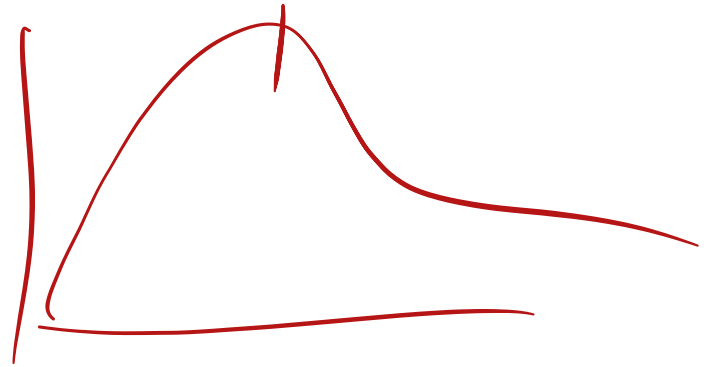
Aside:Bayes Estimators

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.

$\int_{\theta} L(\hat{\theta}, \theta) \, p(\theta | y) \, d\theta$

$||$

Examples Squared error loss

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

$$= E[\theta | y]$$

Do This

$$\frac{d}{d\hat{\theta}} \int (\hat{\theta} - \theta)^2 p(\theta | y) d\theta$$

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

The Bias-Variance Tradeoff

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

Example: IQ scores

$Y_i \overset{iid}{\sim} N(\mu, \sigma=13)$, $\mu=112$

$\mu \sim N(\mu_0, \frac{\sigma}{\sqrt{\kappa_0}})$

- 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

$$Y_i \overset{iid}{\sim} N(\mu, \sigma = 13), \ \mu = 112$$

$$\mu \sim N\left(\mu_0, \frac{\sigma}{\sqrt{K_0}}\right)$$

$$E[\mu \mid y_1, \ldots, y_n] = w\bar{y} + (1-w)\mu_0$$

$$w = \frac{n}{n + K_0}$$

Post. mean estimator: $w\bar{Y} + (1-w)\mu_0$

$K_0 = 0, \ \rightarrow \ \hat{\mu}_{MLE} = \bar{Y}$

$$E_{Y|\mu}\left[(\hat{\mu} - \mu)^2\right] = Var(\hat{\mu}) + Bias(\hat{\mu})^2$$

$$Var(\hat{\mu}_{MLE}) = Var(\bar{Y}) = \frac{\sigma^2}{n}$$

$$Bias(\hat{\mu}_{MLE}) = E[\bar{Y}] - \mu = 0$$

$$MSE(\hat{\mu}_{MLE}) = \frac{\sigma^2}{n}$$

$$\mathrm{Var}(\hat{\mu}_{pm}) = \mathrm{Var}(w\bar{Y} + (1-w)\mu_0) = w^2 \frac{\sigma^2}{n}$$

$$\begin{aligned} \mathrm{Bias}(\hat{\mu}_{pm}) &= E[w\bar{Y} + (1-w)\mu_0] - \mu \\ &= w\mu + (1-w)\mu_0 - \mu \\ &= (1-w)(\mu_0 - \mu) \end{aligned}$$

$$\mathrm{MSE}(\hat{\mu}_{pm}) = w^2 \frac{\sigma^2}{n} + (1-w)^2 (\mu_0 - \mu)^2$$

$$\mathrm{MSE}(\hat{\mu}_{pm}) \; \begin{matrix} > \\ < \\ \approx \\ ?? \end{matrix} \; \mathrm{MSE}(\hat{\mu}_{MLE})$$

Example: IQ scores

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

$\sigma = 15$

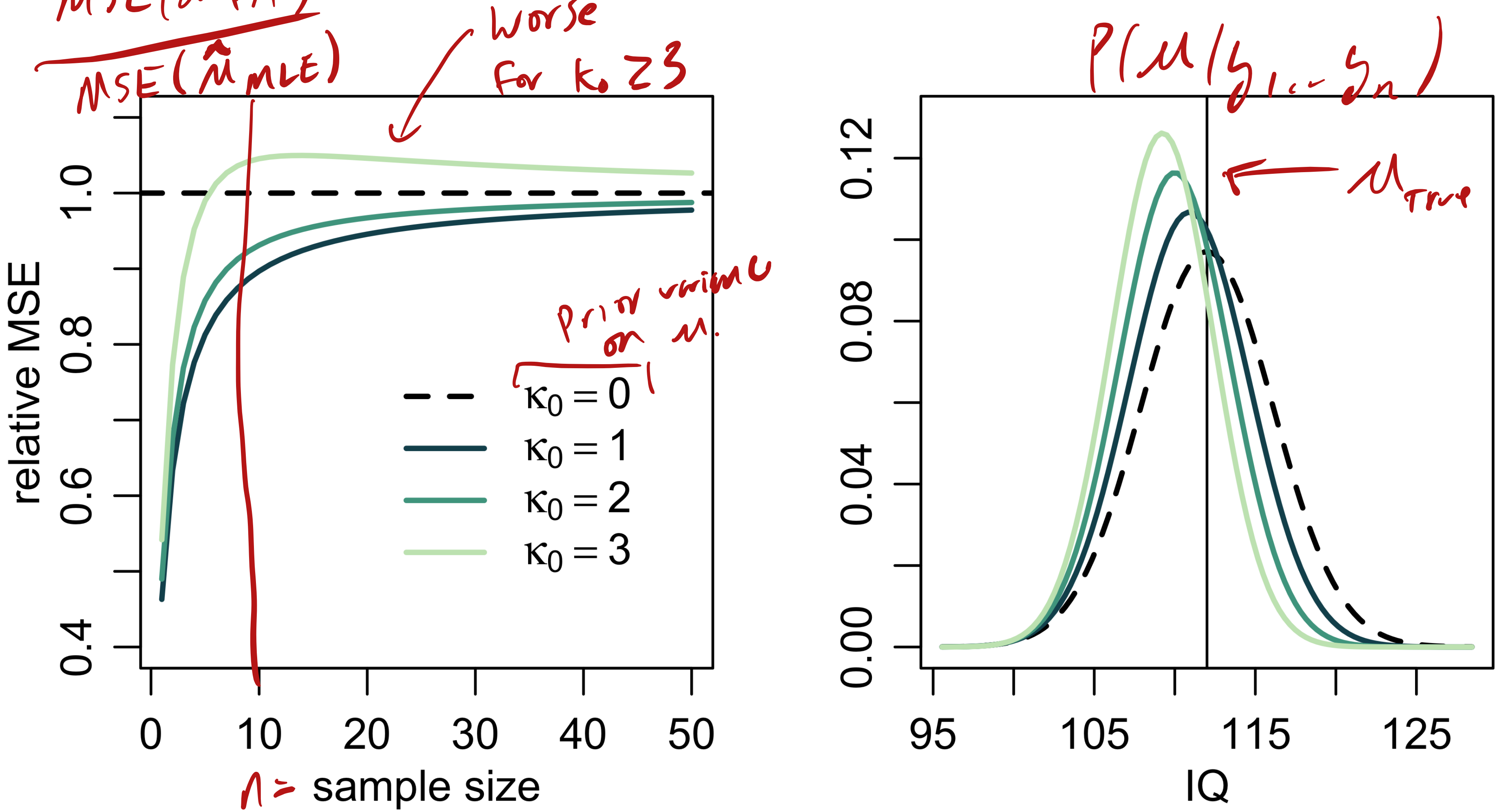
$\mu_0 \quad \mu$

$(100 - 112)^2$

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

$\mu \sim N(\mu_0, \frac{\sigma^2}{\kappa_0})$



The Multivariate Normal Distribution

$$Y_{p\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_p \end{bmatrix} \sim N_p(\mu, \Sigma)$$

$$P(Y = y \mid \mu, \Sigma) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right]$$

$\mu \in \mathbb{R}^p$

$\Sigma \in S_p^+$ cone of positive def. matrices

$\underline{a'\Sigma a > 0}$ $\quad \frac{p(p+1)}{2}$ free parameters.

$$P(\mu \mid \Sigma, y_1, \dots y_n)$$

$$\mu \sim N_p(\mu_0, \Lambda_0)$$

$$L(\mu) \propto \prod_{i=1}^{n} \cancel{(2\pi)^{-1/2}} \cancel{|\Sigma|^{-1/2}} \exp\left[-\frac{1}{2}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right]$$

$$\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right)$$

$$\propto \exp\left(-\frac{1}{2}\mu^T A\mu + \mu^T b\right)$$

$$A = n\Sigma^{-1}, \quad b = n\Sigma^{-1}\bar{y}$$

$$P(\mu) \propto \exp\left(\mu^T\Lambda_0^{-1}\mu^T + \mu^T\Lambda_0^{-1}\mu_0\right)$$

$$P(\mu \mid y_1, \dots y_n) \propto \exp\left(\mu^T A_n \mu + \mu^T \underbrace{(b + \Omega_0^{-1}\mu_0)}_{n\Sigma^{-1}\bar{y} + \Lambda_0^{-1}\mu_0}\right)$$

$$A_n = \Sigma^{-1} + \Lambda_0^{-1}$$

$$\sim N(\mu_n, \Sigma_n)$$

$$\Sigma_n = (n\Sigma^{-1} + \Lambda_0^{-1})^{-1}$$

$$\mu_n = (n\Sigma^{-1} + \Lambda_0^{-1})^{-1}(n\Sigma^{-1}\bar{y} + \Lambda_0^{-1}\mu_0)$$

Special case

$$\Lambda_0^{-1} = k_0\Sigma^{-1}$$

$$\mu_n = (n\Sigma^{-1} + k_0\Sigma^{-1})^{-1}(n\Sigma^{-1}\bar{y} + k_0\Sigma^{-1}\mu_0)$$

$$= \frac{1}{n+k_0}\Sigma(n\Sigma^{-1}\bar{y} + k_0\Sigma^{-1}\mu_0)$$

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

$\frac{\sigma^2}{k_0}$

(univariate analog)

$\mu \sim N(\mu_0, \tau^2)$

Σ unknown, $\mu = 0$ known

$\vec{y}_1, \ldots \vec{y}_n \sim N(0, \Sigma)$

$$\underbrace{\begin{pmatrix} \vec{y}_1 & \vec{y}_2 & \cdots & \vec{y}_n \end{pmatrix}}_{p \times n}$$

ith row
kth column

$$\underset{p \times p}{S} = YY^T = \begin{pmatrix} \sum_{i=1}^n y_{i1}^2 & \left(\sum_{i=1}^n y_{ij} y_{ik}\right) \\ & \ddots & \\ & & \sum y_{ip}^2 \end{pmatrix}$$

sum of squares matrix

$$S \sim \text{Wishart}(n, \Sigma)$$

$$P(S) \propto |S|^{\frac{(n-p-1)}{2}} e^{-\frac{1}{2} tr(\Sigma^{-1} S)}$$

$n > p \rightarrow YY^T$ is pos. def. w.p. 1.

$\rightarrow$ symmetric.

$$\boxed{E[S] = n\Sigma}$$

$$\frac{YY^T}{n} \rightarrow \Sigma$$

$X \sim \text{Wish}, \quad X^{-1} \sim \text{Inv-Wish.}$

$\Sigma \sim \text{Inv-Wish}$ is conjugate prior

$$L(\Sigma) \propto |\Sigma|^{-n/2} \exp\Big(-\tfrac{1}{2} \underbrace{y'\Sigma^{-1}y}_{\text{scalar.}}\Big)$$

$tr(ABC) = tr(BCA)$

$$\propto |\Sigma|^{-n/2} \exp\Big(-\tfrac{1}{2} tr(y'\Sigma^{-1}y)\Big)$$

$$|\Sigma|^{-n/2} \exp\Big(-\tfrac{1}{2} tr(\Sigma^{-1} \underbrace{(yy^T)}_{p \times p,\ = S})\Big)$$

$$\Sigma \sim IW(\nu, \Lambda_0) \to |\Sigma|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2} tr(\Sigma^{-1}\Lambda_0)}$$

$$\to P(\Sigma | y_1, \dots y_n) \sim \underline{IW(n+\nu, S+\Lambda_0)}$$

$$\Sigma \sim IW(\nu, \Lambda_0) \qquad \underline{\nu > p-1}$$

Default: $\Lambda_0 = I$

$IW(\text{small}, I)$

Jeffreys $\propto |\Sigma|^{-\frac{(p+1)}{2}}$

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

of species $\quad \sum \theta_i = 1, \quad \sum_{k=1}^{K} Y_k = n$

- 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:
  - $Y_i \sim Bin(n, \theta_i)$
  - $Y_i + Y_j \sim Bin(n, \theta_i + \theta_j)$

$\propto p(Y \mid \theta) \propto \prod_{k=1}^{K} \theta_k^{y_k}$

- What is $\hat{\theta}_{MLE}$?

$$\hat{\theta}_{k,MLE} = \frac{Y_k}{\sum y_k}$$

Dirichlet Distribution

Generalizes Beta, conjugate for multinomial.

$$\vec{\theta}_K \sim Dir(\alpha_1, \ldots \alpha_K)$$

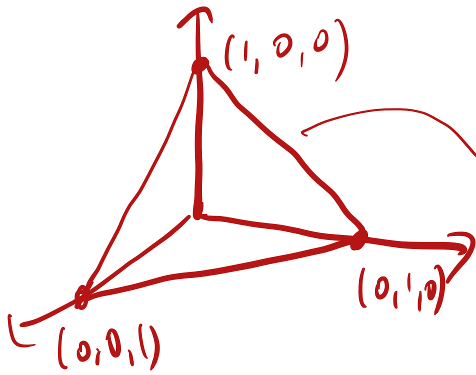
$$E[\theta_k] = \frac{\alpha_k}{\sum_{i=1}^{K} \alpha_i}$$

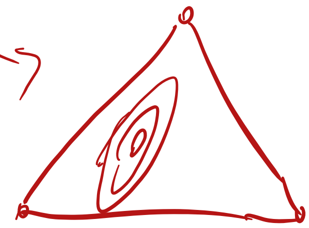
$$P(\theta \mid y) \propto \prod_k \theta_k^{(y_k + \alpha_k - 1)} \longrightarrow Dir(y_1 + \alpha_1, \ldots y_k + \alpha_k)$$

$K = 3$

$(\theta_1, \theta_2, \theta_3)$

$\sum \theta = 1$





$Z_i \sim \mathrm{Gam}(\alpha_i, 1)$

$$\left(\frac{Z_1}{\sum_{k=1}^{K} Z_k}, \ \ldots \ \frac{Z_K}{\sum Z_k} \right) \sim \mathrm{Dir}(\alpha_1, \ldots \alpha_K)$$