

Variational Methods for Latent Variable Problems

Ryan Giordano (for Johns Hopkins Biostats BLAST working group)

Oct, 2021

Massachusetts Institute of Technology

Outline

Outline for today:

- Some examples of latent variable models
- A template: The Neyman-Scott “paradox” and marginalization
- Bayesian versus frequentist approaches to marginalization
- The classical EM algorithm (in brief)

Next week, we will build on these ideas to present more general variational inference.

Randomized controlled trials were run in seven different countries to measure the effect of access to microcredit on business profits. In each country, thousands of businesses were observed. These businesses share common, unobserved attributes of their particular country. [Meager, 2020]

The different levels of profit and microcredit effectiveness in each country are latent variables. We wish to infer the overall average effectiveness of microcredit, which is common to all observations.



Latent variable models: Mouse geonomics

A set of mice were infected with an influenza virus, and the expression level for a large number of genes were measured over time. We wish to cluster together genes that have similarly shaped expression time series. [Luan and Li, 2003]

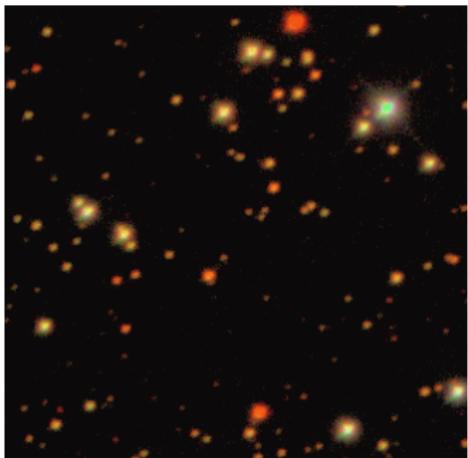
The cluster identities (archetypical time series of expression levels) are common to all observations. Which cluster a particular gene belongs to is a latent variable.

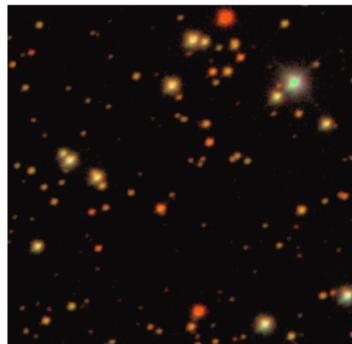


Latent variable models: Astronomical catalogs

The Sloan Digital Sky Survey systematically photographed the night sky from the earth's surface. Astronomers wish to create a catalog of stars and galaxies and their properties that can be searched through and analyzed statistically, e.g. for evidence of dark matter. [Regier et al., 2019]

Each individual image contains distortion from that particular night's atmosphere and telescope configuration. The shape and identity of the astronomical objects are latent variables, and the distortion is common to all astronomical objects in a particular image. The typical shape and variability of the distortion is common to all images.





Each of these models exhibits

- High dimensional “local” latent structure
- Low-dimensional “global” parameters of primary interest
- Possibly complicated dependence between the two (knowledge of the local variables informs the value of the globals, and vice-versa)

The Neyman-Scott “paradox”



...or “Why don’t we always use the maximum likelihood estimator?”

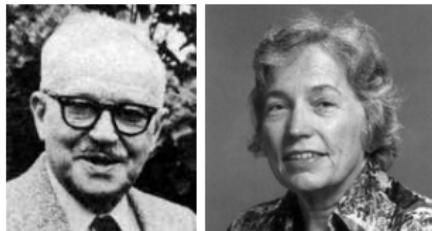
The “paradox” is built around a toy model. For some unknown z_n and θ , draw

$$y_{na}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

$$y_{nb}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

Draw N such pairs, each with its own z_n . The task is to infer θ .

The Neyman-Scott “paradox”



...or “Why don’t we always use the maximum likelihood estimator?”

The “paradox” is built around a toy model. For some unknown z_n and θ , draw

$$y_{na}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

$$y_{nb}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

Draw N such pairs, each with its own z_n . The task is to infer θ .

Observations: $y = (y_{11}, y_{1b}, \dots, y_{Na}, y_{Nb})$

Unknown latent variables: $z = (z_1, \dots, y_N)$

Unknown global parameter: $\theta \in \mathbb{R}$

The Neyman-Scott “paradox”

$$y_{na}|z_n, \theta \sim \mathcal{N}(z_n, \theta) \quad y_{nb}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

Let's use that old workhorse, the maximum likelihood estimator (MLE)!

(**Spoiler:** Something will go wrong.)

The Neyman-Scott “paradox”

$$y_{na}|z_n, \theta \sim \mathcal{N}(z_n, \theta) \quad y_{nb}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

Let's use that old workhorse, the maximum likelihood estimator (MLE)!

(**Spoiler:** Something will go wrong.)

The normal distribution gives (up to constants):

$$\log p(y_{na}, y_{nb}|\theta, z_n) = -\frac{1}{2}\theta^{-1}(y_{na} - z_n)^2 - \frac{1}{2}\log\theta - \frac{1}{2}\theta^{-1}(y_{nb} - z_n)^2 - \frac{1}{2}\log\theta$$

The Neyman-Scott “paradox”

$$y_{na}|z_n, \theta \sim \mathcal{N}(z_n, \theta) \quad y_{nb}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

Let's use that old workhorse, the maximum likelihood estimator (MLE)!

(**Spoiler:** Something will go wrong.)

The normal distribution gives (up to constants):

$$\log p(y_{na}, y_{nb}|\theta, z_n) = -\frac{1}{2}\theta^{-1}(y_{na} - z_n)^2 - \frac{1}{2}\log\theta - \frac{1}{2}\theta^{-1}(y_{nb} - z_n)^2 - \frac{1}{2}\log\theta$$

So the log likelihood is given by

$$\log p(y|\theta, z) = \sum_{n=1}^N \log p(y_{na}, y_{nb}|\theta, z_n)$$

The MLE is given by:

$$\hat{\theta}, \hat{z} := \underset{\theta, z}{\operatorname{argmax}} \log p(y|\theta, z) \quad \Leftrightarrow \quad \left. \frac{\partial \log p(y|\theta, z)}{\partial(\theta, z)} \right|_{\hat{\theta}, \hat{z}} = 0$$

The Neyman-Scott “paradox”

$$y_{na}|z_n, \theta \sim \mathcal{N}(z_n, \theta) \quad y_{nb}|z_n, \theta \sim \mathcal{N}(z_n, \theta)$$

Let's use that old workhorse, the maximum likelihood estimator (MLE)!

(Spoiler: Something will go wrong.)

The normal distribution gives (up to constants):

$$\log p(y_{na}, y_{nb}|\theta, z_n) = -\frac{1}{2}\theta^{-1}(y_{na} - z_n)^2 - \frac{1}{2}\log\theta - \frac{1}{2}\theta^{-1}(y_{nb} - z_n)^2 - \frac{1}{2}\log\theta$$

So the log likelihood is given by

$$\log p(y|\theta, z) = \sum_{n=1}^N \log p(y_{na}, y_{nb}|\theta, z_n)$$

The MLE is given by:

$$\hat{\theta}, \hat{z} := \underset{\theta, z}{\operatorname{argmax}} \log p(y|\theta, z) \quad \Leftrightarrow \quad \left. \frac{\partial \log p(y|\theta, z)}{\partial(\theta, z)} \right|_{\hat{\theta}, \hat{z}} = 0$$

Exercise: Find an expression for \hat{z}_n .

The Neyman-Scott “paradox”

Exercise: Find an expression for \hat{z}_n .

$$\begin{aligned} 0 &= \frac{\partial \log p(y|\theta, z)}{\partial z_n} \bigg|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial \log p(y_{na}, y_{nb}|\theta, z_n)}{\partial z_n} \bigg|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial}{\partial z_n} \left(-\frac{1}{2}\theta^{-1}(y_{na} - z_n)^2 - \frac{1}{2}\log\theta - \frac{1}{2}\theta^{-1}(y_{nb} - z_n)^2 - \frac{1}{2}\log\theta \right) \bigg|_{\hat{z}, \hat{\theta}} \\ &= -\hat{\theta}^{-1}(y_{na} - \hat{z}_n) - \hat{\theta}^{-1}(y_{nb} - \hat{z}_n) \Rightarrow \\ \hat{z}_n &= \frac{1}{2}(y_{na} + y_{nb}). \end{aligned}$$

The Neyman-Scott “paradox”

Exercise: Find an expression for \hat{z}_n .

$$\begin{aligned} 0 &= \frac{\partial \log p(y|\theta, z)}{\partial z_n} \bigg|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial \log p(y_{na}, y_{nb}|\theta, z_n)}{\partial z_n} \bigg|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial}{\partial z_n} \left(-\frac{1}{2}\theta^{-1}(y_{na} - z_n)^2 - \frac{1}{2}\log\theta - \frac{1}{2}\theta^{-1}(y_{nb} - z_n)^2 - \frac{1}{2}\log\theta \right) \bigg|_{\hat{z}, \hat{\theta}} \\ &= -\hat{\theta}^{-1}(y_{na} - \hat{z}_n) - \hat{\theta}^{-1}(y_{nb} - \hat{z}_n) \Rightarrow \\ \hat{z}_n &= \frac{1}{2}(y_{na} + y_{nb}). \end{aligned}$$

Wonderful! This is a very sensible expression, and it doesn't depend on $\hat{\theta}$.

The Neyman-Scott “paradox”

Exercise: Find an expression for \hat{z}_n .

$$\begin{aligned} 0 &= \frac{\partial \log p(y|\theta, z)}{\partial z_n} \bigg|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial \log p(y_{na}, y_{nb}|\theta, z_n)}{\partial z_n} \bigg|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial}{\partial z_n} \left(-\frac{1}{2}\theta^{-1}(y_{na} - z_n)^2 - \frac{1}{2}\log\theta - \frac{1}{2}\theta^{-1}(y_{nb} - z_n)^2 - \frac{1}{2}\log\theta \right) \bigg|_{\hat{z}, \hat{\theta}} \\ &= -\hat{\theta}^{-1}(y_{na} - \hat{z}_n) - \hat{\theta}^{-1}(y_{nb} - \hat{z}_n) \Rightarrow \\ \hat{z}_n &= \frac{1}{2}(y_{na} + y_{nb}). \end{aligned}$$

Wonderful! This is a very sensible expression, and it doesn't depend on $\hat{\theta}$.

Exercise: Using this result, find an expression for $\hat{\theta}$.

Hint: $(y_{na} - \hat{z}_n)^2 = (y_{nb} - \hat{z}_n)^2 = \frac{1}{4}(y_{na} - y_{nb})^2$

The Neyman-Scott “paradox”

Exercise: Find an expression for $\hat{\theta}$.

Hint: $(y_{na} - \hat{z}_n)^2 = (y_{nb} - \hat{z}_n)^2 = \frac{1}{4} (y_{na} - y_{nb})^2$

$$\begin{aligned} 0 &= \frac{\partial \log p(y|\theta, z)}{\partial \theta} \Big|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial}{\partial \theta} \sum_{n=1}^N \left(-\frac{1}{2} \theta^{-1} (y_{na} - z_n)^2 - \frac{1}{2} \log \theta - \frac{1}{2} \theta^{-1} (y_{nb} - z_n)^2 - \frac{1}{2} \log \theta \right) \Big|_{\hat{z}, \hat{\theta}} \\ &= \sum_{n=1}^N \left(\frac{1}{2} \hat{\theta}^{-2} \frac{1}{4} (y_{na} - y_{nb})^2 - \frac{1}{2} \hat{\theta}^{-1} + \frac{1}{2} \hat{\theta}^{-2} \frac{1}{4} (y_{na} - y_{nb})^2 - \frac{1}{2} \hat{\theta}^{-1} \right) \\ &= \hat{\theta}^{-2} \frac{1}{4} \sum_{n=1}^N (y_{na} - y_{nb})^2 - N \hat{\theta}^{-1} \Rightarrow \\ \hat{\theta} &= \frac{1}{4} \frac{1}{N} \sum_{n=1}^N (y_{na} - y_{nb})^2. \end{aligned}$$

The Neyman-Scott “paradox”

Exercise: Find an expression for $\hat{\theta}$.

Hint: $(y_{na} - \hat{z}_n)^2 = (y_{nb} - \hat{z}_n)^2 = \frac{1}{4} (y_{na} - y_{nb})^2$

$$\begin{aligned} 0 &= \frac{\partial \log p(y|\theta, z)}{\partial \theta} \Big|_{\hat{z}, \hat{\theta}} \\ &= \frac{\partial}{\partial \theta} \sum_{n=1}^N \left(-\frac{1}{2} \theta^{-1} (y_{na} - z_n)^2 - \frac{1}{2} \log \theta - \frac{1}{2} \theta^{-1} (y_{nb} - z_n)^2 - \frac{1}{2} \log \theta \right) \Big|_{\hat{z}, \hat{\theta}} \\ &= \sum_{n=1}^N \left(\frac{1}{2} \hat{\theta}^{-2} \frac{1}{4} (y_{na} - y_{nb})^2 - \frac{1}{2} \hat{\theta}^{-1} + \frac{1}{2} \hat{\theta}^{-2} \frac{1}{4} (y_{na} - y_{nb})^2 - \frac{1}{2} \hat{\theta}^{-1} \right) \\ &= \hat{\theta}^{-2} \frac{1}{4} \sum_{n=1}^N (y_{na} - y_{nb})^2 - N \hat{\theta}^{-1} \Rightarrow \\ \hat{\theta} &= \frac{1}{4} \frac{1}{N} \sum_{n=1}^N (y_{na} - y_{nb})^2. \end{aligned}$$

Exercise: Suppose the true parameters are θ_0 and z_0 . What is $\lim_{N \rightarrow \infty} \hat{\theta}$?

Hint: Use the law of large numbers.

The Neyman-Scott “paradox”

Exercise: What is the behavior of $\hat{\theta}$ for large N ? By the law of large numbers,

$$\begin{aligned}\hat{\theta} &= \frac{1}{4} \frac{1}{N} \sum_{n=1}^N (y_{na} - y_{nb})^2 \\ &\xrightarrow[N \rightarrow \infty]{\text{prob}} \frac{1}{4} \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - y_{nb})^2] \\ &= \frac{1}{4} \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n} - (y_{nb} - z_{0n}))^2] \\ &= \frac{1}{4} \left(\mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n})^2] + \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{nb} - z_{0n})^2] + \right. \\ &\quad \left. 2 \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n})(y_{nb} - z_{0n})] \right) \\ &= \frac{1}{4} (\theta_0 + \theta_0 + 0) \\ &= \frac{\theta_0}{2} \neq \theta_0.\end{aligned}$$

The Neyman-Scott “paradox”

Exercise: What is the behavior of $\hat{\theta}$ for large N ? By the law of large numbers,

$$\begin{aligned}\hat{\theta} &= \frac{1}{4} \frac{1}{N} \sum_{n=1}^N (y_{na} - y_{nb})^2 \\ &\xrightarrow[N \rightarrow \infty]{\text{prob}} \frac{1}{4} \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - y_{nb})^2] \\ &= \frac{1}{4} \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n} - (y_{nb} - z_{0n}))^2] \\ &= \frac{1}{4} \left(\mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n})^2] + \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{nb} - z_{0n})^2] + \right. \\ &\quad \left. 2 \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n})(y_{nb} - z_{0n})] \right) \\ &= \frac{1}{4} (\theta_0 + \theta_0 + 0) \\ &= \frac{\theta_0}{2} \neq \theta_0.\end{aligned}$$



The Neyman-Scott “paradox”

Exercise: What is the behavior of $\hat{\theta}$ for large N ? By the law of large numbers,

$$\begin{aligned}\hat{\theta} &= \frac{1}{4} \frac{1}{N} \sum_{n=1}^N (y_{na} - y_{nb})^2 \\ &\xrightarrow[N \rightarrow \infty]{\text{prob}} \frac{1}{4} \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - y_{nb})^2] \\ &= \frac{1}{4} \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n} - (y_{nb} - z_{0n}))^2] \\ &= \frac{1}{4} \left(\mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n})^2] + \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{nb} - z_{0n})^2] + \right. \\ &\quad \left. 2 \mathbb{E}_{p(y|\theta_0, z_0)} [(y_{na} - z_{0n})(y_{nb} - z_{0n})] \right) \\ &= \frac{1}{4} (\theta_0 + \theta_0 + 0) \\ &= \frac{\theta_0}{2} \neq \theta_0.\end{aligned}$$



⇒ The MLE is inconsistent. What went wrong?

The Neyman-Scott “paradox”

The MLE is inconsistent. What went wrong?

$$\begin{aligned} \hat{z}_n &= \frac{1}{2}(y_{na} + y_{nb}) \\ \hat{\theta} &\xrightarrow[N \rightarrow \infty]{prob} \frac{\theta_0}{2} \neq \theta_0 \end{aligned}$$

The Neyman-Scott “paradox”

The MLE is inconsistent. What went wrong?

$$\hat{z}_n = \frac{1}{2}(y_{na} + y_{nb})$$
$$\hat{\theta} \xrightarrow[N \rightarrow \infty]{\text{prob}} \frac{\theta_0}{2} \neq \theta_0$$

- Our estimates for the latent variables \hat{z}_n are quite uncertain (they use only two observations each).
- But our MLE estimate for $\hat{\theta}$ treated the \hat{z}_n as if they were known exactly. I.e., our MLE estimator assumed less dispersion around \hat{z}_n than was truly present.
- Consequently, we *under-estimated* the dispersion θ_0 .
- To avoid this problem, we must *account for the uncertainty* in z_n when estimating θ .

Solution: Marginalize: Integrate out the uncertainty in the z_n .

The Neyman-Scott “paradox”

To marginalize we must:

- Add a distributional assumption $z|\theta \sim p(z|\theta)$.
- Compute the marginal $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$.
- Compute the marginal MLE $\hat{\theta} = \operatorname{argmax}_{\theta} p(y|\theta)$
 - (Contrast with $\hat{\theta}, \hat{z} = \operatorname{argmax}_{\theta, z} p(y|\theta, z)$)

The Neyman-Scott “paradox”

To marginalize we must:

- Add a distributional assumption $z|\theta \sim p(z|\theta)$.
- Compute the marginal $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$.
- Compute the marginal MLE $\hat{\theta} = \operatorname{argmax}_{\theta} p(y|\theta)$
 - (Contrast with $\hat{\theta}, \hat{z} = \operatorname{argmax}_{\theta, z} p(y|\theta, z)$)

Neyman-Scott resolved

Let's let $z_n \sim \mathcal{N}(0, 1)$. Then, by standard properties of the normal,

$$\begin{pmatrix} y_{na} \\ y_{nb} \end{pmatrix} \stackrel{iid}{\sim} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 + \theta & 1 \\ 1 & 1 + \theta \end{pmatrix} \right)$$

Sample covariances of the bivariate normal are consistent, so the marginal MLE

$$\hat{\theta} := \operatorname{argmax}_{\theta} \sum_{n=1}^N \log \int p(y_{na}, y_{nb}|\theta, z_n)p(z_n)dz_n$$

is consistent.

Marginalization: General setup

In general notation, we want to infer θ from

Observations: $y = (y_1, \dots, y_N)$

Unknown latent variables: $z = (z_1, \dots, z_N)$

Unknown global parameter: $\theta \in \mathbb{R}^D$

We have learned that

Bad: $\hat{\theta}, \hat{z} = \operatorname{argmax}_{\theta, z} \log p(y|\theta, z)$

Good: $\hat{\theta} = \operatorname{argmax}_{\theta} \log \int p(y|\theta, z)p(z|\theta)dz.$

Marginalization: General setup

In general notation, we want to infer θ from

Observations: $y = (y_1, \dots, y_N)$

Unknown latent variables: $z = (z_1, \dots, z_N)$

Unknown global parameter: $\theta \in \mathbb{R}^D$

We have learned that

Bad: $\hat{\theta}, \hat{z} = \operatorname{argmax}_{\theta, z} \log p(y|\theta, z)$

Good: $\hat{\theta} = \operatorname{argmax}_{\theta} \log \int p(y|\theta, z)p(z|\theta)dz.$

There are two problems:

- Need to posit $p(z|\theta)$
- Need to compute $\int p(y|\theta, z)p(z|\theta)dz$

We will only deal with the second problem in these two talks, assuming we have a $p(z|\theta)$ we are willing to live with.

In general, the integral is hard!

Bayesian statistics has marginalization built in

Link to optional blog post on frequentist vs Bayesian statistics

Bayesian statistics has marginalization built in

Recall that a Bayesian model posits a full generative process:

$$\theta \sim p(\theta) \quad z|\theta \sim p(z|\theta) \quad y|z, \theta \sim p(y|z, \theta)$$

and forms the posterior

$$p(\theta, z|y) = \frac{p(y|\theta, z)p(z|\theta)p(\theta)}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'}$$

Bayesian statistics has marginalization built in

Recall that a Bayesian model posits a full generative process:

$$\theta \sim p(\theta) \quad z|\theta \sim p(z|\theta) \quad y|z, \theta \sim p(y|z, \theta)$$

and forms the posterior

$$\begin{aligned} p(\theta, z|y) &= \frac{p(y|\theta, z)p(z|\theta)p(\theta)}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'} \\ \Rightarrow p(\theta|y) &= \int p(\theta, z|y)dz = \frac{\int p(y|\theta, z)p(z|\theta)p(\theta) dz}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'} \\ &= \frac{(\int p(y|\theta, z)p(z|\theta)dz) p(\theta)}{\int (\int p(y|\theta', z')p(z'|\theta')dz') p(\theta')d\theta'} \\ &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}. \end{aligned}$$

Bayesian statistics has marginalization built in

Recall that a Bayesian model posits a full generative process:

$$\theta \sim p(\theta) \quad z|\theta \sim p(z|\theta) \quad y|z, \theta \sim p(y|z, \theta)$$

and forms the posterior

$$\begin{aligned} p(\theta, z|y) &= \frac{p(y|\theta, z)p(z|\theta)p(\theta)}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'} \\ \Rightarrow p(\theta|y) &= \int p(\theta, z|y)dz = \frac{\int p(y|\theta, z)p(z|\theta)p(\theta) dz}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'} \\ &= \frac{(\int p(y|\theta, z)p(z|\theta)dz) p(\theta)}{\int (\int p(y|\theta', z')p(z'|\theta')dz') p(\theta')d\theta'} \\ &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}. \end{aligned}$$

⇒ Bayesian methods do not suffer from the Neyman-Scott problem.

- Bayesians are forced to posit $p(z|\theta)$
- Forming the posterior is equivalent to using the marginal $p(y|\theta)$

Bayesian statistics has marginalization built in

Recall that a Bayesian model posits a full generative process:

$$\theta \sim p(\theta) \quad z|\theta \sim p(z|\theta) \quad y|z, \theta \sim p(y|z, \theta)$$

and forms the posterior

$$\begin{aligned} p(\theta, z|y) &= \frac{p(y|\theta, z)p(z|\theta)p(\theta)}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'} \\ \Rightarrow p(\theta|y) &= \int p(\theta, z|y)dz = \frac{\int p(y|\theta, z)p(z|\theta)p(\theta) dz}{\int \int p(y|\theta', z')p(z'|\theta')p(\theta')d\theta' dz'} \\ &= \frac{(\int p(y|\theta, z)p(z|\theta)dz) p(\theta)}{\int (\int p(y|\theta', z')p(z'|\theta')dz') p(\theta')d\theta'} \\ &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}. \end{aligned}$$

⇒ Bayesian methods do not suffer from the Neyman-Scott problem.

- Bayesians are forced to posit $p(z|\theta)$
- Forming the posterior is equivalent to using the marginal $p(y|\theta)$

But the integral is still hard! Full Bayesian solutions typically require Markov Chain Monte Carlo, which is slow and sampling based.

Are there faster alternatives, based on optimization? Hint: Yes.

Marginalization: The EM algorithm

One “frequentist” method for optimizing the marginal likelihood is the famous expectation-maximization (EM) algorithm.

The EM algorithm works / is useful when:

- The joint log probability $\log p(y|\theta, z) + \log p(z|\theta)$ is easy to write down
- The posterior $p(z|y, \theta)$ is easy to compute
- The marginalizing integral $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$ is hard

Marginalization: The EM algorithm

One “frequentist” method for optimizing the marginal likelihood is the famous expectation-maximization (EM) algorithm.

The EM algorithm works / is useful when:

- The joint log probability $\log p(y|\theta, z) + \log p(z|\theta)$ is easy to write down
- The posterior $p(z|y, \theta)$ is easy to compute
- The marginalizing integral $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$ is hard

The EM algorithm alternates between two steps. Starting at an iterate $\hat{\theta}_{(i)}$, repeat until convergence:

The E-step: Compute $Q_{(i)}(\theta) := \mathbb{E}_{p(z|y, \hat{\theta}_{(i)})} [\log p(y|\theta, z) + \log p(z|\theta)]$

The M-step: Compute the next iterate $\hat{\theta}_{(i+1)} := \operatorname{argmax}_{\theta} Q_{(i)}(\theta)$

Marginalization: The EM algorithm

One “frequentist” method for optimizing the marginal likelihood is the famous expectation-maximization (EM) algorithm.

The EM algorithm works / is useful when:

- The joint log probability $\log p(y|\theta, z) + \log p(z|\theta)$ is easy to write down
- The posterior $p(z|y, \theta)$ is easy to compute
- The marginalizing integral $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$ is hard

The EM algorithm alternates between two steps. Starting at an iterate $\hat{\theta}_{(i)}$, repeat until convergence:

The E-step: Compute $Q_{(i)}(\theta) := \mathbb{E}_{p(z|y, \hat{\theta}_{(i)})} [\log p(y|\theta, z) + \log p(z|\theta)]$

The M-step: Compute the next iterate $\hat{\theta}_{(i+1)} := \operatorname{argmax}_{\theta} Q_{(i)}(\theta)$

Note that everything in the E- and M-steps are “easy.” Nevertheless, the iterates $\hat{\theta}_{(i)}$ converge to a (possibly local) optimum of the marginal log likelihood $\log p(y|\theta)$.

Marginalization: The EM algorithm

One “frequentist” method for optimizing the marginal likelihood is the famous expectation-maximization (EM) algorithm.

The EM algorithm works / is useful when:

- The joint log probability $\log p(y|\theta, z) + \log p(z|\theta)$ is easy to write down
- The posterior $p(z|y, \theta)$ is easy to compute
- The marginalizing integral $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$ is hard

The EM algorithm alternates between two steps. Starting at an iterate $\hat{\theta}_{(i)}$, repeat until convergence:

The E-step: Compute $Q_{(i)}(\theta) := \mathbb{E}_{p(z|y, \hat{\theta}_{(i)})} [\log p(y|\theta, z) + \log p(z|\theta)]$

The M-step: Compute the next iterate $\hat{\theta}_{(i+1)} := \operatorname{argmax}_{\theta} Q_{(i)}(\theta)$

Note that everything in the E- and M-steps are “easy.” Nevertheless, the iterates $\hat{\theta}_{(i)}$ converge to a (possibly local) optimum of the marginal log likelihood $\log p(y|\theta)$.

Exercise: Prove that the EM algorithm gives a consistent estimator for the Neyman-Scott paradox.

Next week: Actual variational inference

Next week we will:

- Prove that the EM algorithm works (in a non-standard way)
- Identify a lie I told on the last slide
- Generalize to cases when the posterior $p(z|y, \theta)$ is difficult to compute

This last point will finally bring us to the set of techniques commonly called "variational inference."

Conclusions

Y. Luan and H. Li. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482, 2003.

R. Meager. Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. *LSE working paper*, 2020.

Jeffrey Regier, Keno Fischer, Kiran Pamnany, Andreas Noack, Jarrett Revels, Maximilian Lam, Steve Howard, Ryan Giordano, David Schlegel, Jon McAuliffe, et al. Cataloging the visible universe through bayesian inference in julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104, 2019.