Chris Conlon

NYU Stern

March 1, 2019

Quick Refresh: Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Given a positive test result what is the probability a patient actually has cancer?

Table: Test Accuracy

	Cancer (1%)	No Cancer (99%)
Positive Test	80%	9.6%
Negative Test	20%	90.4%

Quick Refresh: Bayes Rule

Caclulate $Pr(Cancer \& PositiveTest)$ and $Pr(NoCancer \& PositiveTest)$

Table: Joint Probabilities

	Cancer (1%)	No Cancer (99%)
Positive Test	(0.8)(0.01)=0.008	(0.9)(0.096)=0.09504
Negative Test	(0.2)(0.01)=0.002	(0.9)(0.904)=0.89496

$$Pr(Cancer|PositiveTest) = \frac{Pr(Cancer, PosTest)}{Pr(Cancer, PosTest) + Pr(NoCancer, PosTest)} = .008/.10304 = 0.0776$$

Introduction

- Suppose that we toss a coin several times with $x_i \in \{H, T\} = \{1, 0\}$
- $\mathbf{X} = \{H, T, H, H, \ldots\}$.
- Suppose that the probability of heads $Pr(x_i = H) = p$.
- What is the likelihood of an observed sequence of $\mathbf{X}$? where x_i are I.I.D.

$$Pr(x_i|p) = p^{x_i}(1-p)^{1-x_i}$$
$$Pr(\mathbf{X}|p) = p^{\sum_i x_i}(1-p)^{\sum_i (1-x_i)}$$

Introduction: MLE for coin toss

Can construct the log likelihood and find the MLE.

$$\ell(\mathbf{X}|p) = (\sum_i x_i) \ln p + (N - \sum_i x_i) \ln(1-p)$$

$$\frac{\partial \ell(p)}{\partial p} = (\sum_i x_i)\frac{1}{p} - (N - \sum_i x_i)\frac{1}{1-p} = 0$$

$$\frac{1-p}{p} = \frac{(\frac{1}{N} \cdot N - \frac{1}{N} \cdot \sum_i x_i)}{\frac{1}{N} \cdot \sum_i x_i} \rightarrow \hat{p} = \frac{1}{N} \cdot \sum_i x_i$$

Can also construct the properties of $\hat{p}$.

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{N} \cdot \sum_i x_i\right] = \left[\frac{1}{N} \cdot \sum_i \mathbb{E}x_i\right] = \mu_x = p_0$$

$$\mathbb{V}[\hat{p}|\mathbf{X}] = \mathbb{V}\left[\frac{1}{N} \cdot \sum_i x_i\right] = \frac{1}{N^2} \cdot \sum_i \mathbb{V}(x_i) = \frac{N}{N^2} p(1-p)$$

Which gives us a CI of: $\left(\overline{x} \pm 1.96 \cdot \sqrt{\frac{1}{N}\overline{x}(1-\overline{x})}\right)$

Interpretation of Bayesian methods

Basic idea is as follows:

- You observe some data X which has probability $P(x|\theta)$ under parameters $\theta \in \Theta$.
- There is an (assumed) prior distribution $P(\theta)$.
- The object of interest is the posterior distribution $P(\theta|X)$ and/or functionals of $P(\cdot|X)$ such as the mean.
- The posterior distribution is sufficient for any counterfactual you'd want to run:
  - Draw random samples from $P(\cdot|X)$, and simulate your counterfactual.
  - This gives you a distribution of counterfactual outcomes.
  - Report 5th and 95th percentiles.
  - "Inference on counterfactuals is for free."

The same application:

- Start with a (diffuse) initial guess for the distribution of p: $f_P(p)$.
- Incorporate information from likelihood: $f(x_i|p)$
- Construct posterior density estimate $f(p|x_i)$.
  - This doesn't characterize a best estimate $\hat{p}$ but a full distribution.
  - We can calculate $\mathbb{E}[p|x_i]$ or $\mathbb{V}[p|x_i]$ or any other functions of the posterior density.
- Challenge: How to choose initial $f_P(p)$.

One possible guess is the uniform distribution $f(z) = 1$ on $0 \leq z \leq 1$.

- Marginal/Prior Distribution: $f_P(p) = 1$ for $0 \leq p \leq 1$.
- Conditional Distribution/Likelihood: $f_{X|P}(x|p) = p^x(1-p)^{1-p}$
- Joint Distribution :
  $f_{X,P}(x,p) = f_{X|P}(x|p) \cdot f_P(p) = p^x(1-p)^{1-p} \cdot 1 = x \cdot p + (1-x) \cdot (1-p)$
  - This is only defined for $p \in [0,1]$ and $x \in \{0,1\}$. It is zero elsewhere.
- What about Marginal Distribution for x?

$$\begin{aligned}\int_p f_{PX}(x,p)dp &= x \cdot \int_0^1 pdp + (1-x) \cdot \int_0^1 (1-p)dp \\ &= x \cdot \frac{1}{2} + (1-x) \cdot \frac{1}{2} = \frac{1}{2} \propto 1\end{aligned}$$

The object we are usually interested in is the Posterior Distribution

$$f_{P|X}(p|x) = \frac{f_{X|P}(x|p) \cdot f_P(p)}{\int_0^1 f_{X|P}(x|p) \cdot f_P(p) dp} = 2p^x(1-p)^{1-x} \propto p^x(1-p)^{1-x}$$

- We are back at the p.m.f. of the Bernoulli which is maybe comforting.
- This is true because $f_X(x) \propto 1$ and $f_P(p) \propto 1$.
- $f_{P|X}(p|x=0) = (1-p)$ and $f_{P|X}(p|x=1) = p$.

- Let's try a different prior distribution than the uniform we used last time. This time we will use a $Beta(\alpha, \beta)$ distribution:

$$f_P(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$
$$\mathbb{E}[p|\alpha,\beta] = \frac{\alpha}{\alpha+\beta}$$
$$\mathbb{V}[p|\alpha,\beta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- This has the advantage that it plays nicely with the Binomial.
- Consider $\alpha = 16, \beta = 8$. This gives $\mathbb{E}[p] = \frac{2}{3}$ and $\mathbb{SE}[p] = 0.094$.

Consider the case where $x = 1$ (we get one piece of new data).

$$f_P(p) \cdot f_{X|P}(x|p) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}}_{c(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \cdot p \propto p^{\alpha}(1-p)^{\beta-1}$$

- The resulting distribution is now $(p|x = 1) \sim Beta(\alpha + 1, \beta)$.
- Our posterior has mean = 0.68 and SE = 0.091.
- Estimate of mean increases and SE decreases.
- Likewise if $x = 0$ we get $(p|x = 0) \sim Beta(\alpha, \beta + 1)$
- There is a conjugacy relationship between the Beta and the Binomial.

General Case

$$\overbrace{f_{\theta|X}(\theta|x)}^{\text{posterior}} = \frac{\overbrace{f_{X,\theta}(x,\theta)}^{\text{joint}}}{\underbrace{f_X(x)}_{\text{marginal of } x}} = \frac{\overbrace{f_{X|\theta}(x|\theta)}^{\text{likelihood}} \cdot \overbrace{f_\theta(\theta)}^{\text{prior}}}{\int f_{X|\theta}(x|\theta) \cdot f_\theta(\theta) d\theta}$$

There is a shortcut because the denominator doesn't depend on θ

$$f_{\theta|X}(\theta|x) \propto f_{X|\theta}(x|\theta) \cdot f_\theta(\theta) = \mathcal{L}(\theta|x) \cdot f_\theta(\theta)$$

We can cheat because there exists a constant c so that $c \int \mathcal{L}(\theta|x) \cdot f_\theta(\theta) d\theta = 1$.

A Normal Example

Assume $X \sim N(\mu, 1)$ and $\mu \sim N(0, 100)$. What is $f_{\mu|X}(\mu|X = x)$?

$$\begin{aligned} f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2 \cdot 100}\mu^2\right) \\ &= \exp -\frac{1}{2}(x^2 - 2x\mu + \mu^2 + \mu^2/100) \\ &\propto \exp\left(-\frac{1}{2(100/101)}(\mu - (100/101)x)^2\right) \end{aligned}$$

It happens that $(u|x) \sim N(100x/101, 100/101)$.
In general we can't expect a closed form posterior.

Generalization to Multiple Observations

This is straighforward:

$$p(\theta|X_1, \ldots, X_N) \propto \mathcal{L}(\theta|X_1, \ldots, X_N) \cdot p(\theta)$$

- Still depends on: prior, likelihood to construct posterior.
- Can update one observation at a time or all at once.

Bayesian Vs. Frequentist methods

- Once you've got estimates, you've got posterior probability intervals without additional work.
- But Bayesian objects of interest are different from confidence intervals:
  - NOT a sampling experiment.
- There is also a frequentist interpretation:
  - In large samples, in identified models: means, modes, medians of Bayesian posteriors converge to MLE estimates.
- An asymptotic result, the Bernstein-Von Mises Theorem, states that if you have enough data, under some (weak) technical conditions the posterior of a finite-dimensional parameter is asymptotically normal and independent of the prior. (See Van der Vaart for a general statement)

Bernstein von-Mises Theorem

A posterior distribution converges as you get more and more data to a multivariate normal distribution centred at the maximum likelihood estimator with covariance matrix given by $n^{-1}I(\theta_0)^{-1}$, where θ_0 is the true population parameter (Edit: here $I(\theta_0)$ is the Fisher information matrix at the true population parameter value).

Under these conditions (and some more):

1. MLE is consistent
2. Fixed number of parameters
3. θ_0 in interior of Θ (true value of SD can't = 0).
4. The prior density must be non-zero in a neighborhood of θ_0.
5. log-likelihood needs to be smooth (two derivates at the true value and more)

Uses of Bernstein-Von Mises Theorems

- Bayesian posterior mean is asymptotically equivalent to the maximum likelihood estimator.
- In practice, computations of extremum estimators might not converge, might get stuck in local optima, etc.
- It can be a lot easier to compute a mean than to find a mode of a function.
- The B-vM theorem says that "the Ns justify the means."

What is a Conjugate Prior?

For a given likelihood $f_{X|\theta}(x|\theta)$ we can chose a prior $f_\theta(\theta)$ so that the posterior is proportional to a known parametric distribution.

- This makes life easy because now the posterior has a known parametric distribution (normal, beta, gamma, etc.)
- Other than convenience, this alone doesn't tell us that our choice of $f_\theta(\theta)$ is the best prior by any metric.
- Using a non-conjugate prior is entirely defensible, just less convenient.

Beta-Binomial

Prior has hyper parameters (α, β):

$$Pr(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad E[p] = \frac{\alpha}{\alpha+\beta}$$

Likelihood

$$Pr(y = k|n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

Posterior

$$\begin{aligned} f(p|n,k,\alpha,\beta) &\propto p^{k+\alpha}(1-p)^{n-k+\beta} \\ &\sim Beta(\alpha+k, \beta+n-k) \end{aligned}$$

Prior defined on unit simplex when $\sum_{i=1}^{k} p_i = n$

$$Pr\left(p_1, \ldots, p_K; \alpha_1, \ldots, \alpha_K\right) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

Likelihood

$$L(x_1, \ldots, x_k | n, p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \times \cdots \times p_k^{x_k}$$

Posterior

$$\begin{aligned} f(p_1, \ldots, p_k | x_1, \ldots, x_k, \alpha_1, \ldots, \alpha_k) &\propto p_1^{x_1 + \alpha_1} \times \cdots \times p_k^{x_k + \alpha_k} \\ &\sim Dirichlet(\alpha_k + x_k) \end{aligned}$$

Beta-Binomial and Dirichlet-Multinomial

- Dirichlet is the Multinomial generalization of a beta
- In the Beta-Binomial we can write things as if $m = \alpha + \beta$ is the number of pseudo observations and $E[p] = \frac{\alpha}{\alpha+\beta}$.
- In the Dirichlet-Multinomial we can write things as if $m = \sum_k \alpha_k$ and $E[p_1, \ldots, p_k] = \left(\frac{\alpha_1}{\sum_{k=1}^K \alpha_k} \cdots \frac{\alpha_K}{\sum_{k=1}^K \alpha_k} \right)$
- In both cases it is as if we see m observations from before the data and n observations from the data.

Gamma-Poisson

Prior (Gamma) has α occurrences in β intervals

$$Pr(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \text{ for } \theta > 0, \alpha > 0, \beta > 0$$

Likelihood (Poisson) we observe k events in each of the n periods:

$$Pr(x_i = k|\theta) = \frac{\theta^k e^{-\theta}}{k!}$$

Posterior is Gamma

$$\theta \sim Gamma(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$$

Exponential families

- Bayesian methods make heavy use of exponential-family distributions.
  - Class of distributions that includes normal, beta, gamma, exponential distribution, ...
- Definition: If we have a sample $(y_1, \ldots, y_n)$ from an exponential family, the likelihood is proportional to:

$$p(y|\theta) \propto g(\theta)^n \exp\{\sum_{j=1}^{k} c_j \varphi_j(\theta) \overline{h}_j(y)\}, \quad \overline{h}_j(y) = \sum_{i=1}^{n} h_j(y_i).$$

Exponential-family distributions have conjugate priors

- Fact: Exponential family distributions with j parameters θ have j minimal sufficient statistics $\{h_1, \ldots, h_j\}$.
  - A minimal sufficient statistic is a sufficient statistic that can be represented as a function of any other sufficient statistic.
- Result is as follows:

$$\text{Prior (with hyperparameters } \tau\text{): } p(\theta|\tau) \propto g(\theta)^{\tau_0} \exp\{\sum_{j=1}^{k} c_j \varphi_j(\theta) \tau_j\},$$

$$\text{Posterior: } p(\theta|y) \propto g(\theta)^{\tau_0+n} \exp\{\sum_{j=1}^{k} c_j \varphi_j(\theta) \left(h_j + \tau_j\right)\},$$

where h_j is the minimal sufficient statistic.

Normal: Known Variance

Assume $x \sim N(\beta, \sigma^2)$ with σ^2 known. Prior: $\beta \sim N(\mu_0, \tau^2)$. What is $f_{\beta|X}(\beta|X = x)$?

$$
\begin{aligned}
f_{\beta|X}(\beta|x) &\propto \exp\left(-\frac{1}{2\sigma^2}(x-\beta)^2\right) \cdot \exp\left(-\frac{1}{2\cdot\tau^2}(\beta-\mu_0)^2\right) \\
&\propto \exp -\frac{1}{2}\left(\frac{x^2}{\sigma^2} - \frac{2x\beta}{\sigma^2} + \frac{\beta^2}{\sigma^2} + \frac{\beta^2}{\tau^2} - \frac{2\beta\mu_0}{\tau^2} + \frac{\mu_0^2}{\tau^2}\right) \\
&\propto \exp -\frac{1}{2}\left(\beta^2\frac{\sigma^2+\tau^2}{\tau^2\sigma^2} - \beta\frac{2x\tau^2 + 2\mu_0\sigma^2}{\tau^2\cdot\sigma^2}\right) \\
&\propto \exp -\frac{1}{2\left(1/\left(1/\tau^2 + 1/\sigma^2\right)\right)}\left(\left(\beta - \left(x/\sigma^2 + \mu_0/\tau^2\right)/\left(1/\sigma^2 + 1/\tau^2\right)\right)\right)
\end{aligned}
$$

The resulting distribution is Normal with mean and variance (precision):

$$
\mathbb{E}[\beta|X = x] = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \qquad \frac{1}{\mathbb{V}(\beta|X)} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}
$$

Example: normal, known variance

- observe a draw $x \sim N(\beta, \sigma^2)$.
  - β has conjugate prior $\beta \sim N(\mu, \tau^2)$.
  - Suppose σ^2 is known.
- With some algebra, we showed:
  - Posterior: $f_{\beta|x}(\beta|x) \sim N\left(\frac{\overline{\beta}/\tau^2 + x/\sigma^2}{1/\tau^2 + 1/\sigma^2}, \frac{1}{1/\tau^2 + 1/\sigma^2}\right)$.
    - Mean is weighted average of prior mean and data mean.
    - Weights are proportional to **precisions**.
- With multiple draws $x_i \sim N(\beta, \sigma^2)$:
  - Minimal sufficient statistics are $h_1(x) = \overline{x}$, $h_2(x) = \sum x^2$.
  - $\beta|x_1, \ldots, x_n \sim N\left(\frac{\mu/\tau^2 + \overline{x}(n/\sigma^2)}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2}\right)$.

Example: normal, known variance

Despite being a giant mess this makes sense:

$$\mathbb{E}[\beta|h(\mathbf{x}) = (\bar{x}, \sum x_i^2)] = \frac{\frac{\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \qquad \frac{1}{\mathbb{V}(\mu|X)} = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$$

- Posterior mean is a weighted average of prior mean and sample mean.
- Weights depend on precision of two samples.
- Posterior Precision is sum of precision of each sample $\frac{1}{\mathbb{V}(\cdot)}$
- Probably we want to choose a relatively uninformative prior with large τ^2.
- $\tau^2 \to \infty$ implies an improper prior distribution because it no longer integrates to one. But because of $\propto$ still mostly ok.

- We're now ready to estimate the standard linear regression model:

$$y_i = x_i\beta + \epsilon_i,\ \epsilon_i \sim iid\ N(0, \sigma^2)$$

- In matrix form:

$$y \sim N(X\beta, \sigma^2 I_n).$$

- Specify prior as

$$P(\beta, \sigma^2) = P(\sigma^2)P(\beta|\sigma^2)$$

$$P(\sigma^2) \propto (\sigma^2)^{-(v_0/2+1)} \exp\left(-\frac{v_0 s_0^2}{2\sigma^2}\right)$$

$$\beta|\sigma^2 \sim N(\overline{\beta}, \sigma^2 A^{-1}).$$

That is, $\sigma^2 \sim IG(v_0/2, v_0 s_0^2/2)$, or

$$\sigma^2 \sim \frac{v_o s_o^2}{\chi^2_{v_0}}.$$

- Can show that posterior distribution will be of same form:

$$\beta|\sigma^2, y, X \sim N(\tilde{\beta}, \sigma^2(X'X + A)^{-1}),$$
$$\sigma^2|y, X \sim \frac{v_1 s_1^2}{\chi_{v_1^2}},$$

with

$$\begin{aligned} v_1 &= v_0 + n \\ s_1^2 &= \frac{v_0 s_0^2 + n s^2}{v_0 + n} \\ \tilde{\beta} &= (X'X + A)^{-1}(X'X\hat{\beta} + A\overline{\beta}) \\ n s^2 &= (y - x\tilde{\beta})'(y - x\tilde{\beta}) + (\tilde{\beta} - \overline{\beta})'A(\tilde{\beta} - \overline{\beta}). \end{aligned}$$

Linear regression: comments

- No simulation required.
- The Bayes estimator of β is a weighted average of the least squares estimator and the prior mean.
- The bayes estimator of σ^2 is "centered" over s_1^2 which is a weighted average of the prior hyperparameter and a sample quantity.
- As $n \to \infty$ the distribution of the Bayes estimator about the posterior mean converges to the distribution of the MLE estimator about the true value.

- We can extend these results to the following multivariate regression model:

$$\begin{aligned} y_i^1 &= X_i^1 \beta_1 + \epsilon_{i1} \\ &\vdots \\ y_i^k &= X_i^k \beta_k + \epsilon_{ik} \\ \epsilon_i &\sim N(0, \Sigma). \end{aligned}$$

- In vector form:

$$y_i = \beta X_i + \epsilon_i, \ \epsilon_i \sim N(0, \Sigma).$$

- N independent observations $i = 1, \ldots, N$, k-dimensional LHS variable in each observation.

Priors

- We need to place a prior on β, Σ.
  - Write as $p(\beta, \Sigma) = p(\Sigma)p(\beta|\Sigma)$.
  - Start with independent priors $p(\beta, \Sigma) = p(\beta)p(\Sigma)$.
  - Traditional to place a prior on Σ^{-1} instead of Σ.
- Convenient conjugate prior: Wishart distribution on Σ^{-1}, or Inverse Wishart on Σ.
  - Inverse Wishart is a matrix-valued distribution with two parameters, Ψ and v.
  - Ψ is a positive-definite $K \times K$ matrix. $v \in \mathbb{R}$ satisfies $v \geq K - 1$.
  - Density evaluated at X :

$$\frac{|\Psi|^{v/2}}{2^{vp/2}\Gamma_p(v/2)}|X|^{-\frac{v+p+1}{2}}e^{-\frac{1}{2}tr(\Psi X^{-1})},$$

    where Γ_J is the multivariate gamma function.
- Omitting the constant that makes the density integrate to 1,

$$p(\beta|\overline{\beta}, A) \propto |A|^{1/2}\exp(-\frac{1}{2}(\beta - \overline{\beta})'A(\beta - \overline{\beta})).$$

Posterior: $\beta|y, X$

- Notation: $G = \Sigma^{-1}$.
- Take Cholesky decomposition: $G = CC'$.
- Premultiply linear system by C' :

$$C'u_i = C'X_i\beta + C'\epsilon_i$$
$$u_i^* = X_i^*\beta + \epsilon_i^*.$$

- We now have iid errors: $\epsilon_i^* \sim N(0, I)$.
- As before, in the Bayesian linear model:

$$\beta|u, G \sim N(\tilde{\beta}, \Sigma_\beta),$$
$$\Sigma_\beta = (X^{*'}X^* + A)^{-1}$$
$$\tilde{\beta} = \Sigma_\beta(X^{*'}u^* + A\overline{\beta}).$$

- Note: As $X^{*'}$ becomes large relative to A, this estimator approaches the GLS estimator.

Posterior: $\Sigma|y, X$

- The posterior distribution of $\Sigma|y, X$ is given by:

$$G|\beta, w \sim W(v + N, \Psi + \sum_{i=1}^{N} \epsilon_i \epsilon_i'),$$

or

$$\Sigma|\beta, w \sim IW(v + N, \Psi + \sum_{i=1}^{N} \epsilon_i \epsilon_i').$$

What is Emprical Bayes?

- Priors can be an important modeling choice
- But what makes a good prior?
  - Sufficiently diffuse
  - As non-informative as possible
  - Don't tip the scales
  - Don't rule out the truth
- Idea: can we use the data itself to construct a prior?
  - If everything is a function of data, are we back in frequentist paradigm?
  - Can we get benefits of Bayes estimation without unpalatable assumptions?

A (famous) Baseball Example

Suppose we want to estimate batting averages (AVG) for some baseball players

- $AVG = \frac{\#\text{hits}}{\#AtBats}$
- Use data on the first $n = 45$ at bats and hits x_i for the 1970 season.
- Predict the batting average μ_i for the end of the season ($n = 400 - 500$ at bats).
- Obvious estimate is batting average after 45 at bats: $\widehat{\mu}_i^{MLE} = x_i/45$.
- Is there a better estimate?

A Baseball Example

Table 1.1: Batting averages $z_i = \hat{\mu}_i^{(\mathrm{MLE})}$ for 18 major league players early in the 1970 season; μ_i values are averages over the remainder of the season. The James–Stein estimates $\hat{\mu}_i^{(\mathrm{JS})}$ (1.35) based on the z_i values provide much more accurate overall predictions for the μ_i values. (By coincidence, $\hat{\mu}_i$ and μ_i both average 0.265; the average of $\hat{\mu}_i^{(\mathrm{JS})}$ must equal that of $\hat{\mu}_i^{(\mathrm{MLE})}$.)

Name	hits/AB	$\hat{\mu}_i^{(\mathrm{MLE})}$	μ_i	$\hat{\mu}_i^{(\mathrm{JS})}$
Clemente	18/45	.400	**.346**	.294
F Robinson	17/45	.378	**.298**	.289
F Howard	16/45	.356	**.276**	.285
Johnstone	15/45	.333	**.222**	.280
Berry	14/45	.311	**.273**	.275
Spencer	14/45	.311	**.270**	.275
Kessinger	13/45	.289	**.263**	.270
L Alvarado	12/45	.267	**.210**	.266
Santo	11/45	.244	**.269**	.261
Swoboda	11/45	.244	**.230**	.261
Unser	10/45	.222	**.264**	.256
Williams	10/45	.222	**.256**	.256
Scott	10/45	.222	**.303**	.256
Petrocelli	10/45	.222	**.264**	.256
E Rodriguez	10/45	.222	**.226**	.256
Campaneris	9/45	.200	**.286**	.252
Munson	8/45	.178	**.316**	.247
Alvis	7/45	.156	**.200**	.242
Grand Average		.265	**.265**	.265

A (famous) Baseball Example

Probably we can do better than the MLE here:

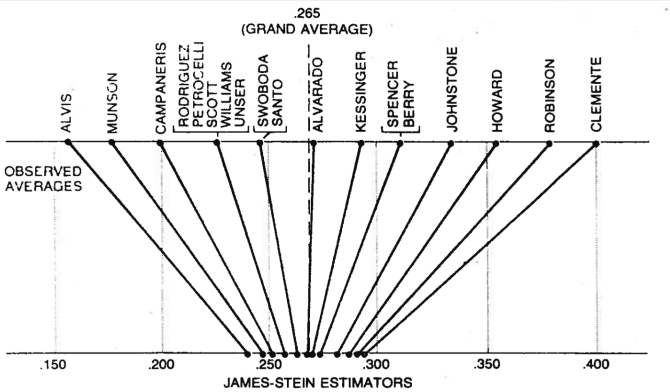
- Thurman Munson wins Rookie of the Year and ends up batting $\mu_i = .316$. If he batted .178 all year, his career would not have lasted long.
- Clemente's .400 seems unlikely to hold up. Last player to hit $> .400$ was Ted Williams .406 in 1941.
- But how?

Idea is to take an average between the observed average y_i and the overall mean $\bar{y}$:

$$\widehat{\mu}_i^{JS} = (1 - \lambda) \cdot \bar{y} + \lambda \cdot y_i, \quad \lambda = 1 - \frac{(m-3)\sigma^2}{\sum_i (y_i - \bar{y})^2}$$

- This has the effect of shrinking y_i towards the prior mean $\bar{y}$.
- In this case the prior mean is just $\bar{y}$ the grand-mean of all players
- How can information about unrelated players inform us about μ_i?
- Also consider proportion of foreign cars in Chicago as an additional y_i, can this help too?
- The shrinkage factor λ depends on sample size and variance, but how is it chosen?

A Baseball Example



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

The Diversion Ratio: Conlon and Mortimer (2018)

Raise price of good j. People leave. What fraction of leavers switch to k?

$$D_{jk} = \frac{\frac{\partial q_k}{\partial p_j}}{\left|\frac{\partial q_j}{\partial p_j}\right|}$$

It's one of the best ways economists have to characterize competition among sellers.

- High Diversion: Close Substitutes → Mergers more likely to increase prices.
- Very low diversion → products may not be in the same market.
  (ie: Katz & Shapiro)
- Demand Derivatives NOT elasticities.
- No equilibrium responses.

Estimating $\overline{D_{jk}}$

Remove a product j measure Δq_j and Δq_k.

$$\overline{D_{jk}} = \frac{\widehat{\Delta q_k}}{|\widehat{\Delta q_j}|} = \frac{E[q_k|Z=1] - E[q_k|Z=0]}{\left|E[q_j|Z=1] - E[q_j|Z=0]\right|} = \frac{E[q_k|Z=1] - E[q_k|Z=0]}{E[q_j|Z=0]}$$

Using a Beta-Binomial Prior

How to restrict $D_{jk} \in [0, 1]$?

$$\Delta q_k | \Delta q_j, D_{jk} \quad \sim \quad Bin(n = \Delta q_j, p = D_{jk})$$

Using a Beta-Binomial Prior

How to restrict $D_{jk} \in [0, 1]$?

$$
\begin{aligned}
\Delta q_k | \Delta q_j, D_{jk} &\sim Bin(n = \Delta q_j, p = D_{jk}) \\
D_{jk} | \beta_1, \beta_2 &\sim Beta(\beta_1, \beta_2) \\
E[D_{jk} | \beta_1, \beta_2, \Delta q_j, \Delta q_k] &= \frac{\beta_1 + \Delta q_k}{\beta_1 + \beta_2 + \Delta q_j} \\
\mu_{jk} &= \underbrace{\frac{\beta_1}{\beta_1 + \beta_2}}_{m_{jk}}, \quad \lambda = \frac{m_{jk}}{m_{jk} + \Delta q_j} \\
\widehat{D_{jk}} &= \lambda \cdot \mu_{jk} + (1 - \lambda) \frac{\widehat{\Delta q_k}}{\widehat{\Delta q_j}}
\end{aligned}
$$

μ_{jk} is prior mean; m_{jk} is no. pseudo-obs; λ weights our prior mean.
When we have a lot of experimental obs, prior receives little weight.

Using a Dirichlet Prior

How do restrict $D_{j\cdot} \in \Delta$?

- Same idea as before, but use Dirichlet Prior.
- Acts like pseudo-observations from the multinomial distribution.
- If we had same number of treated observations for each substitute we would have conjugacy/closed form (We don't).
- Likelihood is $\Delta q_k \sim Binomial(\Delta q_j, D_{jk})$ not multinomial.
- We use $m = 3.05$ pseudo-observations for the Dirichlet prior.
- Estimator is still technically non-parametric. Why?

Shrinkage Estimator, Intuition

- We “shrink” towards the prior mean when we have experimental estimates that are imprecise.
- Idea is very simple: when we have lots of data, use the experimental measure.
- When data are scarce: put more weight on the prior/model-based measure.
  - In practice: FTC/DOJ tend to assume diversion proportional to marketshare
  - Use plain logit (could also use more complicated model)
  - Logit sets the mean of the prior as: $\mu_{jk} = \frac{s_k}{1-s_j}$

Results: Mars

Firm	Product	# Weeks	Δq_k	Δq_j	$\frac{\Delta q_k}{\Delta q_j}$	Beta(J)	Beta(300)	Dirichlet(4.15)
		Snickers Removal						
Mars	M&M Peanut	176	375.52	-954.30	39.35	37.04	30.80	18.40
Mars	Twix Caramel	134	289.60	-702.39	41.23	37.86	29.49	15.88
Pepsi	Rold Gold (Con)	174	161.37	-900.11	17.93	16.84	13.95	7.54
Nestle	Butterfinger	61	72.95	-362.82	20.11	17.07	11.19	4.45
Mars	M&M Milk Chocolate	97	71.76	-457.36	15.69	13.83	9.85	4.14
Kraft	Planters (Con)	136	78.01	-759.87	10.27	9.57	7.80	3.81
Kellogg	Zoo Animal Cracker	177	65.72	-970.22	6.77	6.48	5.68	2.92
Pepsi	Sun Chip	159	45.30	-866.09	5.23	4.98	4.33	2.07
Hershey	Choc Hershey (Con)	41	29.78	-179.57	16.58	12.17	6.30	2.01
	Outside Good	180	460.89	-970.22	47.50			23.12
		M&M Peanut Removal						
Mars	Snickers	218	296.58	-1239.29	23.93	22.90	19.91	16.47
Mars	Twix Caramel	176	110.93	-1014.32	10.94	10.39	8.88	6.76
Mars	M&M Milk Chocolate	99	73.47	-529.58	13.87	12.46	9.18	6.26
Nestle	Raisinets	181	71.82	-1001.14	7.17	6.82	5.82	4.37
Kraft	Planters (Con)	190	61.42	-1046.10	5.87	5.62	4.90	3.60
Hershey	Twizzlers	62	32.98	-332.99	9.90	8.32	5.32	3.35
Kellogg	Rice Krispies Treats	46	22.37	-220.17	10.16	7.90	4.43	2.51
Pepsi	Frito	160	37.25	-902.42	4.13	3.95	3.47	2.37
	Outside Good	218	606.18	-1238.49	48.95			36.35

Results: Kellogg's

Firm	Product	# Weeks	Δq_k	Δq_j	$\frac{\Delta q_k}{\Delta q_j}$	Beta(J)	Beta(300)	Dirichlet(4.15)
		Animal Crackers Removal						
Pepsi	Rold Gold (Con)	132	114.39	-440.80	25.95	22.90	16.21	9.89
Mars	Snickers	145	92.44	-483.63	19.11	17.26	13.04	7.58
Mars	M&M Peanut	142	77.72	-469.44	16.55	14.98	11.43	6.47
Kellogg	CC Famous Amos	144	66.18	-478.20	13.84	12.40	9.15	5.39
Pepsi	Baked Chips (Con)	134	62.55	-447.60	13.97	12.46	9.13	5.27
Mars	Twix Caramel	110	50.17	-338.97	14.80	12.75	8.74	4.58
Sherwood	Ruger Wafer (Con)	119	48.20	-368.65	13.07	11.28	7.63	4.28
Hershey	Choc Herhsey (Con)	30	33.60	-132.57	25.34	17.14	7.86	3.81
Kellogg	Rice Krispies Treats	13	23.52	-37.80	62.22	23.24	7.16	2.99
Kar's Nuts	Kar Sweet&Salty Mix	95	30.06	-334.50	8.99	7.72	5.27	2.73
Misc	Popcorn (Con)	56	25.72	-226.89	11.34	8.92	5.08	2.61
Kraft	Planters (Con)	114	28.05	-380.25	7.38	6.53	4.78	2.43
Mars	M&M Plain	73	22.67	-295.07	7.68	6.47	4.26	2.15
	Outside Good	145	240.52	-482.91	49.81			21.98
		Famous Amos Removal						
Pepsi	Sun Chip	139	143.60	-355.68	40.37	34.39	22.66	15.75
Kraft	Planters (Con)	121	82.11	-332.61	24.69	20.89	13.68	8.75
Hershey	Choc Hershey (Con)	38	48.60	-66.84	72.72	36.93	13.36	7.18
Pepsi	Frito	119	49.88	-313.21	15.93	13.44	8.85	5.32
Misc	Rasbry Knotts	133	46.62	-345.38	13.50	11.45	7.49	4.81
Pepsi	Grandmas Choc Chip	95	39.99	-259.21	15.43	12.51	7.62	4.49
Pepsi	Dorito Buffalo Ranch	72	38.11	-224.24	17.00	13.28	7.53	4.43
Pepsi	Chs PB Frito Cracker	34	26.87	-83.65	32.13	18.16	7.14	3.74
Kellogg	Choc Sandwich FA	57	27.97	-122.04	22.91	15.06	6.84	3.69
Pepsi	Rold Gold (Con)	147	32.62	-392.22	8.32	7.40	5.54	3.19
Kraft	Oreo Thin Crisps	29	20.73	-43.29	47.89	19.20	6.12	3.05
Mars	Combos (Con)	98	23.56	-274.54	8.58	7.03	4.34	2.61

Comparison of Assumptions: Snickers Experiment

	Total	Assn 1	Assn 2	Assn 3 ($m = K$)	Assn 4 (m = 4.15)
Products with $D_{jk} < 0$	51	24	26	0	0
Products with $0 \leq D_{jk} \leq 10$	51	13	15	43	48
Products with $10 \leq D_{jk} \leq 20$	51	5	5	5	2
Products with $D_{jk} > 20$	51	9	5	3	1
Sum of all positive D_{jk}s	51	402.84	301.95	265.41	98.72
Sum of all negative D_{jk}s	51	-238.90	-239.07	0.00	0.00

Note: Table includes only products for which there were at least 50 sales of the focal product in control weeks, on average.

How do we estimate these models?

A few options

- With conjugate priors we can closed forms
- Can do it by hand if we have sufficient statistics
  - Clear in beta-binomial or poisson-gamma relationship.
  - Mostly not the case.
- Mostly we do what is known as Markov Chain Monte Carlo
- The goal is to draw from $f(\theta|\mathbf{X})$ or to compute moments of the distribution $E_f[g(\theta)]$.

- Suppose we want to calculate a function

$$E_f[g(x)] = \int g(x)f(x)dx$$

- How do we do it?
  1. Draw from $\hat{x}_s \sim f(x)$
  2. Caclulate $g(\hat{x}_s)$
  3. Repeat for $s = 1 \ldots, S$
  4. Calculate $E[\hat{g}] = \frac{1}{S}\sum_{s=1}^{S} g(\hat{x}_s)$

Monte Carlo Example

- Let's integrate $g(x) = \phi(x)$ (normal pdf) over $f(x) = Unif(0, 1)$ from $[0, 1]$.
- We know the answer is $\Phi(1) - \Phi(0)$.

```
Integral <- function(n){
    X <- runif(n)
    Y <- exp(-X^2/2)/sqrt(2*pi)
    Int <- sum(Y)/n
    Error <- Int-(pnorm(1)-pnorm(0))
    list(Int,Error)}
```

MCMC

- Linear regression is nice because it has closed forms (even in Bayesian framework).
- In most nonlinear models, it's hard to write down the posterior in closed form.
- It's often easier to sample from a distribution than to characterize it.

Gibbs Sampling

The first building block is known as the Gibbs Sampler

- Suppose that $p(x, y)$ is a p.d.f or p.m.f that is hard to sample directly from.
- But suppose that $p(x|y)$ or $p(y|x)$ are easy to sample from.
- Gibbs sampler says:
  1. Initialize (x_0, y_0).
  2. Randomly draw $y_1 \sim g(y|x_0)$.
  3. Randomly draw $x_1 \sim f(x|y_1)$.
  4. Randomly draw $y_2 \sim g(y|x_1)$.
  5. Rinse and Repeat.
- This sequence $(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots$ is a Markov Chain.
- Why? Because $(x_k, y_k)|(x_{k-1}, y_{k-1})$ doesn't depend on x_{k-h} or y_{k-h} for $h \geq 2$.
  - This does not mean that (x_3, y_3) and (x_1, y_1) are I.I.D!

Comments

- Can prove that the Gibbs sampler converges to the posterior distribution of $\theta|X$, regardless of starting value θ^0, under some regularity conditions.
  - In practice, you throw out a chunk of observations at the beginning of a run as "burn-in."
- For any statistic you want to compute, it would suffice to draw from the posterior distribution of (u, β, Σ) and simulate the statistic at these points.
  - In practice: you pick every mth point along your chain.
- No explicit optimization!

- Consider the following model:

$$u_{ij} = X_{ij}\beta + \epsilon^*_{ij},\ j = 1, \ldots, J-1$$
$$u_{io} = \epsilon_{io}.$$

In matrix form:

$$u_i = X_i\beta + \epsilon_i,\ \epsilon_i \sim N(0, \Sigma).$$

We observe

$$y_{ij} = 1(u_{ij} \geq u_{ik} \forall k),$$

as well as X.

Key insight

- We want to be able to sample from

$$p(\beta, \Sigma | y_1, \ldots, y_N, X).$$

- If we knew u we could estimate β, Σ via linear regression.
- We don't know u, but if we can sample from its marginal distribution and treat it as data, we can alternate "regression" steps and new draws of u.

McCulloch and Rossi (JoE 1994)

- Estimation procedure as follows.
  - Start with a guess of β, Σ.
  - Iterate the following steps:
    1. Draw $u|\beta, \Sigma$, subject to constraints imposed by y_i :
    - If $y_{ij} = 1$ then we must have $u_{ij} \geq u_{ik}$ for all $k \neq j$.
    1. With u in hand, derive β, Σ from posterior distribution given their priors, X, and u. (regression step)
  - Repeating steps (a) and (b) creates a markov chain which converges to the joint distribution of β, Σ, u.
  - Estimator is an example of a Gibbs sampler.

Estimation outline

- Impose Normal and IW priors for β, Σ respectively.
- Iterate the following steps:
  - Step 1: $N * J$ conditional distributions $u_{ij}|u_{i,-j}, \beta, \Sigma, X, y$
  - Step 2: $\beta|u, \Sigma, X, y$
  - Step 3: $\Sigma|u, \beta, X, y$
- Steps (2) and (3) are exactly as in the multivariate regression model.

Step 1 details

- for each i, j, $u_{ij}|u_{i,-j}, \beta, G, y_i$ is distributed according to a truncated univariate normal distribution.
  1. If $u = \begin{bmatrix} u_{ij} \\ u_{i,-j} \end{bmatrix}$ is normally distributed with mean $\mu = \begin{bmatrix} \mu_{ij} \\ \mu_{i,-j} \end{bmatrix}$ and variance $\Sigma = \begin{bmatrix} \Sigma_{jj} & \Sigma_{j,-j} \\ \Sigma_{-j,j} & \Sigma_{-j,-j} \end{bmatrix}$, then

     $$u_{ij}|u_{i,-j} \sim N(\mu_{ij} + \Sigma_{-j,j}\Sigma_{-j,-j}^{-1}\left(u_{i,-j} - \mu_{i,-j}\right),\ \Sigma_{jj} - \Sigma_{-j,j}\Sigma_{-j,-j}^{-1}\Sigma_{j,-j}).$$

     Here, $\mu = X_i\beta$.
  2. Truncation points:
     - 2.1 If $y_{ij} = 1$, then we know $u_{ij} \geq \max_{k \neq j} u_{ik}$.
     - 2.2 Otherwise, $u_{ij} \leq \max_{k \neq j} u_{ik}$.
  3. To draw from a distribution F truncated at $[l, u]$, draw a uniform random number x and compute

     $$u_{ij} = F^{-1}(xF(l) + (1 - x)F(u)).$$

- Treating u as an additional unknown parameter is an example of **data augmentation**.
- Well-chosen auxiliary parameters can make Bayesian estimation tractable.
- Try augmenting the data with $\epsilon \equiv u - X\beta$ instead of u!

Example (High-probability island)

Let

$$S = \{0, 1\}^{100}.$$

Suppose that the zero vector has probability $P(0) = 1/2$, and all other vectors are equally likely, i.e. occur with probability $P(s) = \frac{1}{2} \cdot \frac{1}{2^{100}-1}$.
Consider a Gibbs sampler with an initial value of $s = 0$. Suppose that we are updating the jth component. There are two possible outcomes, e_j and 0. We have:

$$p(e_j|0) = \frac{1}{2^{100}-1} \Big/ \left(\frac{1}{2^{100}-1} + 1\right) \approx \frac{1}{2^{100}}$$

$$p(0|0) = 1 \Big/ \left(\frac{1}{2^{100}-1} + 1\right) \approx 1.$$

Once 0 is reached, it will take approximately 2^{100} draws on average to draw a value other than 0. Therefore it will take *many* draws to estimate $P(0)$.

When does Gibbs sampling fail?

Another failure mode occurs when there are multiple high-probability islands.

Example (Two islands)

Let

$$S = \{0, 1\}^2.$$

and suppose

$$P(0, 0) = \frac{1}{2}$$
$$P(1, 1) = \frac{1}{2}$$
$$P(0, 1) = 0$$
$$P(1, 0) = 0.$$

In this case a sampler that starts at $(0, 0)$ or $(1, 1)$ will never reach the other high-probability state.

- This example is an extreme case.
- But Gibbs samplers may mix poorly when the underlying density has multiple separated peaks.

Hierarchical models

- Bayesian approach to random effects and random coefficients.
  - Actually much more general than this.
  - Hierarchical models have some parameters that are distributed according to distributions which in turn have priors.
- Example (random coefficients, McCulloch and Rossi): In the multinomial probit model, we could impose structure on the covariance matrix as follows:

$$
\begin{aligned}
u_i &= X_i\beta_i + \epsilon_i \\
\epsilon_i &\sim N(0, I) \\
\beta_i &\sim N(\overline{\beta}, \Sigma_{\beta_i}) \\
\overline{\beta} &\sim N\left(\overline{\overline{\beta}}, \Sigma_{\overline{\beta}}\right) \\
\Sigma_{\overline{\beta}} &\sim IW(\alpha, \nu).
\end{aligned}
$$

In this case, β_i are normally-distributed random coefficients. The parameters of $F(\beta_i)$ in turn have prior distributions.

Random coefficients: estimation

- Augment data w/ random coefficients β_i as well as u_i.
- Use Gibbs sampler:
  - Update $u|\beta_i, X$ exactly as before, except that variance of ϵ is fixed at $N(0, I)$.
  - Update $(\beta_i|u, \overline{\beta}, \Sigma_{\beta_i}, X_i)$, one observation i at a time.
  - Update $\overline{\beta}, \Sigma_{\beta_i}|\{\beta_i\}_{i=1,\ldots,N}$ as in Bayesian linear regression.
- See McCulloch and Rossi for estimation procedure.

Metropolis-Hastings

- Gibbs samplers require you to know conditional densities.
- Often, all you know is a function that is proportional to the (conditional) density.
  - Example: you want to draw from $f(x|x \in R)$ for some region R, and you know $f(x)$, but it is difficult to characterize the region R.
  - In this case it's easy to compute $f(x) \cdot 1(x \in R)$ but difficult to compute $\frac{f(x)1(x \in R)}{\int_R f(x)dx}$.
- Metropolis-Hastings simulators are algorithms to sample from a density when you know a function proportional to it.
- You can even use MH for conditional distributions within a Gibbs sampler (known as MH-within-Gibbs).

MH ALGORITHM

- Suppose you know a function $\pi(x|data)$ which is proportional to $f(x|data)$.
- The Metropolis-Hastings simulator creates a sequence that converges to the stationary distribution $\pi(\cdot)$.
- Algorithm is as follows:
  1. Fix a proposal density $q(x'|x)$.
  2. Start at $x = x^{(0)}$.
  3. For iteration $i = 1, \ldots, T$ do:
     - Propose $x' \sim q(x'|x^{(i-1)})$
     - Define the acceptance probability

$$\alpha(x'|x) = \min\left\{1, \frac{q(x'|x^{(i-1)})\pi(x'|data)}{q(x^{(i-1)}|x')\pi(x^{(i-1)}|data)}\right\}.$$

     - Update x according to:

$$x^{(i)} = \begin{cases} x' & \text{with probability } \alpha(x'|x) \\ x^{(i-1)} & \text{with probability } 1 - \alpha(x'|x) \end{cases}.$$

Choosing a proposal density

- Symmetric proposal density

$$q(x'|x) = q(x|x')$$

is a popular choice.

  - This option is known as **random walk Metropolis-Hastings**.
  - Normal distribution is a popular choice within this class:

$$q(x'|x) = \phi(x' - x;\ 0, \sigma_q^2).$$

  - For any symmetric $q(\cdot|\cdot)$:

$$\alpha(x'|x) = \min\left\{1, \frac{\pi(x'|data)}{\pi(x|data)}\right\}.$$

  - Aim for acceptance rate$\approx 1/3$ by choice of scale of proposal distribution?
    - Too low and too many draws are needed.
    - Too high and the chain mixes slowly.

Other proposal densities

1. Independent proposal density:

$$\alpha(x'|x) = \min\left\{1, \frac{q(x')\pi(x'|data)}{q(x)\pi(x|data)}\right\}.$$

   - Can be convenient if some choice of $q(\cdot)$ is easy to sample from **and** results in an easy-to-evaluate expression for $q(x)\pi(x)$.

2. Other proposal densities that simplify the expression $\frac{q(x'|x^{(i-1)})\pi(x'|data)}{q(x^{(i-1)}|x')\pi(x^{(i-1)}|data)}$.

Other methods when your "density" doesn't integrate to 1

- MH is not the only method for sampling when all you know is a function proportional to the density.
- Slice sampling (Neal 2003) can be practical on some problems.
- Also: Hamiltonian Monte Carlo (aka "Hybrid Monte Carlo")
  - Requires gradient, but can be much faster to converge than MH.

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo combines ideas from physics with the Metropolis-Hastings algorithm to sample from the posterior with much less autocorrelation than random-walk MH algorithms.
My slides follow Neal (2012), “MCMC using Hamiltonian Dynamics”

Hamiltonian

Think of a marble rolling around on a surface. The marble has:

- a d-dimensional position vector, q
- a d-dimensional momentum vector, p

Define the *Hamiltonian* as follows:

$$H(q, p) = U(q) + K(p)$$

where

- $U(q)$ is the potential energy (depends on position).
- $K(p)$ is the kinetic energy (depends on how fast the marble is moving).

The partial derivatives of the Hamiltonian determine how p and q evolve.

$$\frac{\partial q_i}{\partial t} = \frac{\partial H}{\partial p_i}$$
$$\frac{\partial p_i}{\partial t} = -\frac{\partial H}{\partial q_i}$$

These equations imply that the Hamiltonian is invariant.

Details

$$H(q, p) = U(q) + K(p)$$

1. Typically, $K(p) = p'M^{-1}p/2$, where M is a symmetric, positive-definite "mass matrix". (This is the formula for kinetic energy.)
2. Observe that $K(p) = p'M^{-1}p/2$ is also the log-pdf of a Normal with mean 0 and covariance M.
3. In physics, $U(\cdot)$ would represent the height of the surface at different locations.
4. We're going to choose $U(\cdot)$ to have a probabiliistic interpretation.

Interpretation as probability

- Fact (from statistical mechanics): Given a "temperature" T, Hamiltonian dynamics imply a joint distribution of states, given by

$$Pr(q,p) = \frac{1}{Z}\exp\left(\frac{-H(q,p)}{T}\right)$$

  - Z is the constant needed to make $Pr(\cdot)$ integrate to 1.
- In our case, taking $T = 1$, we have

$$Pr(q,p) = \frac{1}{Z}\exp\left(-U(q)\right)\exp\left(-K(p)\right).$$

- Goal: pick $U(\cdot)$ so that the marginal of $Pr(\cdot)$ is the posterior over q.
- Achieve this by picking

$$U(q) = -\log\left(\pi(q)L(q|D)\right)$$

  where:
  - $\pi(q)$: prior density
  - $L(q|D)$: likelihood given data D.

Hamiltonian Monte Carlo algorithm

Assume

1. prior $\pi(q)$ on parameters q
2. Gaussian prior w/ diagonal covariance matrix on p, **independent of** $\pi(\cdot)$
3. diagonal mass matrix.

Iterate as follows:

0 At beginning of iteration t, algorithm has state (q_t, p_t).

1 Draw a new p, iid, from independent Gaussian distribution. Consider state (q_t, p).

2 Simulate a trajectory using Hamiltonian dynamics. Arrive at (q', p').

3 Use the (standard) Metropolis-Hastings acceptance rule for a symmetric proposal density:

$$\alpha = \min[1, \exp(-H(q', p') + H(q_t, p_t))].$$

With probability α, the new state is (q', p'); otherwise it's (q_t, p_t) again.

Simulating Hamiltonian dynamics

- Hamiltonian is conserved under (continuous) Hamiltonian dynamics.
- To implement on a computer, need to discretize. Some subtlety to simulation/discretization.
- Naive methods could cause simulation errors to blow up.
- Use “leapfrog” algorithm or refinements thereof. This algorithm keeps Hamiltonian (exactly) constant. See details in Neal (2012).

Hamiltonian Monte Carlo: comments

1. Equations of motion involve gradient of log-likelihood. Need to provide this.
2. Why is the proposal density symmetric? Can prove that "forward" trajectory is as likely as "backward".
3. If we skipped step (1), would always have $H(q_t, p_t) = H(q', p')$. Only change in likelihood of (p, q) comes from new momentum draw in step (1).
4. Step 1 is necessary. If total energy is fixed, particle will never reach states with likelihood below a given finite value.)
5. Well-tested implementation in STAN (domain-specific language; callable from `R`/`Python`/`Julia`; provides automatic differentiation). Fantastic if you can phrase your problem in its language.
6. If you can shoehorn your problem into STAN, do it!

Stan Code

```
% Main Specification: Dirichlet Prior
data {
  int<lower=1> J;               // number of products, including outside good
  int<lower=1> N[J];            // number of trials
  int<lower=0> y[J];            // number of successes for each product j
  vector[J] priors;    // mean of the distribution of alpha
}

parameters {
  simplex[J] theta;
}

model {
  theta ~ dirichlet(priors);
  for (j in 1:J) {
    y[j] ~ binomial(N[j], theta[j]);
  }
}
```

An MCMC approach to classical estimation (Chernozhukov and Hong, JoE 2003)

- It turns out that MCMC can be useful outside of Bayesian framework.
- CH consider case of extremum estimation
  - Review: want to maximize a criterion function $L_n(\theta)$, where n is sample size.
  - Leading example: GMM

$$L_n(\theta) = -\frac{1}{2}\left(n^{-1/2}\sum_{i=1}^{n} m_i(\theta)\right)' W_n(\theta)\left(n^{-1/2}\sum_{i=1}^{n} m_i(\theta)\right)$$

for moment function $m_i(\theta)$, weight matrix $W_n(\theta)$.

Laplace-type estimators

- Consider an objective function $L_n(\theta)$.
- Pick a prior $\pi(\theta)$ strictly positive and continuous over Θ.
- The following transformation is known as a **quasi-posterior:**

$$p_n(\theta) \equiv \frac{\pi(\theta)\exp(L_n(\theta))}{\int_\Theta \pi(\theta)\exp(L_n(\theta))d\theta}.$$

- Quasi-posterior mean is an example of a **Laplace-type estimator**:

$$\int_\Theta \theta p_n(\theta)d\theta$$

  - Formally: LTEs solve

$$\hat{\theta} = \arg\inf_\zeta \int \rho_n(\theta - \zeta)p_n(\theta)d\theta$$

  - Different choices of ρ_n give posterior means, medians, τth quantiles.

Computing LTEs

$$p_n(\theta) \equiv \frac{\pi(\theta)\exp(L_n(\theta))}{\int_\Theta \pi(\theta)\exp(L_n(\theta))d\theta}.$$

- Even if $L_n(\theta)$ is not a likelihood, $p_n(\cdot)$ is a proper density over Θ.
- Goal: compute mean.
- Can do so via MCMC:

$$\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B}\theta^{(i)}$$

- In practice:
  - Constant term $\int_\Theta \pi(\theta)\exp(L_n(\theta))d\theta$ may be difficult to compute.
  - Use Metropolis-Hastings with $f(\theta) = \pi(\theta)\exp(L_n(\theta))$.

$$p_n(\theta) \equiv \frac{\pi(\theta)\exp(L_n(\theta))}{\int_\Theta \pi(\theta)\exp(L_n(\theta))d\theta}.$$

1. If $L_n(\cdot)$ is log-likelihood, then $p_n(\theta)$ is the (Bayesian) posterior.
2. Paper gives two procedures for inference.
   - 2.1 Can estimate variance along your sequence, use delta method.
   - 2.2 Under stronger conditions, to estimate a 95% CI, can use quantiles of sequence $\theta^{(i)}$.
     - 2.2.1 Valid in GMM iff you use best estimate of optimal weight matrix.
     - 2.2.2 See paper for details.