Multiple parameter models II

April 9, 2023

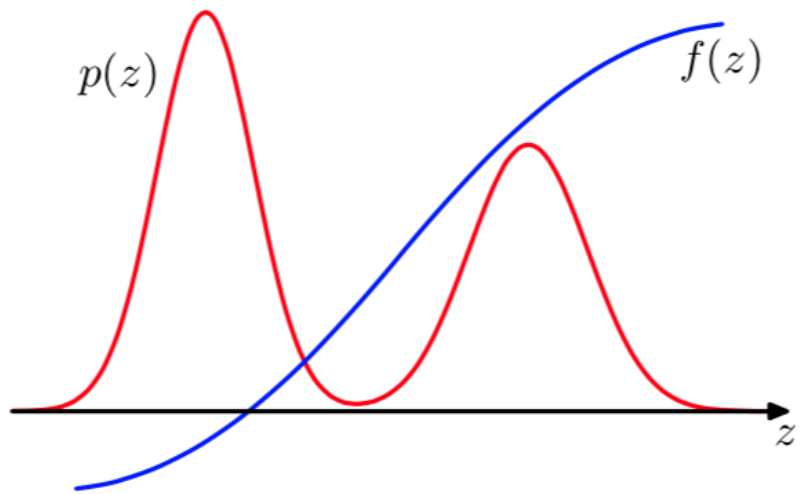
Overview

Introduction

- The fundamental problem that we wish to address involves finding the expectation of some function $f(z)$ with respect to a probability distribution $p(z)$

$$E[f] = \int f(z)p(z)dz$$

Schematic illustration of a function $f(z)$ whose expectation is to be evaluated with respect to a distribution $p(z)$.



Introduction

- The general idea behind sampling methods is to obtain a set of samples $z^{(\ell)}$ drawn independently from the distribution $p(z)$

$$\hat{f} = \frac{1}{L} \sum_{\ell=1}^{L} f(z^{(\ell)})$$

  With a moderate (10 to 20) L, $E[\hat{f}] = E[f]$
- The problem is the samples $\{z^{(\ell)}\}$ might not be independent.

Standard distributions

- Suppose that z is uniformly distributed over the interval $(0, 1)$, we transform the values of z using some function $f(\cdot)$ so that $y = f(z)$.
- The distribution of y will be governed by

$$p(y) = p(z)\left|\frac{dz}{dy}\right|$$

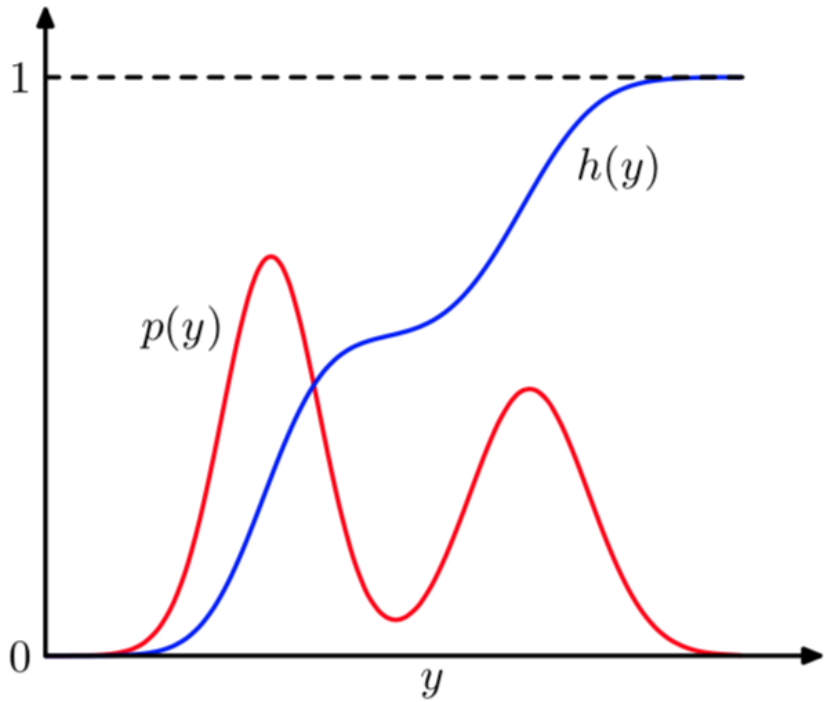
- Our goal is to choose the function $f(z)$ such that the resulting values of y have some specific desired distribution $p(y)$

$$z = h(y) \equiv \int_{-\infty}^{y} p(\hat{y})\mathrm{d}\hat{y}$$

thus $y = h^{-1}(z)$

Standard distributions

- Geometrical interpretation of the transformation method for generating nonuniformly distributed random numbers. $h(y)$ is the indefinite integral of the desired distribution $p(y)$. If a uniformly distributed random variable z is transformed using $y = h^{-1}(z)$, then y will be distributed according to $p(y)$.



Standard distributions

- consider the exponential distribution

$$p(y) = \lambda exp(-\lambda y)$$

In this case, $h(y) = 1 - exp(-\lambda y)$ so the transformation from uniform distributed z using $y = -\lambda^{-1} ln(1 - z)$, then y will have an exponential distribution.

Rejection Sampling

- Sample from relatively complex distributions, subject to certain constrains.
- Suppose we wish to sample from a distribution $p(z)$ that is not one of the simple, standard distributions considered so far, and that sampling directly from $p(z)$ is difficult.
- We are easily to evaluate $p(z)$ for any given value of z, up to some normalizing constant Z so that

$$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$

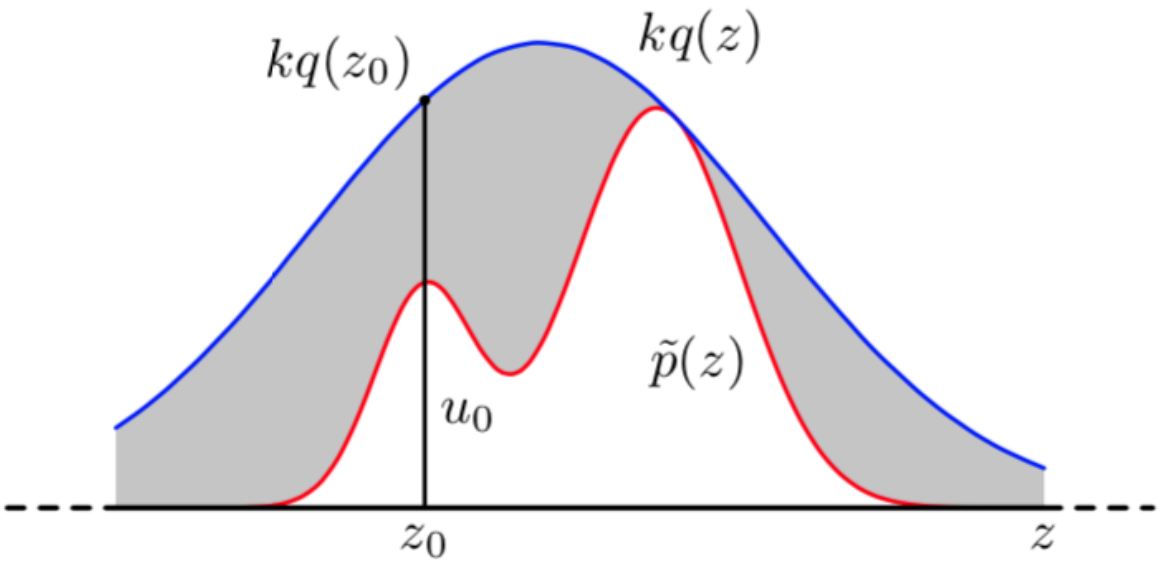
where $\widetilde{p}(z)$ can readily be evaluated, but Z_p is unknown.

Rejection Sampling

- In order to apply rejection sampling, we need some simpler distribution $q(z)$, sometimes called a proposal distribution, from which we can readily draw samples.
- We next introduce a constant k whose value is chosen such that $kq(z) \geq \widetilde{p}(z)$ for all values of z.
- The function $kq(z)$ is called the comparison function and is illustrated for a univariate distribution.
- Each step of rejection sampler involves generating two random numbers.
  - first, we generate a number z_0 from the distribution $q(z)$.
  - Nest, we generate a number u_0 from a uniform distribution over $[0, kq(z_0)]$.
- This pair of random numbers has uniform distribution under the curve of the function $kq(z)$.
- Finally, if $u_0 > \widetilde{p}(z_0)$ then the sample is rejected, otherwise u_0 is retained.

Rejection Sampling

In the rejection sampling method, samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\widetilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\widetilde{p}(z)$.



One of the key point is we need to select the k as small as possible.

Importance Sampling

- The technique of importance sampling provides a framework for approximating expectations directly but does not itself provide a mechanism for drawing samples from distribution $p(z)$.

$$\mathbb{E}[f] \simeq \sum_{l=1}^{L} p\left(\mathbf{z}^{(l)}\right) f\left(\mathbf{z}^{(l)}\right)$$

- As in the case of rejection sampling, importance sampling is based on the use of a proposal distribution $q(z)$ from which it is easy to draw samples.

Importance Sampling

- We can express the expectation in the form the expectation in the form of a finite sum over samples $\{z^{(\ell)}\}$ drawn from $q(z)$

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} \\ &\simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p\left(\mathbf{z}^{(l)}\right)}{q\left(\mathbf{z}^{(l)}\right)}f\left(\mathbf{z}^{(l)}\right)\end{aligned}$$

- The quantities $r_\ell = p(z^{(\ell)})/q(z^{(\ell)})$ are known as importance weights.

Importance Sampling

- It will often be the case that the distribution $p(z)$ can only be evaluated up to a normalization constant, so that $p(z) = \widetilde{p}(z)/Z_p$ where $\widetilde{p}(z)$ can be evaluated easily.
- Similarly, we may wish to use an importance sampling distribution $q(z) = \widetilde{q}(z)/Z_q$, which has the sample property.

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\mathrm{d}\mathbf{z} \\ &\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l f\left(\mathbf{z}^{(l)}\right)\end{aligned}$$

Importance Sampling

- we can use the same sample set to evaluate the ratio Z_p/Z_q with the result

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \widetilde{p}(\mathbf{z})\mathrm{d}\mathbf{z} = \int \frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})} q(\mathbf{z})\mathrm{d}\mathbf{z}$$

$$\simeq \frac{1}{L}\sum_{l=1}^{L} \widetilde{r}_l$$

- Hence

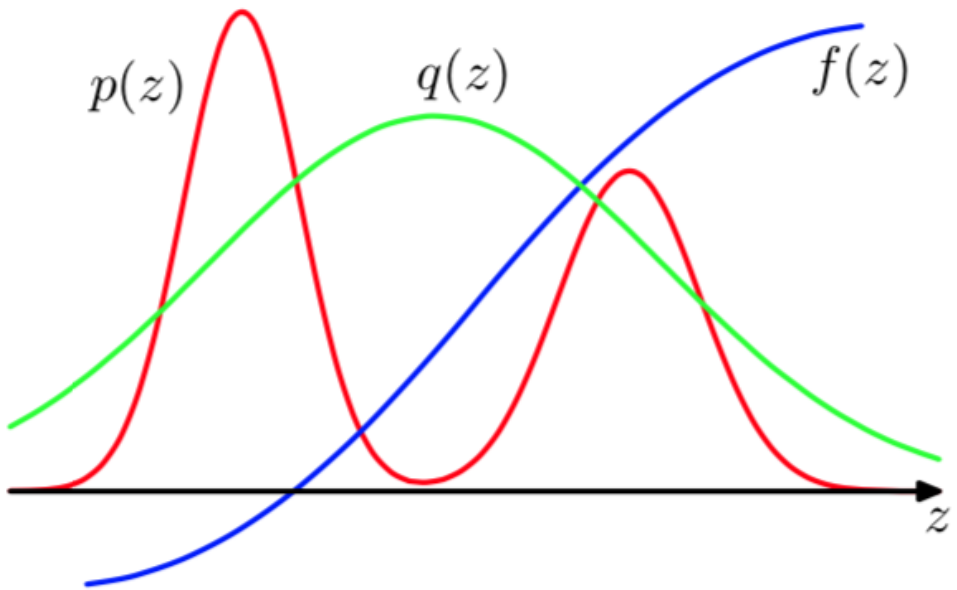
$$\mathbb{E}[f] \simeq \sum_{l=1}^{L} w_l f\left(\mathbf{z}^{(l)}\right)$$

in which

$$w_l = \frac{\widetilde{r}_l}{\sum_m \widetilde{r}_m} = \frac{\tilde{p}\left(\mathbf{z}^{(l)}\right)/q\left(\mathbf{z}^{(l)}\right)}{\sum_m \tilde{p}\left(\mathbf{z}^{(m)}\right)/q\left(\mathbf{z}^{(m)}\right)}$$

Importance Sampling

Importance sampling addresses the problem of evaluating the expectation of a function $f(z)$ with respect to a distribution $p(z)$ from which it is difficult to draw sampled directly. Instead, samples $\{z^{(\ell)}\}$ are drawn from a simpler distribution $q(z)$, and the corresponding terms in the summation are weighted by the ratios $p(z^{(\ell)})/q(z^{(\ell)})$.



MCMC

- As with rejection and importance sampling, we again sample from a proposal distribution. This time, however,we maintain a record of the current state $z^{(\tau)}$, and the proposal distribution $q(z|z^{(\tau)})$ depends on this current state, and so the sequence of samples $z^{(1)}, z^{(2)}, \ldots$ forms a Markov chain.
- if we write $p(z) = \widetilde{p}(z)/Z_p$, we will assume that $\widetilde{p}(z)$ can readily be evaluated for any given value of z.
- The proposal distribution itself is chosen to be sufficiently simple that it is straightforward to draw a candidate sample $z*$ from the proposal distribution and then accept the sample according to an appropriate criterion.

MCMC

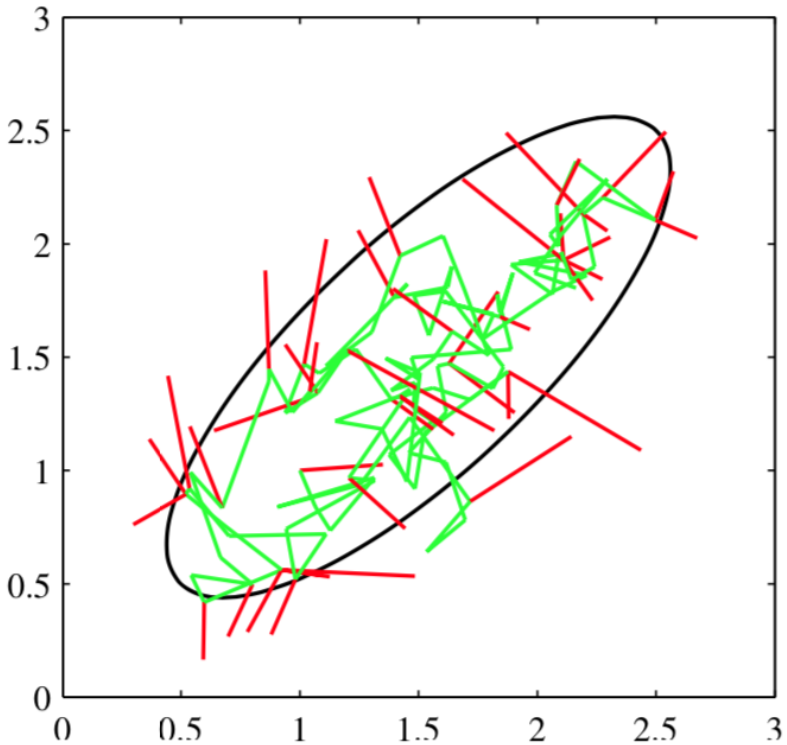
- In the basic Metropolis algorithm, we assume that the proposal distribution is symmetric, that is $q(z_A|z_B) = q(z_B|z_A)$ for all values of z_A and z_B.
- Then we have

$$A\left(\mathbf{z}^{\star}, \mathbf{z}^{(\tau)}\right) = \min\left(1, \frac{\tilde{p}\left(\mathbf{z}^{\star}\right)}{\widetilde{p}\left(\mathbf{z}^{(\tau)}\right)}\right)$$

- This can be achieved by choosing a random number u with uniform distribution over the unit interval $(0, 1)$ and then accepting the sample if $A\left(\mathbf{z}^{\star}, \mathbf{z}^{(\tau)}\right) > u$

MCMC

- A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution whose standard deviation is 0.2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected.



MCMC

- The generalized Metropolis-Hastings algorithm does not require the proposal distribution to be symmetric.
- At step τ of the algorithm, in which the current state is $z^{(\tau)}$, we draw a sample $z*$ from the distribution $q_k(z|z^{(\tau)})$ and then accept it with probability

$$A_k\left(\mathbf{z}^{\star}, \mathbf{z}^{(\tau)}\right) = \min\left(1, \frac{\tilde{p}\left(\mathbf{z}^{\star}\right) q_k\left(\mathbf{z}^{(\tau)} \mid \mathbf{z}^{\star}\right)}{\widetilde{p}\left(\mathbf{z}^{(\tau)}\right) q_k\left(\mathbf{z}^{\star} \mid \mathbf{z}^{(\tau)}\right)}\right)$$

Here k labels the members of the set of possible transitions being considered.

- We can show that $p(z)$ is an invariant distribution of the Markov chain defined by the Metropolis-Hastings algorithm by showing that detailed balance, defined by

$$p^{\star}(\mathbf{z})T\left(\mathbf{z},\mathbf{z}'\right)=p^{\star}\left(\mathbf{z}'\right)T\left(\mathbf{z}',\mathbf{z}\right)$$

  is satisfied.

- Then we have

$$\begin{aligned}p(\mathbf{z})q_k\left(\mathbf{z}\mid\mathbf{z}'\right)A_k\left(\mathbf{z}',\mathbf{z}\right)&=\min\left(p(\mathbf{z})q_k\left(\mathbf{z}\mid\mathbf{z}'\right),p\left(\mathbf{z}'\right)q_k\left(\mathbf{z}'\mid\mathbf{z}\right)\right)\\&=\min\left(p\left(\mathbf{z}'\right)q_k\left(\mathbf{z}'\mid\mathbf{z}\right),p(\mathbf{z})q_k\left(\mathbf{z}\mid\mathbf{z}'\right)\right)\\&=p\left(\mathbf{z}'\right)q_k\left(\mathbf{z}'\mid\mathbf{z}\right)A_k\left(\mathbf{z},\mathbf{z}'\right)\end{aligned}$$

  as required.

MCMC

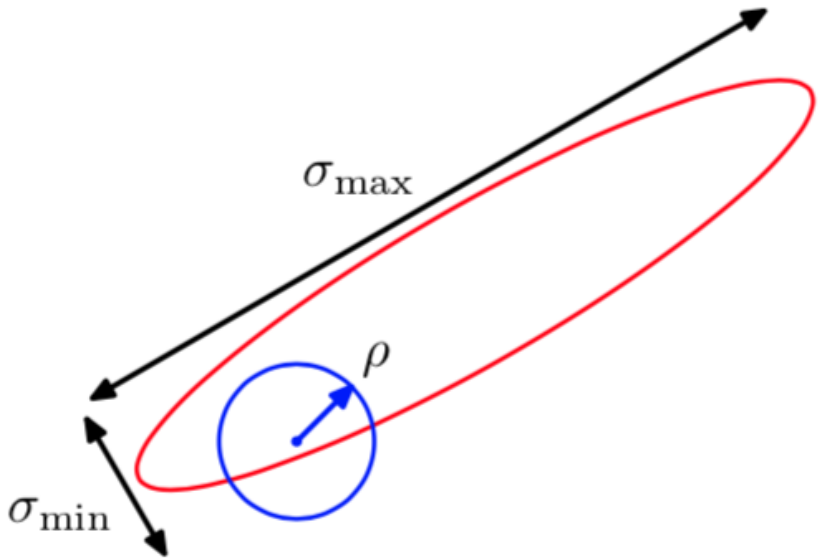
- The specific choice of proposal distribution can have a marked effect on the performance of the algorithm.
- For continuous state spaces, a common choice is a Gaussian centered on the current state, leading to an important trade-off in determining the variance parameter of this distribution.
  - small variance: high proportion of accepted transitions, but slow random walk.
  - big variance: high rejection rate.
- Consider a multivariate distribution $p(z)$ having strong correlations between the components of z. The scale ρ of the proposal distribution should be as large as possible without incurring high rejection rates.

MCMC

- Schematic illustration of the use of an isotropic Gaussian proposal distribution (blue circle) to sample from a correlated multivariate Gaussian distribution (red ellipse) having very different standard deviations in different directions, using the Metropolis-Hastings algorithm.

MCMC

- In order to keep the rejection rate low, the scale ρ of the proposal distribution should be on the order of the smallest standard deviation σ_{min}, which leads to random walk behavior in which the number of steps separating states that are approximately independent is of order $(\sigma_{max}/\sigma_{min})^2$ where σ_{max} is the largest standard deviation.



Multinomial model for categorical data

- The multinomial distribution is used to describe data for which each observation is one of k possible outcomes.
- If y is the vector of counts of the number of observations of each outcome, then the likelihood of

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}$$

  where the sume of the probabilities, $\sum_{j=1}^{k} \theta_j$, is 1.
- The conjugate prior distribution of binomial distribution is beta distribution
- And binomial distribution is a two-category multinomial distribution.
- What might the conjugate prior distribution of multivariate distribution look like?

Multinomial model for categorical data

- The conjugate prior distribution is a multivariate generalization of the beta distribution known as Dirichlet

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1},$$

where the distribution is restricted to nonnegative θ_js with $\sum_{j=1}^{k} \theta_j = 1$. Please conduct the posterior distribution.

- For more information,

Dirichlet	$\theta \sim \text{Dirichlet}(\alpha_1, .., \alpha_k)$ $p(\theta) = \text{Dirichlet}(\theta\|\alpha_1, .., \alpha_k)$	'prior sample sizes' $\alpha_j > 0$; $\alpha_0 \equiv \sum_{j=1}^{k} \alpha_j$

$p(\theta) = \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}$ $\theta_1, .., \theta_k \geq 0; \sum_{j=1}^{k} \theta_j = 1$	$\text{E}(\theta_j) = \frac{\alpha_j}{\alpha_0}$ $\text{var}(\theta_j) = \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}$ $\text{cov}(\theta_i, \theta_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}$ $\text{mode}(\theta_j) = \frac{\alpha_j - 1}{\alpha_0 - k}$

Multinomial model for categorical data

- The resulting posterior distribution for the θ_js is also Dirichlet with parameters $\alpha_j + y_j$.
- The prior distribution expressed on the scale of α is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{k}(\alpha_j - 1)$ observations with $\alpha_j - 1$ observations of the jth outcome category.
- The noninformative Dirichlet prior distribution could be set as $\alpha_j = 1$ for all j.
- Can we set the $\alpha_j = 2$ for all j as the prior distribution? What is the difference?

Dirichlet Example

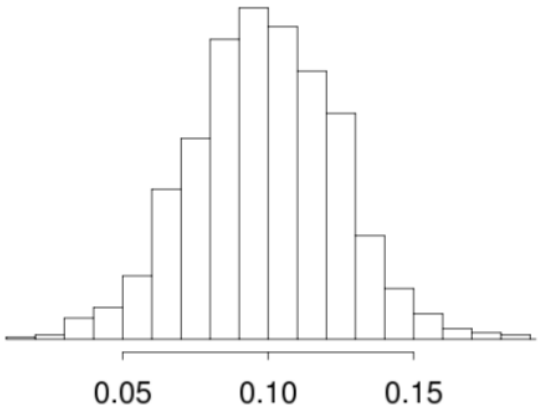
Pre-election Polling

- We consider a survey question with three possible responses. In late October, 1988, a survey was conducted by CBS News of 1447 adults in the US to find out their preferences in the upcoming presidential election.
- $y_1 = 727$ supported George Bush, $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or expressed no opinion.
- With the i.i.d. assumption of the observations, then the data (y_1, y_2, y_3) follow a multinomial distribution with parameters $(\theta_1, \theta_2, \theta_3)$.
- The estimand of interest is $\theta_1 - \theta_2$, the population difference in support for the two major candidates.

Dirichlet Example

Pre-election Polling

- Noninformative prior: $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the posterior is Dirichlet(728, 584, 138).
- The posterior distribution of $\theta_1 - \theta_2$ could be computed by integration.
- But it is simpler to draw 1000 points $(\theta_1, \theta_2, \theta_3)$ from the posterior distribution and then compute $\theta_1 - \theta_2$ for each, the results are below:



Multivariate normal model with known variance

- y is a random vector with d dimension following a multivariate normal distribution whose likelihood function for n observations would be

$$p(y_1, \ldots, y_n | \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right).$$

- Suppose Σ is known, the conjugate prior distribution for μ would be a normal distribution $\mu \sim N(\mu_0, \Lambda_0)$
- Recall your previous knowledge, what is the posterior distribution of μ?

Multivariate normal model with known variance

- The posterior distribution of μ would be

$$p(\mu|y,\Sigma) \propto \exp\left(-\frac{1}{2}\left((\mu-\mu_0)^T\Lambda_0^{-1}(\mu-\mu_0)+\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)\right),$$

- With rearranging,

$$\begin{aligned} p(\mu|y,\Sigma) &\propto \exp\left(-\frac{1}{2}(\mu-\mu_n)^T\Lambda_n^{-1}(\mu-\mu_n)\right) \\ &= \mathrm{N}(\mu|\mu_n,\Lambda_n), \end{aligned}$$

- where

$$\begin{aligned} \mu_n &= (\Lambda_0^{-1}+n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0+n\Sigma^{-1}\overline{y}) \\ \Lambda_n^{-1} &= \Lambda_0^{-1}+n\Sigma^{-1}. \end{aligned}$$

Posterior predictive distribution for new data

- For a new observation $\tilde{y} \sim N(\mu, \Sigma)$ if μ and Σ are known.
- But since μ is unknown, we first find the joint distribution of $\tilde{y}$ and μ that $p(\tilde{y}, \mu|y) = N(\tilde{y}|\mu, \Sigma)N(\mu|\mu_n, \Lambda_n)$, which means the marginal distribution of $\tilde{y}$ is also a normal distribution.

$$\begin{aligned} \mathrm{E}(\tilde{y}|y) &= \mathrm{E}(\mathrm{E}(\tilde{y}|\mu, y)|y) \\ &= \mathrm{E}(\mu|y) = \mu_n, \end{aligned}$$

$$\begin{aligned} \mathrm{var}(\tilde{y}|y) &= \mathrm{E}(\mathrm{var}(\tilde{y}|\mu, y)|y) + \mathrm{var}(\mathrm{E}(\tilde{y}|\mu, y)|y) \\ &= \mathrm{E}(\Sigma|y) + \mathrm{var}(\mu|y) = \Sigma + \Lambda_n. \end{aligned}$$

Conjugate inverse-Wishart family of prior distributions

- Recall the conjugate distribution for the univariate normal with unknown mean and variance is the normal-inverse-χ^2 distribution.
- The inverse-Wishart distribution, a multivariate generalization of the scaled inverse-χ^2, could be used to describe the prior distribution of the matrix Σ.
- The conjugate prior distribution for (μ, Σ), the normal-inverse-Wishart, has hyperparameters $(\mu_0, \Lambda_0/k_0; \nu_0, \Lambda_0)$

$$\begin{aligned} \Sigma &\sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1}) \\ \mu|\Sigma &\sim \text{N}(\mu_0, \Sigma/\kappa_0), \end{aligned}$$

Conjugate inverse-Wishart family of prior distributions

- The corresponding joint prior density is

$$p(\mu, \Sigma) \propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)\right).$$

- The parameters ν_0 and Λ_0 describe the degrees of freedom and the scale matrix for the inverse-Wishart distribution on Σ
- The remaining parameters are the prior mean μ_0 and the number of prior measurements k_0 on the Σ scale.

Conjugate inverse-Wishart family of prior distributions

- By multiplying the prior to the multivariate normal likelihood, the posterior hyperparameters would be

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\overline{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{y} - \mu_0)(\overline{y} - \mu_0)^T,
\end{aligned}$$

- where

$$S = \sum_{i=1}^{n}(y_i - \overline{y})(y_i - \overline{y})^T.$$

Conjugate inverse-Wishart family of prior distributions

Other results from the univariate normal easily generalize to the multivariate case.

- The marginal posterior distribution of μ is multivariate $t_{\nu_n - d + 1}(\mu_n, \Lambda_n / (k_n(\nu_n - d + 1)))$.
- The posterior predictive distribution of a new observation $\tilde{y}$ is also multivariate t with an additional factor of $k_n + 1$ in the numerator of the scale matrix.
- Samples from the joint posterior distribution of (μ, Σ) could be drawn by first drawing Σ from inverse-Wishart and then drawing μ from the conditional normal.

Different noninformative prior distributions

- Inverse-Wishart with $d+1$ degrees of freedom: setting $\Sigma \sim Inv - Wishart_{d+1}(\mathbf{I})$ has the appealing feature that each of the correlation in Σ has marginally a uniform prior distribution
- Inverse-Wishart with $d-1$ degrees of freedom: the prior would be

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2},$$

- The corresponding posterior is

$$\begin{aligned} \Sigma|y &\sim \text{Inv-Wishart}_{n-1}(S^{-1}) \\ \mu|\Sigma, y &\sim \text{N}(\overline{y}, \Sigma/n). \end{aligned}$$

Scaled inverse-Wishart model

- The scaled inverse-Wishart model is try to control the flexibility of the information containing in the prior distribution of Σ
- The scaled inverse-Wishart model for Σ has the form

$$\Sigma = \mathrm{Diag}(\xi)\Sigma_\eta \mathrm{Diag}(\xi),$$

in which Σ_η is given an inverse-Wishart prior distribution (one choice is Inv-$Wishart_{d+1}(\mathbf{I})$) and the scale parameter ξ can be given weakly informative prior themselves.

Analysis of a boiassay experiment

- Here is an example of a nonconjugate model for a bioassay experiment. In the development of drugs, acute toxicity tests or bioassay experiments are commonly performed on animals.
- Such experiments proceed by administering various dose levels of the compound to batches of animals. The animals' responses are typically characterized by a dichotomous outcome: for example, alive and dead, tumor or not tumor.
- An experiment of this kind fives rise to data of the form

$$(x_i, n_i, y_i); i = 1, \ldots, k,$$

where x_i represents the ith of k does level (often measured on a logarithmic scale) given to n_i animals, of which y_i subsequently respond with positive outcome.

Analysis of a boiassay experiment

- The real data is below: 20 animals were tested, five at each of four does levels.

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Table 3.1: *Bioassay data from Racine et al. (1986).*

Example: Modeling the dose-response relation

Assumptions:

- the outcomes of the five animals within each group i is i.i.d.
- the model structure of each group i is the same as

$$y_i|\theta_i \sim Bin(n_i, \theta_i)$$

- the model parameters among groups are i.i.d.
- using noninformative prior density function for θ

$$p(\theta_1, \ldots, \theta_4) \propto 1$$

  would lead to independent beta posterior distribution.
- The i.i.d. θ_i has serious flaw, the dose level x_i would be systematically related to the death probability.

Example: Modeling the dose-response relation

- The simplest model of the does-response relation is the relation of θ_i to x_i is linear that $\theta_i = \alpha + \beta x_i$.
- But since the probability should between 0 and 1, the logistic relation would be used

$$logit(\theta_i) = log \frac{\theta_i}{1-\theta_i} \alpha + \beta x_i$$

solving for θ_i gives

$$\theta_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

- Then the corresponding likelihood would be

$$p(y_i|\alpha,\beta,n_i,x_i) \propto [\text{logit}^{-1}(\alpha+\beta x_i)]^{y_i}[1-\text{logit}^{-1}(\alpha+\beta x_i)]^{n_i-y_i}.$$

Example: Modeling the dose-response relation

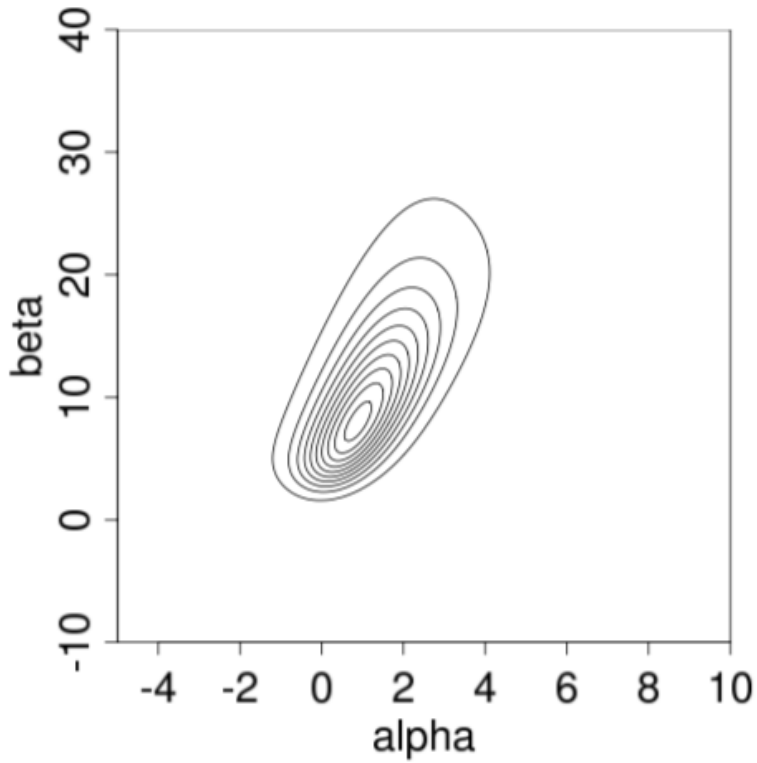
- The model is characterized by the parameters α and β, whose joint posterior distribution is

$$\begin{aligned} p(\alpha, \beta | y, n, x) &\propto p(\alpha, \beta | n, x) p(y | \alpha, \beta, n, x) \\ &\propto p(\alpha, \beta) \prod_{i=1}^{k} p(y_i | \alpha, \beta, n_i, x_i). \end{aligned}$$

We consider the sample sizes n_i and dose levels x_i as fixed for this analysis.

Example: Modeling the dose-response relation

- With the proportional posterior distribution function we can have the contour graph by letting the function equals to some constants.



Example: Prior and rough Estimation

- The prior setup for (α, β) is independent and locally uniform, that is $p(\alpha, \beta) \propto 1$.
- The rough MLE result of $(\hat{\alpha}, \hat{\beta}) = (0.8, 7.7)$ with standard error of 1.0 and 4.9 based on the data in Table 3.1.
- The corresponding log likelihood function would be

$$L(\alpha, \beta) = \sum_{i=1}^{4} -n_i log(1 + e^{\alpha+\beta x_i}) + \sum_{i=1}^{4} y_i(\alpha + \beta x_i)$$

- The estimation could be done through letting the first order derivative equal to zero.

$$\sum_{i=1}^{4}(y_i - n_i\theta_i)1 = 0$$

and

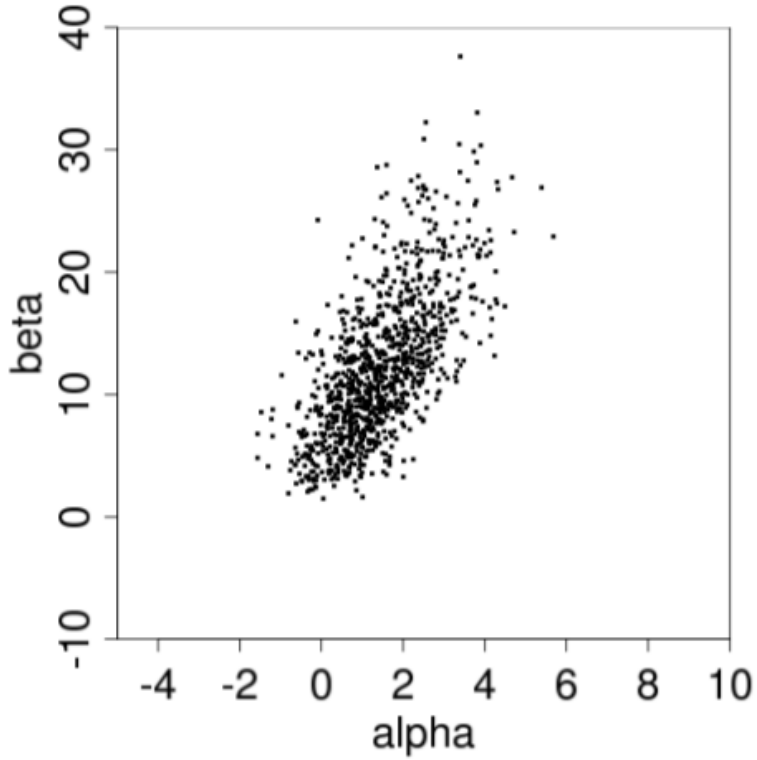
$$\sum_{i=1}^{4}(y_i - n_i\theta_i)x_i = 0$$

Example: Sampling from the joint posterior distribution

- More information related to MLE for logistic regression could be found from the Teams additional reading.
- Since we have already learned MCMC algorithm, the two-dimension Metropolis-Hastings algorithm is easy to be applied to select samples $(\alpha^{(s)}, \beta^{(s)})$.
- The sampling method mentioned in textbook just provides another option as a reference.
- (1) find the marginal distribution of $p(\alpha|y)$ by treating β as discrete variable and doing the weighted summation. (2) draw $\alpha^{(s)}$ from the marginal distribution. (3) draw $\beta^{(s)}$ from conditional distribution given $\alpha^{(s)}$. (4) randomly shift $(\alpha^{(s)}, \beta^{(s)})$ to simulate the continuous random variables.

Example: Sampling from the joint posterior distribution

- The sampling results could be shown as a scatter plot.

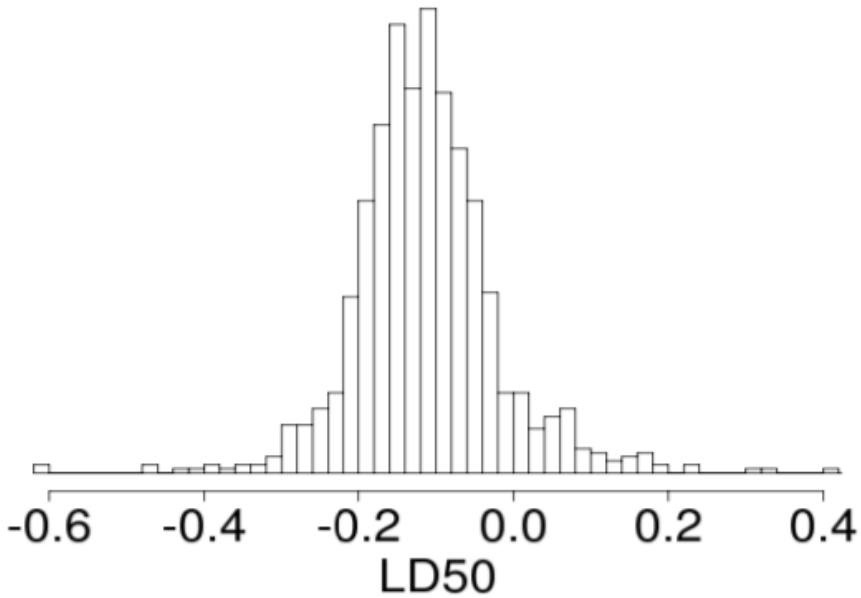


Example: The posterior distribution of the LD50

- LD50: the dose level at which the probability of death is 50%

$$\text{LD50: } E(\frac{y_i}{n_i}) = logit^{-1}(\alpha + \beta x_i) = 0.5$$

where $\alpha + \beta x_i = 0$, so $x_i = \alpha/\beta$. Through sampling:



Example: Explanation of the LD50

- if $\beta < 0$, in which case increasing the dose does not cause the probability of death to increase. It should be allowed but hard to explain.
- the posterior probability that $\beta > 0$ is the drug is harmful.
- the posterior distribution for the LD50 conditional on $\beta > 0$ is roughly estimated to exceed 0.999 according to the positive values of β from the 1000 sampling.