Lecture 3: Multiparameter Models

Professor Alexander Franks

1/22/24

Announcements

- Today: Chapter 3
- Next: Chapter 4 (can skim) and Chapter 5

- HW 1 due Sunday
- OH Friday 2-3pm.
- #3 Drop Alaska + Hawaii.

$\hat{\theta}_{MLE}$ to $\hat{\theta}_{posterior\ mean}$.

Bayesian inference in the normal model

- Assume $y_1, \ldots y_n \sim N(\mu, \sigma^2)$ with σ^2 a known constant
- Let's start with a Jeffreys' (improper prior): $p(\mu) \propto \text{const}$
- What is the posterior distribution $p(\mu \mid y_1, \ldots y_n, \sigma^2)$?

$$\sqrt{I(\mu)} \propto n/\sigma^2 \propto \text{const}$$

$$P(\mu \mid y_1, \ldots y_n) \propto \underbrace{e^{-\frac{\sum (y_i - \mu)^2}{2\sigma^2}}}_{L(\mu)} \times \text{const}$$

$$\propto e^{\frac{-n\mu^2 - 2\sum y_i \mu + \text{const}}{2\sigma^2}} \quad 1/n \quad 1/n$$

$$\propto \quad e^{\frac{-(\mu-\bar{y})^2}{2\sigma^2/n}}$$

$$P(\mu \mid y_1, \ldots y_n, \sigma^2) \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

$$\hat{\mu}_{MLE} = \bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\sqrt{\sigma^2/n}\,(\mu - \bar{y}) \sim N(0,1)$$

Bayesian inference in the normal model

- Assume $y_1, \ldots y_n \sim N(\mu, \sigma^2)$ with σ^2 a known constant
- The normal prior distribution is conjugate for μ in the normal sampling model
- Sampling distribution, prior distribution and posterior distribution are all normal.
- Assume the prior is $p(\mu) \sim N(\mu_0, \sigma^2/\kappa_0)$ $\sim \tau^2$
- What are the parameters of the posterior $p(\mu \mid y_1, \ldots y_n, \sigma^2)$?

Conjugate

$$P(\mu \mid y_1, \dots, y_n) \propto L(\mu)\, P(\mu)$$

$$\propto e^{-\frac{(\mu-\bar{y})^2}{2\sigma^2/n}}\, e^{-\frac{(\mu-\mu_0)^2}{2\sigma^2/k_0}}$$

$$\propto \exp\left[-\frac{1}{2}\frac{n+k_0}{\sigma^2}\mu^2 - 2\left(\frac{n\bar{y}}{\sigma^2} + \frac{k_0}{\sigma^2}\mu_0\right)\mu + \text{const}\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{n+k_0}{\sigma^2}\right)\left(\mu - \frac{\frac{n\bar{y}}{\sigma^2} + \frac{k_0}{\sigma^2}\mu_0}{\frac{(n+k_0)}{\sigma^2}}\right)^2\right]$$

Mean: $\dfrac{n/\sigma^2}{\frac{(n+k_0)}{\sigma^2}}\bar{y} + \dfrac{\frac{k_0}{\sigma^2}}{\frac{n+k_0}{\sigma^2}}\mu_0$

$$= \frac{n}{n+k_0}\bar{y} + \frac{k_0}{n+k_0}\mu_0$$

$$= w\hat{\mu}_{MLE} + (1-w)\mu_0$$

$$P(\mu \mid y_1, \dots, y_n) \sim N\left(w\bar{y} + (1-w)\mu_0, \frac{\sigma^2}{n+k_0}\right)$$

A conjugate prior for the normal likelihood

- The normal distribution is conjugate for the normal likelihood
  - Often called the "normal-normal model"
- $Y_i \sim N(\mu, \sigma^2)$ and $\mu \sim N(\mu_0, \sigma^2/\kappa_0)$ implies that the posterior distribution $p(\mu \mid y)$ is also normally distributed:

$$\mu \mid Y \sim N(\mu_n, \tau_n^2)$$

where $\mu_n = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}$ and $\tau_n^2 = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}$

The posterior mean and pseudo-counts

$\tau^2 = \frac{\sigma^2}{K_0}$

$$\mu_n = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\mu_0 + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\bar{y}$$
$$= (1 - w)\mu_0 + w\bar{y}$$

where $w = \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}$

Fisher Weight.

$y_i \overset{ind}{\sim} N(\mu, \sigma_i^2)$, MLE for μ?

$\hat{\mu}_{MLE} = \sum \frac{1}{\sigma_i^2} y_i \Big/ \sum 1/\sigma_i^2$

Known mean, unknown variance

- Assume we have n mean-zero normal observations with variance σ^2

$$L(\sigma^2) \propto \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(-y_i-\mu)^2}{2\sigma^2}}$$

- Define $d_i = (y_i - \mu)$ for notational convenience
- What is $p(\sigma^2 \mid \mu, d_1, \ldots d_n)$?

denom.

$$L\left(\sigma^2\right) \propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\sum_{i=1}^{n} d_i^2 / \left(2\sigma^2\right)\right\}, \ \sigma^2 > 0$$

$$P(\sigma^2) \propto \frac{1}{\sigma^2}$$

* $p(\sigma) = 1/\sigma$
* $p(\log \sigma) \propto$ const

$$P_{\log\sigma}(\log\sigma) \propto \text{const} \qquad \sigma^2 = \exp(2\log\sigma)$$

$$P_{\sigma^2}(\sigma^2) \propto \cancel{P_{\log\sigma}(\quad)} \left| \frac{d\log\sigma}{d\sigma^2} \right|$$

$$\frac{d\,\tfrac{1}{2}\log(\sigma^2)}{d\sigma^2} \propto 1/\sigma^2$$

Known mean, unknown variance

- Assume we have n mean-zero normal observations with variance σ^2
- Jeffreys prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$ (careful!)
- The posterior:
  $p(\sigma^2 \mid y) \propto \left(\sigma^2\right)^{-n/2-1} \exp\left\{-\sum_{i=1}^{n} d_i^2 / \left(2\sigma^2\right)\right\}$
- This distribution called an *inverse-Gamma* distribution.
  - If $X \sim \mathrm{Gamma}(a, b)$ then $\frac{1}{X} \sim \mathrm{Inv\text{-}Gamma}(a, b)$.

$\text{Inv-Gam}\left(n/2, \frac{\sum d_i^2}{2}\right)$

$$\Downarrow$$

$$\underline{\text{Scaled Inv-}}\chi^2\left(n, \frac{\sum d_i^2}{n}\right)$$

$$\chi^2_k = \sum_k z_i^2 \quad , \quad z_i \sim N(0,1)$$

$$\sigma^2 \chi^2_k \qquad \chi^2_k \equiv \text{Gam}\left(k/2, 1/2\right)$$

A conjugate prior distribution

- In general, the conjugate prior distribution has the form:

$$p(\sigma^2) \propto (\sigma^2)^{-k_1} e^{\frac{k_2}{\sigma^2}}$$

- Inverse-gamma: $p(y \mid a, b) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp\left(-\frac{b}{y}\right)$
  - The mean of an inverse-Gamma is b/(a-1)

Joint inference for the mean and variance

In the normal model we typically factorize the prior distribution $p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2)$.

Specifically:

$$\sigma^2 \sim Inv - Gamma(\nu_0/2, \nu_0/2\sigma_0^2)$$
$$\mu|\sigma^2 \sim \text{ normal } \left(\mu_0, \sigma^2/\kappa_0\right)$$
$$Y_1, \ldots, Y_n|\mu, \sigma \sim \text{ i.i.d. normal } \left(\mu, \sigma^2\right)$$

- ν_0 is interpreted as the prior sample size
- σ_0^2 is a prior sample variance

$$P(\mu, \sigma^2 \mid y_1, \dots y_n) \propto \left[\prod^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{\sigma^2}}\right] e^{-\frac{(\mu-\mu_0)^2}{2\sigma^2/\kappa_0}} (\sigma^2)^{-\frac{\nu_0}{2}-1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}}$$

$\kappa_0 \to 0$ implies what about $P(\mu \mid \sigma^2)$

$\nu_0 \to 0$ implies what about $P(\sigma^2)$

$P(\mu \mid \sigma^2, y_1, \dots y_n)$ "Conditional post"

$P(\sigma^2 \mid y_1, \dots y_n)$ "Marginal post."

Joint inference for the mean and variance

- We write down $p(\mu, \sigma^2 | \, y_1, \ldots y_n) \propto L(\mu, \sigma^2) p(\mu, \sigma^2)$. Now what?
- Estimands of potential interest:
  - $E[\mu \mid y_1, \ldots y_n]$
  - $E[\sigma \mid y_1, \ldots y_n]$
  - $E[\underbrace{\sigma/\mu}_{\text{CV.}} \mid y_1, \ldots y_n]$ (coefficient of variation)

Posterior inference for arbitrary functions

Method of transformations

1. Find the inverse, $\cancel{\theta = g^{-1}(\gamma) = \frac{e^{\gamma}}{1+e^{\gamma}}}$

2. Compute $\frac{dg^{-1}(\gamma)}{d\gamma}$

$\gamma = \frac{\sigma}{\mu}$

3. Find

$$p_{\gamma}(\gamma \mid y_1, \ldots y_n) = \left| \frac{dg^{-1}(\gamma)}{d\gamma} \right| \times p_{\theta}(g^{-1}(\gamma) \mid y_1, \ldots y_n)$$

$f\left(\frac{\sigma}{\mu}\right) p\left(\frac{\sigma}{\mu} \mid y_1, \ldots y_n\right)$

Posterior inference for arbitrary functions

For any $\gamma = g(\theta)$ we have

$$E(g(\theta)|y) = \int g(\theta)p(\theta|y)d\theta$$

Examples:

- $E[\gamma \mid y] = \int \log(\frac{\theta}{1-\theta})p(\theta \mid y)d\theta$
- $Pr(\theta \in R \mid y) = E(I[\theta \in R]|y)$

Monte Carlo

$$\int g(\theta)p_\theta(\theta \mid y_1, \ldots y_n)d\theta = \int \gamma p_\gamma(\gamma \mid y_1, \ldots y_n)d\gamma$$

Law of the Unconscious Statstician

LE CASINO
CASINO MONTE-CARLO
CASINO NIGHT
À CIEL OUVERT
ROCK

Monte Carlo Method

- $\overline{\theta} = \sum_{s=1}^{S} \theta^{(s)}/S \to \mathrm{E}[\theta|y_1, \ldots, y_n]$
- $\sum_{s=1}^{S} \left(\theta^{(s)} - \overline{\theta}\right)^2/(S-1) \to \mathrm{Var}[\theta|y_1, \ldots, y_n]$
- $\#\left(\theta^{(s)} \leq c\right)/S \to \Pr(\theta \leq c|y_1, \ldots, y_n)$
- the α-percentile of $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\} \to \theta_\alpha$

Almost sure convergence due to the SLLN

Monte Carlo Error

$$E[g(\theta) \mid y_1, \ldots y_n],$$

$$\frac{\sum g(\theta_i)}{S} \xrightarrow[\text{CLT}]{d} N\left(E[g(\theta) \mid y_1, \ldots y_n], \frac{Var(g(\theta) \mid y_1, \ldots y_n)}{S}\right)$$

assuming $\theta_i \overset{iid}{\sim} P(\theta_i \mid \ldots)$

- If posterior samples are independent then:

$$p(\overline{g(\theta)}) \to N\left(g(\theta), \frac{\mathrm{Var}(g(\theta) \mid y_1, \ldots y_n)}{S}\right)$$

-

- Just because we can write down the (proportional) posterior, doesn't mean we can sample from the distribution!
  - With exponential families and conjugate priors, sampling algorithms are available
  - For more "complicated" distributions we need new tools (e.g. MCMC)

Back to the normal model

Example: midge wing length

- Modeling wing length of different specifies of midge (small, two-winged flies)
- From prior studies: mean wing length close to 1.9mm.
- Prior mean for μ is $\mu_0 = 1.9$ and Jeffreys prior for σ^2.
- Prior sample sizes: choose $\kappa_0 = 1$
- $(\bar{y}, s^2) = (1.804, 0.0169)$ are the sufficient statistics

Working with the log posterior

- As always, we will write down $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$
- In code, we always work with the **log-posterior** for numerical reasons
  - Mathematically it makes no difference, but computationally it is important
  - $L(\theta) \propto \prod p(y_i \mid \theta)$ is very small for moderate sample size (underflow)
  - $\ell(\theta) = \sum log(p(y_i \mid \theta))$ is numerically stable
- Monte Carlo methods only require that we can evaluate the log posterior

Grid approximation

```r
log_normal_posterior <- Vectorize(function(mu, sigma) {

  ### log likelihood
  sum(dnorm(y, mu, sigma, log=TRUE)) +
  ## plus log prior
  dnorm(mu, mu0, sigma/sqrt(k0), log=TRUE) +
  -2*log(sigma)

})


post_grid <- as_tibble(
  expand.grid(seq(1.6, 2.0, by=0.001),
              seq(0.01, 0.25, by=0.001)))
colnames(post_grid) <- c("mu", "s")
```

Grid approximation

```
post_grid %>%
  mutate(log_density = log_normal_posterior(mu, s)) %>%
  mutate(density = exp(log_density - max(log_density))) %>%
  ggplot() +
  geom_raster(aes(mu, s, fill=density)) +
  xlim(c(1.7, 1.95)) + ylim(c(0.05, 0.2)) +
  xlab(expression(mu)) +
  ylab(expression(sigma)) +
  theme_bw(base_size=16
           ) +
  scale_fill_continuous(type="viridis")
```

0.20
0.15
0.10
0.05
σ
1.70
1.75
1.80
1.85
1.90
1.95
μ
density
1.00
0.75
0.50
0.25
0.00

Contour Plot (Standard Deviation)

```r
post_grid %>%
  mutate(density = exp(log_normal_posterior(mu, s))) %>%
  ggplot() +
  geom_contour(aes(mu, s, z=density, colour=stat(level))) +
  xlim(c(1.7, 1.95)) + ylim(c(0.05, 0.2)) +
  xlab(expression(mu)) + ylab(expression(sigma)) +
  ggtitle("Posterior Contours") +
  theme_bw(base_size=16) +
  scale_color_continuous(type="viridis") +
  geom_hline(yintercept=s, linetype="dashed") +
  geom_vline(xintercept=ybar, linetype="dashed")
```

Posterior Contours
0.20
0.15
0.10
0.05
σ
1.70
1.75
1.80
1.85
1.90
1.95
μ
level
80000
60000
40000
20000

Contour Plot (Precision)

```r
post_grid$prec <- 1/post_grid$s^2

post_grid %>%
  mutate(density = exp(log_normal_posterior(mu, sqrt(1/prec)))) %>%
  ggplot() +
  geom_contour(aes(mu, prec, z=density, colour=stat(level))) +
  xlim(c(1.7, 1.95)) + ylim(c(25, 200)) +
  xlab(expression(mu)) + ylab(expression(1/sigma^2)) +
  ggtitle("Posterior Contours") +
  theme_bw(base_size=16) +
  scale_color_continuous(type="viridis") +
  geom_hline(yintercept=1/s^2, linetype="dashed") +
  geom_vline(xintercept=ybar, linetype="dashed")
```

$P(\mu, 1/\sigma^2 \mid \text{data})$

Sampling from the joint posterior

- Contour and raster plots allow us to visualize the posterior (in two dimensions)
  - Grid sampling approximates posterior samples
  - Need to know approximately where the high posterior density is (not easy)
- When we have more than 2 parameters visualization isn't feasible
- How do we summarize the posterior?
  - e.g. posterior means, posterior probabilities, intervals, etc..

Visualizing Posterior Samples

```
post_grid %>%
  mutate(density = exp(log_normal_posterior(mu, sqrt(1/prec)))) %>%
  mutate(density = density / sum(density)) -> post_grid

sample_indices <- sample(1:nrow(post_grid), size=10000, replace=TRUE, prob=

post_grid[sample_indices, ] %>%
  mutate(prec = 1/s^2) %>%
  ggplot() +
  geom_point(aes(x=mu, y=prec)) +
  theme_bw(base_size=16)
```

200
100
prec
1.7
1.8
1.9
2.0
mu

Visualizing Posterior Samples

```
post_grid %>%
  mutate(density = exp(log_normal_posterior(mu, sqrt(1/prec)))) %>%
  mutate(density = density / sum(density)) -> post_grid

sample_indices <- sample(1:nrow(post_grid), size=10000, replace=TRUE, prob=

post_grid[sample_indices, ] %>%
  mutate(prec = 1/s^2) %>%
  ggplot() +
  ggdensity::geom_hdr(aes(x=mu, y=prec)) +
  theme_bw(base_size=16)
```

Visualizing Posterior Samples

Direct Sampling

- $p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y) p(\sigma^2 \mid y)$
- Need to know how to sample from $p(\sigma^2 \mid y) = \int p(\mu, \sigma^2 \mid y) d\mu$ and $p(\mu \mid \sigma^2, y)$
- With the proposed conjugate priors, this integral is tractable

Direct Sampling

- $p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y) p(\sigma^2 \mid y)$
- Need to know how to sample from $p(\sigma^2 \mid y) = \int p(\mu, \sigma^2 \mid y) d\mu$ and $p(\mu \mid \sigma^2, y)$
- With the proposed conjugate priors, this integral is tractable

$$\left\{1/\sigma^2 \mid y_1, \ldots, y_n\right\} \sim \text{gamma}\left(\nu_n/2, \nu_n \sigma_n^2/2\right), \text{ where}$$

$$\nu_n = \nu_0 + n$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0\right.$$

Direct Sampling

```r
## posterior parameters
kn <- k0 + n
nun <- nu0 + n
mun <- (k0 * mu0 + n * ybar) / kn
s2n <- (nu0*s20 + (n-1)*s2  + k0*n / kn * (ybar - mu0)^2) / nun

nsamps <- 10000
prec_samps <- rgamma(nsamps, nun/2, nun*s2n/2)
mu_samps <- rnorm(nsamps, mun, sqrt((1/prec_samps) / kn))

tibble(mu=mu_samps, prec=prec_samps) %>%
  ggplot() +
  geom_point(aes(x=mu, y=prec)) +
  #ggdensity::geom_hdr(aes(x=mu, y=prec)) +
  theme_bw(base_size=16)
```

Direct Sampling

Bayes Estimators

Why the posterior mean?

- Often times we need to make a “decision” by providing a single estimate
- The posterior provides a full distribution over θ, which can be summarized in infinitely many ways
- Specify a *loss function* which describes the cost of estimating $\hat{\theta}$ when the truth is θ

Bayes Estimators

- The *loss function*: $L(\hat{\theta}, \theta)$
  - Squared error: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
  - Absolute error: $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$
- The *Bayes risk* is the posterior expected loss: $E_{\theta|y}[L(\hat{\theta}, \theta)]$
- Summarize the posterior by minimizing the Bayes risk.
- An estimator $\hat{\theta}$ is said to be a Bayes estimator if it minimizes the Bayes risk among all estimators.

Examples Squared error loss

$$\min_{\hat{\theta}} E_{\theta|y}(\hat{\theta} - \theta)^2 = \min_{\hat{\theta}} \int (\hat{\theta} - \theta)^2 p(\theta \mid y) d\theta$$

The Bias-Variance Tradeoff

- The prior distribution (usually) makes your estimator biased…
- But the prior distribution also (usually) reduces the variance!
- Bayes estimators are (usually) consistent
- Example: compute the frequentist mean and variance of the posterior mean.

Example: IQ scores

- Scoring on IQ tests is designed to yield a N(100, 15) distribution for the general population
- We observe IQ scores for a sample of n individuals from a particular town and estimate μ, the town-specific IQ score
- If we lacked knowledge about the town, a natural choice would be $\mu_0 = 100$
- Suppose the true parameters for this town are $\mu = 112$ and $\sigma = 13$
  - The town is smarter on average than the general population

Example: IQ scores

- What is the mean squared error of the MLE? MSE of the posterior mean?

Example: IQ scores

- What is the mean squared error of the MLE? MSE of the posterior mean?

- $\mathrm{MSE}\left[\hat{\mu}_{MLE}\right] = \mathrm{Var}\left[\hat{\mu}_{MLE}\right] = \frac{\sigma^2}{n} = \frac{169}{n}$

Example: IQ scores

- What is the mean squared error of the MLE? MSE of the posterior mean?

- $\mathrm{MSE}\left[\hat{\mu}_{MLE}\right] = \mathrm{Var}\left[\hat{\mu}_{MLE}\right] = \frac{\sigma^2}{n} = \frac{169}{n}$

- $\mathrm{MSE}\left[\hat{\mu}_{PM}|\theta_0\right] = w^2 \frac{169}{n} + (1-w)^2 144$

Example: IQ scores

- What is the mean squared error of the MLE? MSE of the posterior mean?

- $\mathrm{MSE}\left[\hat{\mu}_{MLE}\right] = \mathrm{Var}\left[\hat{\mu}_{MLE}\right] = \frac{\sigma^2}{n} = \frac{169}{n}$

- $\mathrm{MSE}\left[\hat{\mu}_{PM}|\theta_0\right] = w^2 \frac{169}{n} + (1-w)^2 144$

- Reminder: $w = \frac{n}{\kappa_0 + n}$. For what values of n and κ_0 is the MSE smaller for the posterior mean estimator than the maximum likelihood?

Example: IQ scores

The Multivariate Normal Distribution

Dirichlet-Multinomial

Metagenomics example

- Metagenomics is the study of genetic material recovered directly from environmental samples
- Map counts of genetic material to counts of microbial species
- Assume species are sampled with replacement
  - Observed sample is a multinomial distribution
- Total counts isn't meaningful (hard to control how much total sample)
- Relative counts are meaningful

Multinomial Density

- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ with $\sum_i = 1^k \theta_i = 1$
- If $Y = (Y_1, \ldots, Y_K) \sim Mult(n, \theta)$, then:
  - $Y_i \sim Bin(n, \theta_i)$
  - $Y_i + Y_j \sim Bin(n, \theta_i + \theta_j)$
- What is $\hat{\theta}_{MLE}$?

Dirichlet Distribution