1781, 1812)의 값.

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 6쪽

$\frac{a+s}{a+b+n} = \frac{a+b}{a+b+n}\cdot\frac{a}{a+b} + \frac{n}{a+b+n}\cdot\frac{s}{n}$. 사전분포 $\mathrm{Beta}(a,b)$는 성공 a번, 실패 b번을 이미 본 것처럼 작용한다. 관측 n개와 "가상의 관측" $a+b$개의 비율로 무게가 나뉘므로 $a+b$를 사전 표본 크기라 부른다. 균등 사전분포는 가상의 관측 2개(성공 1, 실패 1)다.

연습 1.2 문제 6쪽

정리 1.2(b)로 사후분포는 $\Gamma(1+122,\ 율\ 1+200) = \Gamma(123,\ 201)$이다. 사후평균은 $123/201 = 0.612$, 95% 신용구간은 $[0.509,\ 0.725]$이다. 관측 200개에 비해 사전분포의 가상 관측이 1개라 MLE 0.61과 거의 같다.

연습 1.3 문제 6쪽

$\mathrm{Beta}(251528, 241946)$의 치우침(왜도)은 -1.1×10^{-4}로 왼쪽 꼬리가 조금 더 길다. $p \le \frac{1}{2}$는 왼쪽 꼬리이므로 정확한 확률이 정규근사보다 크다(1.146 대 1.113, 비 1.030). 왜도 γ의 효과는 Edgeworth 전개에서 대략 $\gamma(z^3-3z)/6$의 상대 보정으로 나타나는데, $z\approx13.6$이면 $z^3\approx2500$이라 10^{-4}의 왜도도 수 퍼센트의 차이를 만든다. 꼬리가 멀수록 정규근사의 상대오차가 커진다.

연습 1.4 문제 7쪽

$P(X_{n+1}=1\mid \mathbf{x}) = \int p\,\pi(p\mid\mathbf{x})dp = E(p\mid\mathbf{x}) = \frac{a+s}{a+b+n}$. $a=b=1$이면 $\frac{s+1}{n+2}$로 Laplace의 계승규칙이다.

2절

연습 2.1 문제 13쪽

$\delta_{a,b} - p = \frac{S-np}{n+a+b} + \frac{a-(a+b)p}{n+a+b}$ 이고 첫 항의 평균은 0, 분산은 $\frac{np(1-p)}{(n+a+b)^2}$ 이므로 위험 공식이 나온다. $a=b=c$ 이면 분자는 $np - np^2 + c^2(1-2p)^2 = c^2 + (n-4c^2)p - (n-4c^2)p^2$ 이고 $c=\sqrt{n}/2$ 에서 p 항이 사라져 $n/4$ 가 남는다. MLE가 나은 곳은 $\frac{p(1-p)}{20} < 0.00835$, 곧 $p(1-p) < 0.167$ 이므로 $p < 0.212$ 또는 $p > 0.788$ 이다.

연습 2.2 문제 13쪽

$\bar{X}$ 는 불편이고, $E\bar{X}^2 < \infty$ 이다. Bayes 추정량이라면 Bayes 위험이 0이어야 하지만 $\bar{X}$ 의 위험은 $1/n > 0$ 이다. 사후평균은 $\frac{n\bar{X}}{n+1/\tau^2}$ 이고 Bayes 위험은 사후분산 $\frac{1}{n+1/\tau^2} = \frac{\tau^2}{n\tau^2+1}$ 이다. $\tau \to \infty$ 이면 $1/n$ 으로, $\bar{X}$ 의 위험으로 간다. $\bar{X}$ 는 Bayes 추정량들의 극한(일반화 Bayes)이다.

연습 2.3 문제 13쪽

$\pi(\sigma) \propto 1/\sigma$ 를 σ^2 으로 바꾸면 $\pi(\sigma^2) \propto 1/\sigma^2$ 이다. $T = \sum X_i^2$ 이라 하면 사후밀도는 $(\sigma^2)^{-1}(\sigma^2)^{-n/2}e^{-T/(2\sigma^2)}$ 이고, $U = T/\sigma^2$ 으로 바꾸면 $(d\sigma^2 \propto U^{-2}dU)$ $U^{n/2-1}e^{-U/2}$, 곧 χ^2_n 이다. 따라서 $P\big(\chi^2_{n,0.975} \le T/\sigma^2 \le \chi^2_{n,0.025}\big) = 0.95$ 를 σ^2 에 대해 풀면 6회의 구간과 같은 식이 된다. 사후확률의 문장과 신뢰구간의 문장이 숫자까지 일치하는 예다.

연습 2.4 문제 13쪽

$\psi = \log\frac{p}{1-p}$ 이면 $\frac{d\psi}{dp} = \frac{1}{p(1-p)}$ 이므로 ψ 의 평평한 밀도는 p 에서 $\frac{1}{p(1-p)}$ 이다. $s=0$ 이면 사후밀도는 $p^{-1}(1-p)^{n-1}$ 이고 $\int_0^\epsilon p^{-1}dp = \infty$ 이므로 정규화할 수 없다.

연습 2.5 문제 13쪽

베타분포의 평균과 중앙값의 차이는 대략 왜도 × 표준편차/6이다. 왜도가 -1.1×10^{-4}, 표준편차가 7.1×10^{-4} 이므로 차이는 약 1.3×10^{-8} 이다. 노트북의 정확한 값도 -1.31×10^{-8} 이다. 사후분포가 거의 대칭이므로 제곱손실과 절대손실이 같은 답을 낸다.

연습 2.6 문제 13쪽

(a) $E(\delta\theta) = E\big[\theta E(\delta \mid \theta)\big] = E[\theta\cdot\theta] = E\theta^2$. 첫 등호는 반복기댓값(3회 정리 2.5)이고 θ 는 θ 가 주어지면 상수라 밖으로 나온다. (b) $E(\delta\theta) = E\big[\delta E(\theta \mid \mathbf{X})\big] = E[\delta\cdot\delta] = E\delta^2$. δ 는 $\mathbf{X}$ 의 함수이므로 $\mathbf{X}$ 가 주어지면 상수다. (c) $E(\delta-\theta)^2 = E\delta^2 - 2E(\delta\theta) + E\theta^2 = E\delta^2 - 2E\delta^2 + E\delta^2$ 에서 (a)로 $E\theta^2 = E(\delta\theta) = E\delta^2$ 이므로 0이다. 둘째 적률의 유한성은 세 기댓값이 모두 유한해 전개와

반복기댓값을 쓸 수 있게 한다($|\delta\theta| \le \frac{1}{2}(\delta^2+\theta^2)$). (d) Bayes 위험이 0이 아니면, 곧 $E(\delta-\theta)^2>0$이면, δ는 불편이거나 Bayes이거나 둘 중 하나만일 수 있다. 정규 평균처럼 모든 추정량의 Bayes 위험이 양수인 모형에서는 불편추정량이 Bayes 추정량이 되지 못한다.

3절

연습 3.1 문제 16쪽

$10\log_{10}3 = 4.77$, $10\log_{10}5 = 6.99$이고 합은 11.76 deciban, 곧 가능도비 15다. 사후 오즈는 $\frac{1}{20}\times 15 = \frac{3}{4}$, 사후확률은 $3/7 = 0.43$이다.

연습 3.2 문제 16쪽

$\bar{x}$가 충분통계량이므로 그 분포만 비교하면 된다. H_0에서 $\bar{X}\sim\mathcal{N}(0,1/n)$, H_1에서 주변분포는 $\mathcal{N}(0,\tau^2+1/n)$이다. 밀도의 비는

$$B_{01} = \sqrt{\frac{\tau^2+1/n}{1/n}}\exp\!\left\{-\frac{\bar{x}^2}{2}\left(n-\frac{1}{\tau^2+1/n}\right)\right\} = \sqrt{1+n\tau^2}\exp\!\left\{-\frac{z^2}{2}\cdot\frac{n\tau^2}{1+n\tau^2}\right\}.$$

$\tau\to\infty$이면 지수 부분은 $e^{-z^2/2}$로 가고 $\sqrt{1+n\tau^2}\to\infty$이므로 $B_{01}\to\infty$다. 대립가설을 한없이 흐리게 하면 어떤 자료도 귀무가설을 지지한다(Bartlett의 역설). 비정칙 사전분포로는 Bayes 인수를 정의할 수 없다.

연습 3.3 문제 16쪽

$I(\frac{1}{2}) = 4$, $\pi_1\equiv 1$, $z = 2\sqrt{n}(\hat{p}-\frac{1}{2}) = 13.640$이므로 $\log B_{10} \approx \frac{z^2}{2} - \frac{1}{2}\log\frac{4n}{2\pi} = 93.03 - 6.33 = 86.70$, 상용로그로 37.65다. 정확한 값 37.656과 소수 둘째 자리까지 맞는다.

연습 3.4 문제 16쪽

처음 k번 중 흰 공이 j번 나왔다면 항아리에는 흰 공 $a+j$개, 검은 공 $b+k-j$개가 있다. 특정 순서의 확률은 분자에 흰 공 차례마다 $a, a+1, \dots, a+s-1$, 검은 공 차례마다 $b, \dots, b+n-s-1$이 나오고 분모는 $(a+b)(a+b+1)\cdots(a+b+n-1)$이므로 순서에 무관하다. 그 값은 $\frac{\Gamma(a+s)\Gamma(b+n-s)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(a+b)}{\Gamma(a+b+n)} = \frac{B(a+s,b+n-s)}{B(a,b)} = \int_0^1 p^s(1-p)^{n-s}\mathrm{Beta}(a,b)(dp)$이다.

4절

연습 4.1 문제 20쪽

$\pi P = \pi$에서 $\pi_1 a = \pi_2 b$이므로 $\pi = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)$이고, 이 식이 바로 상세 균형이다. P의 고윳값은 1과 $1-a-b$이므로 $P^n = \mathbf{1}\pi + (1-a-b)^n R$ (R 고정된 행렬)이다.

연습 4.2 문제 20쪽

각 열의 합이 $0.9 + 0.1 = 1$이므로 전이행렬이 이중확률행렬이고 균등분포가 정상분포다. 그러나 $\pi_1 k(1,2) = 0.3 \neq \pi_2 k(2,1) = 0.0\overline{3}$이다. 확률이 한 방향으로 순환하므로 가역이 아니다.

연습 4.3 문제 21쪽

MH 비는 $\frac{\pi(y)g(x)}{\pi(x)g(y)} = \frac{w(y)}{w(x)}$이다. w는 7회의 중요도 가중치다. 중요도 표집이 가중치로 평균을 고친다면, 독립 MH는 가중치가 큰 점에 더 오래 머무는 방식으로 고친다. g의 꼬리가 가벼우면 w가 꼬리에서 한없이 커지고, 연쇄가 그런 점에 들어가면 매우 오래 갇힌다. 정상분포는 맞지만 수렴이 기하급수적이지 않다.

연습 4.4 문제 21쪽

$B = 0$이면 $\hat{V} = \frac{n-1}{n} W$이고 $\hat{R} = \sqrt{(n-1)/n}$이다. $\hat{R} > 1$이 되려면 $B/n > W/n$, 곧 연쇄 평균들의 흩어짐 $\frac{1}{m-1}\sum(\bar{\theta}_j - \bar{\theta})^2$이 연쇄 안의 분산을 n으로 나눈 값 W/n보다 커야 한다. 연쇄들이 독립 표본이라면 연쇄 평균의 분산은 W/n 정도이므로, 그보다 큰 흩어짐은 연쇄들이 아직 다른 곳에 있다는 증거다.

연습 4.5 문제 21쪽

좌표 j의 갱신 핵 K_j는 받아들임 확률 1의 MH이므로 정리 4.5로 $\pi(\theta)K_j(\theta,\theta') = \pi(\theta')K_j(\theta',\theta)$이다. 무작위 순서의 핵은 $K = \frac{1}{d}\sum_j K_j$이고, 상세 균형은 선형이므로 합에도 성립한다. 체계적 순서 $K_1 K_2 \cdots K_d$의 수반은 $K_d \cdots K_1$이라 일반적으로 자기 자신과 다르다.

연습 4.6 문제 21쪽

(a) 조건부밀도의 정의 $\pi(\theta_j \mid \theta_{-j}) = \pi(\theta)/\pi(\theta_{-j})$에서 $\pi(\theta) = \pi(\theta_{-j})\pi(\theta_j \mid \theta_{-j})$이다. θ'은 θ_{-j}를 공유하므로 같은 식에서 $\pi(\theta') = \pi(\theta_{-j})\pi(\theta'_j \mid \theta_{-j})$이다. (b) 제안은 좌표 j만 바꾸므로 $q(\theta,\theta') = \pi(\theta'_j \mid \theta_{-j})$, 거꾸로 $q(\theta',\theta) = \pi(\theta_j \mid \theta_{-j})$이다. MH 비는

$$\frac{\pi(\theta')q(\theta',\theta)}{\pi(\theta)q(\theta,\theta')} = \frac{\pi(\theta_{-j})\pi(\theta'_j \mid \theta_{-j}) \cdot \pi(\theta_j \mid \theta_{-j})}{\pi(\theta_{-j})\pi(\theta_j \mid \theta_{-j}) \cdot \pi(\theta'_j \mid \theta_{-j})} = 1$$

이므로 받아들임 확률은 $\min\{1,1\}=1$이다. (c) 정리 4.5로 각 K_j는 π에 대해 상세 균형을 만족하고, 정리 4.3로 $\pi K_j=\pi$이다. $\pi K_1=\pi$, $\pi K_2=\pi$이면 $\iint \pi(x)K_1(x,y)K_2(y,z)\,dy\,dx = \int\Big[\int \pi(x)K_1(x,y)dx\Big]K_2(y,z)\,dy = \int \pi(y)K_2(y,z)\,dy = \pi(z)$이다. 피적분함수가 음이 아니므로 적분 순서를 바꿀 수 있다(Tonelli). 따라서 한 바퀴 $K_1\cdots K_d$도 π를 보존한다.

연습 4.7 문제 21쪽

(a) (X,Y)가 평균 0, 분산 1, 상관 ρ의 이변량 정규이면 $Y\mid X=x\sim\mathcal{N}(\rho x,1-\rho^2)$이고 $X\mid Y=y\sim\mathcal{N}(\rho y,1-\rho^2)$이다(5회 정리 3.13). 따라서 Y_{t-1}은 ρX_{t-1}에 분산 $1-\rho^2$의 독립 잡음을 더한 것이고, X_t도 같다. 매 단계 새 표준정규를 뽑으므로 잡음들은 서로, 그리고 과거와 독립이다. (b) 대입하면 $X_t=\rho\big(\rho X_{t-1}+\sqrt{1-\rho^2}\epsilon_{t-1}\big)+\sqrt{1-\rho^2}\epsilon'_t=\rho^2X_{t-1}+\eta_t$, $\eta_t=\rho\sqrt{1-\rho^2}\epsilon_{t-1}+\sqrt{1-\rho^2}\epsilon'_t$이다. 두 잡음이 독립이므로 $\operatorname{Var}\eta_t=\rho^2(1-\rho^2)+(1-\rho^2)=(1-\rho^2)(1+\rho^2)=1-\rho^4$이다. (c) 정상 상태에서 $\operatorname{Cov}(X_t,X_{t-k})=\rho^{2k}\operatorname{Var}X$이므로 시차 k의 자기상관은 ρ^{2k}이다. $1+2\sum_{k\ge1}\rho^{2k}=1+\frac{2\rho^2}{1-\rho^2}=\frac{1+\rho^2}{1-\rho^2}$이고, N_{eff}/N은 그 역수 $\frac{1-\rho^2}{1+\rho^2}$이다.

5절

연습 5.1 문제 26쪽

$E[1/V]=\int_0^\infty v^{-1}\frac{v^{p/2-1}e^{-v/2}}{2^{p/2}\Gamma(p/2)}dv=\frac{2^{p/2-1}\Gamma(p/2-1)}{2^{p/2}\Gamma(p/2)}=\frac{1}{2(p/2-1)}=\frac{1}{p-2}$. 피적분함수는 0 근처에서 $v^{p/2-2}$이므로 $p/2-2\le-1$, 곧 $p\le2$이면 적분이 발산한다. $1\cdot2$차원에서 $\|\mathbf{X}\|$가 0 근처에 올 확률이 충분히 커서 $1/\|\mathbf{X}\|^2$의 평균이 무한하다는 것이 James–Stein이 3차원에서 시작하는 이유다.

연습 5.2 문제 26쪽

(a) $g(z)=z$: $E[(Z-\theta)Z]=1$이므로 $EZ^2=\theta^2+1$. $g(z)=z^2$: $E[(Z-\theta)Z^2]=2EZ=2\theta$이므로 $EZ^3=\theta EZ^2+2\theta=\theta^3+3\theta$. (b) 왼쪽은 $E[Z\operatorname{sign}(Z)]=E|Z|=\sqrt{2/\pi}=0.80$, 오른쪽은 $g'=0$(거의 모든 곳)이므로 0이다. g가 0에서 뛰어 절대연속이 아니다. 증명의 $\int_0^u g'=g(u)-g(0)$이 깨진다. James–Stein의 $\mathbf{g}$는 원점 한 점에서만 특이하고, 원점을 지나지 않는 거의 모든 좌표직선 위에서 매끄러우며, 도함수가 $\|\mathbf{x}\|^{-2}$ 정도라 $p\ge3$에서 적분가능하다. 이 조건에서 다차원 보조정리가 성립한다(Stein 1981).

연습 5.3 문제 26쪽

$E[1/\|\mathbf{X}\|^2]=E\big[E(1/\chi^2_{p+2K}\mid K)\big]=E[1/(p+2K-2)]$(연습 5.1). $1/x$는 볼록이므로 Jensen으로

$E[1/\|\mathbf{X}\|^2] \ge 1/E\|\mathbf{X}\|^2 = 1/(p+\|\theta\|^2)$이고, 위험은 $p - \frac{(p-2)^2}{p+\|\theta\|^2}$ 이하다. $t = \|\theta\|^2$이라 하면

$$p - \frac{(p-2)^2}{p+t} - \frac{pt}{p+t} = \frac{p^2 + pt - (p-2)^2 - pt}{p+t} = \frac{4(p-1)}{p+t} \le \frac{4(p-1)}{p} < 4.$$

선형 축소의 하한은 $\min_c E\|c\mathbf{X}-\theta\|^2 = \min_c\{c^2 p + (1-c)^2 t\}$이다. James–Stein은 θ를 모르고도 이 하한에서 4 이내에 있다.

연습 5.4 문제 27쪽

두 추정량은 $A = \{\|\mathbf{X}\|^2 < p-2\}$ 밖에서 같다. A에서 $c = \frac{p-2}{\|\mathbf{X}\|^2} - 1 > 0$이고 손실의 차이는 $\|-c\mathbf{X}-\theta\|^2 - \|\theta\|^2 = c^2\|\mathbf{X}\|^2 + 2c\mathbf{X}'\theta$이다. 첫 항은 양수다. $\|\mathbf{X}\| = r$이 주어지면 $\mathbf{X}$는 반지름 r의 구 위에서 $e^{\mathbf{x}'\theta}$에 비례하는 밀도를 가지므로, $\mathbf{x}$와 그것을 θ에 수직인 평면에 대칭시킨 점을 짝지으면 $\mathbf{x}'\theta > 0$인 쪽의 밀도가 더 크다. 따라서 $E[\mathbf{X}'\theta \mid \|\mathbf{X}\|] \ge 0$이고, A는 $\|\mathbf{X}\|$로만 정해지므로 $E[(c^2\|\mathbf{X}\|^2 + 2c\mathbf{X}'\theta)\mathbf{1}_A] > 0$이다.

연습 5.5 문제 27쪽

$\mathbf{X} \sim \mathcal{N}_p(\theta, \sigma^2 I)$에서 보조정리는 $E[(\mathbf{X}-\theta)'\mathbf{g}] = \sigma^2 E[\nabla\cdot\mathbf{g}]$이다. $\delta = \mathbf{X} - aS^2\mathbf{X}/\|\mathbf{X}\|^2$의 위험은 S^2과 $\mathbf{X}$의 독립성으로

$$p\sigma^2 - 2a\sigma^2 E[S^2](p-2)E\|\mathbf{X}\|^{-2} + a^2 E[S^4]E\|\mathbf{X}\|^{-2} = p\sigma^2 - \sigma^4 E\|\mathbf{X}\|^{-2}\Big[2a(p-2) - a^2\frac{\nu+2}{\nu}\Big]$$

이다($ES^2 = \sigma^2$, $ES^4 = \sigma^4(\nu+2)/\nu$). 괄호는 $a = \frac{(p-2)\nu}{\nu+2}$에서 최대 $\frac{(p-2)^2\nu}{\nu+2} > 0$이다. $\nu \to \infty$이면 σ^2을 아는 경우로 돌아간다.

연습 5.6 문제 27쪽

(a) 정리 1.2(c)에 $n=1$, $\sigma^2=1$, $m=0$을 넣으면 정밀도는 $1+1/\tau^2$이고 사후평균은 $\frac{X_i}{1+1/\tau^2} = \frac{\tau^2}{1+\tau^2}X_i = (1-B)X_i$, 사후분산은 $\frac{\tau^2}{1+\tau^2} = 1-B$이다. θ_i들과 X_i들이 좌표마다 독립이라 θ_i의 사후분포는 X_i만으로 정해진다. 정리 2.1(a)로 각 좌표의 최소 사후 기대손실은 사후분산이고, 합의 기댓값이 Bayes 위험 $p(1-B)$이다. (b) $X_i = \theta_i + \varepsilon_i$에서 $\theta_i \sim \mathcal{N}(0,\tau^2)$와 $\varepsilon_i \sim \mathcal{N}(0,1)$이 독립이므로 $X_i \sim \mathcal{N}(0, 1+\tau^2)$이고 서로 독립이다. $\|\mathbf{X}\|^2 = (1+\tau^2)V$, $V = \sum_i X_i^2/(1+\tau^2) \sim \chi^2_p$이므로 $E\big[\frac{p-2}{\|\mathbf{X}\|^2}\big] = \frac{p-2}{1+\tau^2}E[1/V] = \frac{1}{1+\tau^2} = B$이다(연습 5.1). (c) Bayes 위험은 $E_\pi R(\theta,\delta_{\mathrm{JS}}) = p - (p-2)^2 E_\pi\big[E_\theta(1/\|\mathbf{X}\|^2)\big]$이고, 반복기댓값으로 안쪽은 $\mathbf{X}$의 주변분포에 대한 기댓값 $E[1/\|\mathbf{X}\|^2] = \frac{B}{p-2}$이다. 따라서 $p-(p-2)B = p(1-B) + 2B$이다.

6절

연습 6.1 문제 30쪽

$c = 1 - \frac{15\times 0.004332}{0.0825} = 1 - 0.788 = 0.212$. Clemente: $0.2654 + 0.212 \times (0.400 - 0.2654) = 0.2654 + 0.0285 = 0.294$. Munson: $0.2654 + 0.212 \times (0.178 - 0.2654) = 0.2654 - 0.0185 = 0.247$.

연습 6.2 문제 30쪽

$y_i = \mu + (\theta_i - \mu) + (y_i - \theta_i)$에서 두 괄호는 독립인 $\mathcal{N}(0,\tau^2), \mathcal{N}(0,\sigma^2)$이므로 $y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2+\tau^2)$. 따라서 $E\frac{1}{k-1}\sum(y_i-\bar y)^2 = \sigma^2+\tau^2$이다. 타율 자료에서 $\frac{0.0825}{17} = 0.004854$, $\hat\tau^2 = 0.004854 - 0.004332 = 0.000521$, $\hat\tau = 0.023$이다. 계층 Bayes의 사후중앙값 0.030보다 작다. 사후분포는 τ가 더 클 가능성까지 반영한다.

연습 6.3 문제 30쪽

$\mathbf{1}/\sqrt{p}$를 첫 열로 하는 직교행렬 Q로 $\mathbf{W} = Q'\mathbf{X}$라 하면 $W_1 = \sqrt{p}\bar X$, $(W_2,\dots,W_p)$는 $\mathbf{X} - \bar X\mathbf{1}$의 좌표이고 $\|\mathbf{X} - \bar X\mathbf{1}\|^2 = \sum_{j\ge 2} W_j^2$이다. 추정량은 W_1을 그대로 두고 $(W_2,\dots,W_p)$에 $p-1$차원의 James–Stein$(p-1-2 = p-3)$을 쓴 것이다. 위험은 $1 + \left[(p-1) - (p-3)^2 E\|\mathbf{W}_{2:p}\|^{-2}\right] < p$ $(p-1 \ge 3$, 곧 $p \ge 4)$.

연습 6.4 문제 30쪽

θ를 적분한 사후밀도는 $\frac{1}{\tau}\prod_i \varphi_{\sqrt{\sigma^2+\tau^2}}(y_i - \mu)$이다. μ로 적분한 값 $h(\tau)$는 $\tau \to 0$일 때 양의 유한한 극한 $h(0)$(분산 σ^2의 정규 가능도를 μ로 적분한 것)을 갖는다. 따라서 $\int_0^{\epsilon} h(\tau)/\tau\, d\tau \ge \frac{h(0)}{2}\int_0^{\epsilon'} d\tau/\tau = \infty$. 사후분포는 비정칙이다. 평평한 $\pi(\tau) \propto 1$에서는 같은 적분이 $\int_0 h(\tau)d\tau < \infty$이고, $\tau \to \infty$ 쪽은 $h(\tau) \sim \tau^{-(k-1)}$로 $k \ge 3$이면 유한하다.

7절

연습 7.1 문제 34쪽

사후밀도는 $p^{s-1}(1-p)^{n-s-1}$이고, $\int_0^1$이 유한할 필요충분조건은 $s-1 > -1$, $n-s-1 > -1$, 곧 $0 < s < n$이다. 이때 사후분포는 Beta$(s, n-s)$이고 평균은 s/n이다.

연습 7.2 문제 34쪽

$\|\mathbf{X}\|^2 = (a + Z_1)^2 + \sum_{j\ge 2} Z_j^2$이고 대수의 법칙으로 $\sum_{j\ge 2} Z_j^2/p \to 1$이다. 축소 인자는 $1 -$

$\frac{p-2}{p-1+(a+Z_1)^2+o_P(p)} \to 0$이므로 첫 성분은 0으로 가고 위험은 a^2으로 간다. 합계 위험은 $p\|\theta\|^2/(p+\|\theta\|^2)$ 근처에서 많아야 4 위(연습 5.3), 곧 a^2+4 이하로 p에 비해 아주 작다. $p=50$, $a=5$에서는 축소 인자가 대략 $\frac{a^2+1}{p+a^2}=0.35$라 첫 성분의 위험이 11.2에 머문다.

연습 7.3 문제 34쪽

$\sigma \to 0$이면 $\pi(y)/\pi(x) \to 1$이라 거의 다 받아들이지만 $|X_{t+1}-X_t| \le \sigma|\epsilon|$이 작아 시차 1 자기상관이 1로 간다. $\sigma \to \infty$이면 제안이 대부분 밀도가 거의 0인 곳에 떨어져 거절되고, 연쇄가 같은 값에 머물러 역시 자기상관이 1로 간다. 두 극단 모두 $N_{\text{eff}} \to 0$이다.

8절

연습 8.1 문제 38쪽

$E(\theta \mid x)=\frac{\int \theta\cdot\theta^x e^{-\theta}/x!\,\pi(d\theta)}{m(x)}=\frac{(x+1)\int \theta^{x+1}e^{-\theta}/(x+1)!\,\pi(d\theta)}{m(x)}=(x+1)\frac{m(x+1)}{m(x)}$. 운전자 N명 중 작년에 사고를 x번 낸 사람의 수를 N_x라 하면 $m(x)\approx N_x/N$이므로 $(x+1)N_{x+1}/N_x$가 추정값이다. 사전분포를 몰라도 되고, 사고를 많이 낸 사람의 올해 기대 사고 수가 작년보다 작다는 평균으로의 회귀가 자동으로 들어간다.

연습 8.2 문제 38쪽

$\eta \ge 2\sigma^2$이면 계수 $a=1-\eta/\sigma^2 \le -1$이므로 $|a|\ge 1$이고 분산이 줄지 않는다(엄밀히는 $|a|>1$에서 발산, $a=-1$에서 잡음이 쌓인다). $\mathcal{N}(\mathbf{0},\Sigma)$에서 Σ의 고유벡터 방향마다 AR(1)이 되므로 $\eta < 2\lambda_{\min}(\Sigma)$가 필요하다. 그 걸음으로 가장 긴 방향(분산 $\lambda_{\max}$)을 가로지르려면 $\lambda_{\max}/\eta \approx \kappa$번의 걸음이 든다. 13회 정리 8.1의 경사하강이 조건수 κ에 묶이는 것과 같다.

연습 8.3 문제 38쪽

$X'X=I$, $\lambda=\sigma^2/\tau^2$이면 사후분포는 $\mathcal{N}\big(\frac{z_j}{1+\lambda},\frac{\sigma^2}{1+\lambda}\big)$이므로 구간이 나온다. $z_j\sim\mathcal{N}(\beta_j,\sigma^2)$이면 포함은 $|Z-\lambda\beta_j/\sigma|\le 1.96\sqrt{1+\lambda}$ ($Z\sim\mathcal{N}(0,1)$)과 같다. $\lambda=1$이면 $\beta_j/\sigma=0,1,2,3$에서 포함확률이 0.994, 0.962, 0.780, 0.410이다. 사전분포가 맞다면 β_j에 대해 평균한 포함확률은 정확히 0.95이지만, 큰 계수에서는 구간이 0 쪽으로 끌려가 크게 빗나간다.

연습 8.4 문제 38쪽

Tweedie의 공식 $E(\beta\mid z)=z+(\log m)'(z)$를 미분하면 $1+(\log m)''(z)$이다. $\varphi''(u)=(u^2-1)\varphi(u)$이므로 $m''/m=E[(\beta-z)^2\mid z]-1$, $(m'/m)^2=(E[\beta-z\mid z])^2$이고 $(\log m)''=\mathrm{Var}(\beta\mid z)-1$이다. 따라서 $\frac{d}{dz}E(\beta\mid z)=\mathrm{Var}(\beta\mid z)>0$으로 순증가한다. 사전분포와 잡음이 모두 대칭이라 m

은 짝함수, $(\log m)'$은 홀함수이고 $E(\beta \mid 0) = 0$이다. 순증가하는 홀함수이므로 $z \neq 0$이면 0이 아니다. lasso의 MAP가 $|z| \le \lambda$에서 정확히 0인 것과 대조된다.

연습 8.5 문제 39쪽

(a) 가능도는 $\exp\{-\|\mathbf{y}-X\boldsymbol{\beta}\|^2/(2\sigma^2)\}$, 사전밀도는 $\exp\{-\|\boldsymbol{\beta}\|^2/(2\tau^2)\}$에 비례한다. 로그를 더하면 $-\frac{1}{2\sigma^2}\|\mathbf{y}-X\boldsymbol{\beta}\|^2 - \frac{1}{2\tau^2}\|\boldsymbol{\beta}\|^2 = -\frac{1}{2\sigma^2}\big(\|\mathbf{y}-X\boldsymbol{\beta}\|^2 + \frac{\sigma^2}{\tau^2}\|\boldsymbol{\beta}\|^2\big)$이므로 $\lambda = \sigma^2/\tau^2$이다. (b) $\|\mathbf{y}-X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}'(X'X+\lambda I)\boldsymbol{\beta} - 2\boldsymbol{\beta}'X'\mathbf{y} + \mathbf{y}'\mathbf{y}$이다. $\hat{\boldsymbol{\beta}} = A^{-1}X'\mathbf{y}$로 두면 $(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'A(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'A\boldsymbol{\beta} - 2\boldsymbol{\beta}'A\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'A\hat{\boldsymbol{\beta}}$이고 $A\hat{\boldsymbol{\beta}} = X'\mathbf{y}$이므로 두 식의 차는 $\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'A\hat{\boldsymbol{\beta}}$로 $\boldsymbol{\beta}$에 무관하다. A는 대칭이고, $\mathbf{v}\neq\mathbf{0}$이면 $\mathbf{v}'A\mathbf{v} = \|X\mathbf{v}\|^2 + \lambda\|\mathbf{v}\|^2 > 0$이라 양정치이므로 가역이다. (c) 로그 사후밀도가 $-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(A/\sigma^2)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+$상수이므로 정밀도 행렬 A/σ^2, 평균 $\hat{\boldsymbol{\beta}}$인 정규분포, 곧 $\mathcal{N}_p(\hat{\boldsymbol{\beta}}, \sigma^2A^{-1})$이다 (5회 정리 3.15). 정규밀도는 평균에서 최대이므로 사후평균과 MAP가 모두 $\hat{\boldsymbol{\beta}} = (X'X+\lambda I)^{-1}X'\mathbf{y}$, ridge 추정량이다.

연습 8.6 문제 39쪽

(a) $\log\pi(\theta) = -\theta^2/(2\sigma^2)+$상수이므로 $\nabla\log\pi(\theta) = -\theta/\sigma^2$이고 $\theta_{t+1} = \theta_t - \eta\theta_t/\sigma^2 + \sqrt{2\eta}\xi_t = a\theta_t + \sqrt{2\eta}\xi_t$이다. (b) $0<\eta<2\sigma^2$이면 $0<\eta/\sigma^2<2$이므로 $-1<a<1$이다. (c) $\theta_t \sim \mathcal{N}(0,v)$와 ξ_t가 독립이면 $a\theta_t + \sqrt{2\eta}\xi_t \sim \mathcal{N}(0, a^2v+2\eta)$이다. 정상이면 $v = a^2v + 2\eta$, 곧 $v = \frac{2\eta}{1-a^2}$이다. $1-a^2 = \frac{2\eta}{\sigma^2} - \frac{\eta^2}{\sigma^4} = \frac{2\eta}{\sigma^2}\big(1-\frac{\eta}{2\sigma^2}\big)$이므로 $v = \frac{\sigma^2}{1-\eta/(2\sigma^2)}$이다. $|a|<1$이면 $\operatorname{Var}\theta_t$는 어디서 출발하든 이 값으로 수렴하므로 정상분포는 하나다.