Bayesian Inference for Dirichlet-Multinomials

Mark Johnson

Macquarie University
Sydney, Australia

IPAM "Summer School"
Updated slides available from
http://web.science.mq.edu.au/~mjohnson/Talks.htm

MACQUARIE UNIVERSITY

Random variables and "distributed according to" notation

- A *probability distribution* F is a non-negative function whose values sum (integrate) to 1.
- A random variable X is *distributed according to* F, written $X \sim F$, iff:

$$\mathrm{P}(X = x) \; = \; F(x) \text{ for all } x$$

- You'll sometimes see the notion

$$X \mid Y \; \sim \; F$$

  which means "X is distributed conditonal on Y according to F", i.e.,

$$\mathrm{P}(X \mid Y) \; = \; F(X \mid Y).$$

Outline

Introduction to Bayesian Inference

Sampling with Markov Chains

The Gibbs sampler

Bayes' rule

$$\mathrm{P}(\mathrm{Hypothesis} \mid \mathrm{Data}) = \frac{\mathrm{P}(\mathrm{Data} \mid \mathrm{Hypothesis})\, \mathrm{P}(\mathrm{Hypothesis})}{\mathrm{P}(\mathrm{Data})}$$

- Bayesian's use Bayes' Rule to *update beliefs in hypotheses in response to data*
- $\mathrm{P}(\mathrm{Hypothesis} \mid \mathrm{Data})$ is the *posterior distribution*,
- $\mathrm{P}(\mathrm{Hypothesis})$ is the *prior distribution*,
- $\mathrm{P}(\mathrm{Data} \mid \mathrm{Hypothesis})$ is the *likelihood*, and
- $\mathrm{P}(\mathrm{Data})$ is a normalising constant sometimes called the *evidence* (often intractable to calculate)

Discrete distributions

- A *discrete distribution* has a finite set of outcomes $1, \ldots, m$
- A discrete distribution is parameterized by a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, where $\mathrm{P}(X = j | \boldsymbol{\theta}) = \theta_j$ (so $\sum_{j=1}^{m} \theta_j = 1$)
  - Example: An m-sided die, where θ_j = prob. of face j
- Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ and each $X_i | \boldsymbol{\theta} \sim \textsc{Discrete}(\boldsymbol{\theta})$. Then:

$$\mathrm{P}(\mathbf{X}|\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} \textsc{Discrete}(X_i; \boldsymbol{\theta}) \;=\; \prod_{j=1}^{m} \theta_j^{N_j}$$

  where *N_j is the number of times j occurs in $\mathbf{X}$*.
- Goal of next few slides: compute posterior distribution $\mathrm{P}(\boldsymbol{\theta}|\mathbf{X})$

Multinomial distributions

- Suppose $X_i \sim \textsc{Discrete}(\boldsymbol{\theta})$ for $i = 1, \ldots, n$, and N_j is the number of times j occurs in $\mathbf{X}$
- Then $\mathbf{N}|n, \boldsymbol{\theta} \sim \textsc{Multi}(\boldsymbol{\theta}, n)$, and

$$\mathrm{P}(\mathbf{N}|n, \boldsymbol{\theta}) \;=\; \frac{n!}{\prod_{j=1}^{m} N_j!} \prod_{j=1}^{m} \theta_j^{N_j}$$

  where $n!/\prod_{j=1}^{m} N_j!$ is the number of sequences of values with occurence counts $\mathbf{N}$
- The vector $\mathbf{N}$ is known as a *sufficient statistic* for $\boldsymbol{\theta}$ because it supplies as much information about $\boldsymbol{\theta}$ as the original sequence $\mathbf{X}$ does.

Dirichlet distributions

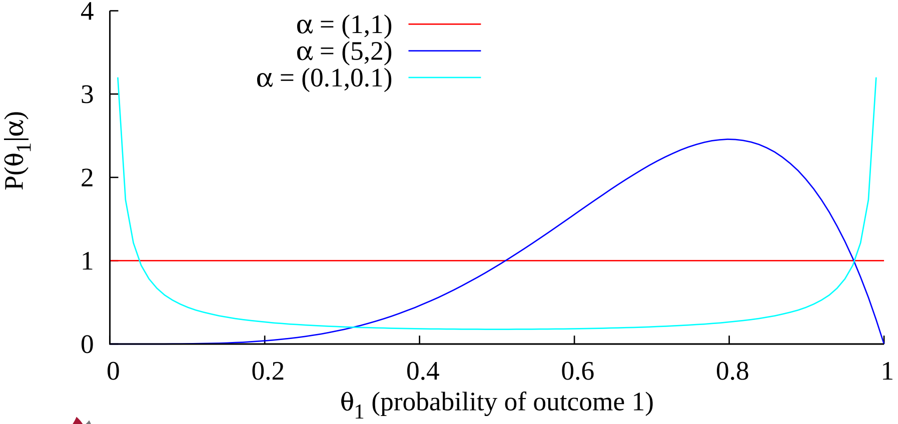
- *Dirichlet distributions* are probability distributions over multinomial parameter vectors
  - called *Beta distributions* when $m = 2$
- Parameterized by a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ where $\alpha_j > 0$ that determines the shape of the distribution

$$\begin{aligned} \mathrm{DIR}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) &= \frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \\ C(\boldsymbol{\alpha}) &= \int_{\Delta} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \, d\boldsymbol{\theta} = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)} \end{aligned}$$

- Γ is a generalization of the factorial function
- $\Gamma(k) = (k-1)!$ for positive integer k
- $\Gamma(x) = (x-1)\Gamma(x-1)$ for all x

Plots of the Dirichlet distribution

$$\mathrm{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \;=\; \frac{\Gamma(\sum_{j=1}^{m} \alpha_j)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1}$$

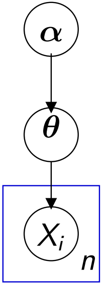


Dirichlet distributions as priors for $\boldsymbol{\theta}$

- Generative model:

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\alpha} &\sim \textsc{Dir}(\boldsymbol{\alpha}) \\ X_i \mid \boldsymbol{\theta} &\sim \textsc{Discrete}(\boldsymbol{\theta}), \quad i = 1, \ldots, n \end{aligned}$$

- We can depict this as a Bayes net using *plates*, which indicate *replication*



Inference for $\boldsymbol{\theta}$ with Dirichlet priors

- Data $\mathbf{X} = (X_1, \ldots, X_n)$ generated i.i.d. from $\textsc{Discrete}(\boldsymbol{\theta})$
- Prior is $\textsc{Dir}(\boldsymbol{\alpha})$. By Bayes Rule, posterior is:

$$
\begin{aligned}
\mathrm{P}(\boldsymbol{\theta}|\mathbf{X}) &\propto \mathrm{P}(\mathbf{X}|\boldsymbol{\theta})\,\mathrm{P}(\boldsymbol{\theta}) \\
&\propto \left(\prod_{j=1}^{m} \theta_j^{N_j}\right) \left(\prod_{j=1}^{m} \theta_j^{\alpha_j - 1}\right) \\
&= \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}, \text{ so} \\
\mathrm{P}(\boldsymbol{\theta}|\mathbf{X}) &= \textsc{Dir}(\mathbf{N} + \boldsymbol{\alpha})
\end{aligned}
$$

- So *if prior is Dirichlet* with parameters $\boldsymbol{\alpha}$, then *posterior is Dirichlet* with parameters $\mathbf{N} + \boldsymbol{\alpha}$

$\Rightarrow$ can regard Dirichlet parameters $\boldsymbol{\alpha}$ as *“pseudo-counts” from “pseudo-data”*

MACQUARIE UNIVERSITY

"Integrated out" or "collapsed" Dirichlet-multinomials

$$\begin{array}{rcll} \boldsymbol{\theta} & \mid & \boldsymbol{\alpha} & \sim \textsc{Dir}(\boldsymbol{\alpha}) \\ X_i & \mid & \boldsymbol{\theta} & \sim \textsc{Discrete}(\boldsymbol{\theta}), \quad i = 1, \ldots, n \end{array}$$

- *Integrate out* $\boldsymbol{\theta}$ to directly calculate probability of $\mathbf{X}$

$$\begin{aligned} \mathrm{P}(\mathbf{X}|\boldsymbol{\alpha}) &= \int \mathrm{P}(\mathbf{X}, \boldsymbol{\theta} \mid \alpha)\, d\boldsymbol{\theta} = \int_{\Delta} \mathrm{P}(\mathbf{X} \mid \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta} \mid \boldsymbol{\alpha})\, d\boldsymbol{\theta} \\ &= \int_{\Delta} \left(\prod_{j=1}^{m} \theta_j^{N_j} \right) \left(\frac{1}{C(\boldsymbol{\alpha})} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \right) d\boldsymbol{\theta} \\ &= \frac{1}{C(\boldsymbol{\alpha})} \int_{\Delta} \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}\, d\boldsymbol{\theta} \\ &= \frac{C(\mathbf{N} + \boldsymbol{\alpha})}{C(\boldsymbol{\alpha})}, \text{ where } C(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{m} \alpha_j)} \end{aligned}$$

Predictive distribution for Dirichlet-Multinomial

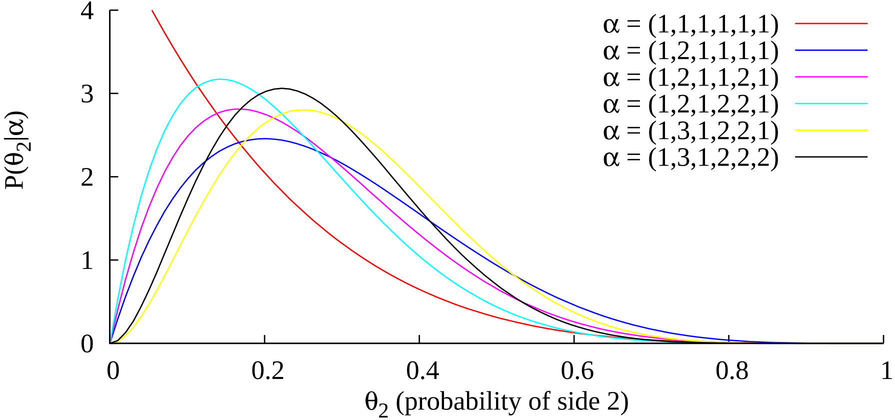
- The *predictive distribution* is the distribution of observation X_{n+1} given observations $\mathbf{X} = (X_1, \ldots, X_n)$ and prior $\mathrm{Dir}(\boldsymbol{\alpha})$

$$
\begin{aligned}
\mathrm{P}(X_{n+1} = k \mid \mathbf{X}, \boldsymbol{\alpha}) &= \int_\Delta \mathrm{P}(X_{n+1} = k \mid \boldsymbol{\theta})\, \mathrm{P}(\boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\alpha})\, d\boldsymbol{\theta} \\
&= \int_\Delta \theta_k\, \mathrm{Dir}(\boldsymbol{\theta} \mid \mathbf{N} + \boldsymbol{\alpha})\, d\boldsymbol{\theta} \\
&= \frac{N_k + \alpha_k}{\sum_{j=1}^{m} N_j + \alpha_j}
\end{aligned}
$$

MACQUARIE UNIVERSITY

Example: rolling a die

- Data $\mathbf{X} = (2, 5, 4, 2, 6)$; prior $= \textsc{Dir}((1, 1, 1, 1, 1, 1))$



MACQUARIE UNIVERSITY

Outline

MACQUARIE UNIVERSITY

Inference in complex models

- If the model is simple enough we can calculate the posterior exactly (conjugate priors)
- When the model is more complicated, we can only approximate the posterior
- *Variational Bayes* calculate the function closest to the posterior within a class of functions
- *Sampling algorithms* produce samples from the posterior distribution
  - *Markov chain Monte Carlo algorithms* (MCMC) use a Markov chain to produce samples
  - A *Gibbs sampler* is a particular MCMC algorithm
- *Particle filters* are a kind of *on-line* sampling algorithm (on-line algorithms only make one pass through the data)

Why sample?

- Setup: Model has variables $\mathbf{X}$, whose value $\mathbf{x}$ we observe, and variables $\mathbf{Y}$, whose value we don't know
  - $\mathbf{Y}$ includes any *parameters* we want to estimate, such as $\boldsymbol{\theta}$
- Goal: compute the *expected value* of some function f:

$$\mathrm{E}[f|\mathbf{X}=\mathbf{x}] \;=\; \sum_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\,\mathrm{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x})$$

- Suppose we can produce n samples $\mathbf{y}^{(t)}$, where $\mathbf{Y}^{(t)} \sim \mathrm{P}(\mathbf{Y} \mid \mathbf{X}=\mathbf{x})$. Then we can estimate:

$$\mathrm{E}[f|\mathbf{X}=\mathbf{x}] \;=\; \frac{1}{n}\sum_{t=1}^{n} f(\mathbf{x},\mathbf{y}^{(t)})$$

- Example: word-tagging. $\mathbf{X}$ is vector of words, $\mathbf{Y}$ is vector of tags. Set $f(\mathbf{x},\mathbf{y}) = 1$ if $y_1 = \mathrm{Noun}$, and zero otherwise. Then $\mathrm{E}[f|\mathbf{X}=\mathbf{x}]$ is prob. that word x_1 is tagged Noun.

Markov chains

- A (first-order) *Markov chain* is a distribution over random variables $S^{(0)}, \ldots, S^{(n)}$ all ranging over the same *state space* $\mathcal{S}$, where:

$$\mathrm{P}(S^{(0)}, \ldots, S^{(n)}) \;=\; \mathrm{P}(S^{(0)}) \prod_{t=0}^{n-1} \mathrm{P}(S^{(t+1)} | S^{(t)})$$

  $S^{(t+1)}$ is *conditionally independent* of $S^{(0)}, \ldots, S^{(t-1)}$ given $S^{(t)}$
- A Markov chain in *homogeneous* or *time-invariant* iff:

$$\mathrm{P}(S^{(t+1)} = s' | S^{(t)} = s) \;=\; P_{s',s} \quad \text{for all } t, s, s'$$

  The matrix P is called the *transition probability matrix* of the Markov chain
- If $\mathrm{P}(S^{(t)} = s) = \pi_s^{(t)}$ (i.e., $\pi^{(t)}$ is a vector of state probabilities at time t) then:
  - $\pi^{(t+1)} \;=\; P\,\pi^{(t)}$
  - $\pi^{(t)} \;=\; P^t\,\pi^{(0)}$

Ergodicity

- A Markov chain with tpm P is *ergodic* iff there is a positive integer m s.t. all elements of P^m are positive (i.e., there is an m-step path between any two states)
- Informally, an ergodic Markov chain "forgets" its past states
- Theorem: For each homogeneous ergodic Markov chain with tpm P there is a *unique limiting distribution* D_P, i.e., as n approaches infinity, the distribution of S_n converges on D_P
- D_P is called the *stationary distribution* of the Markov chain

Using a Markov chain for inference of $\mathrm{P}(Y)$

- Set the state space $\mathcal{S}$ of the Markov chain to the range of $\mathbf{Y}$ ($\mathcal{S}$ may be *astronomically large*)
- Find a tpm P such that $\mathrm{P}(\mathbf{Y} \mid \mathbf{X}) = D_P$
- "Run" the Markov chain, i.e.,
  - Pick $\mathbf{y}^{(0)}$ somehow
  - For $t = 0, 1, \ldots$:
    - sample $\mathbf{y}^{(t+1)}$ from $\mathrm{P}(\mathbf{Y}^{(t+1)} \mid \mathbf{Y}^{(t)}{=}\mathbf{y}^{(t)}, \mathbf{X}{=}\mathbf{x})$, i.e., from $P_{\cdot,\mathbf{y}^{(t)}}$
  - After discarding the first *burn-in* samples, use remaining samples to calculate statistics
- *WARNING:* in general the samples $\mathbf{y}^{(t)}$ are *not independent*

MACQUARIE UNIVERSITY

Outline

Introduction to Bayesian Inference

Sampling with Markov Chains

The Gibbs sampler

MACQUARIE UNIVERSITY

The Gibbs sampler

- The Gibbs sampler is useful when:
  - $\mathbf{Y}$ is multivariate, i.e., $\mathbf{Y} = (Y_1, \ldots, Y_m)$, and
  - easy to sample from $\mathrm{P}(Y_j | \mathbf{Y}_{-j})$
- The *Gibbs sampler* for $\mathrm{P}(Y)$ is the tpm $P = \prod_{j=1}^m P^{(j)}$, where:

$$P^{(j)}_{\mathbf{y}',\mathbf{y}} = \begin{cases} 0 & \text{if } \mathbf{y}'_{-j} \neq \mathbf{y}_{-j} \\ \mathrm{P}(Y_j = y'_j | \mathbf{Y}_{-j} = \mathbf{y}_{-j}) & \text{if } \mathbf{y}'_{-j} = \mathbf{y}_{-j} \end{cases}$$

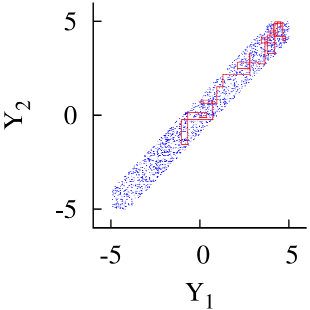
- Informally, *the Gibbs sampler cycles through each of the variables Y_j, replacing the current value y_j with a sample from $\mathrm{P}(Y_j | \mathbf{Y}_{-j} = \mathbf{y}_{-j})$*
- There are *sequential scan* and *random scan* variants of Gibbs sampling

A simple example of Gibbs sampling

$$\mathrm{P}(Y_1, Y_2) = \begin{cases} c & \text{if } |Y_1| < 5, |Y_2| < 5 \text{ and } |Y_1 - Y_2| < 1 \\ 0 & \text{otherwise} \end{cases}$$

- The Gibbs sampler for $\mathrm{P}(Y_1, Y_2)$ samples repeatedly from:

$$\mathrm{P}(Y_2|Y_1) = \textsc{Uniform}(\max(-5, Y_1 - 1), \min(5, Y_1 + 1))$$
$$\mathrm{P}(Y_1|Y_2) = \textsc{Uniform}(\max(-5, Y_2 - 1), \min(5, Y_2 + 1))$$



Sample run

Y_1	Y_2
0	0
0	-0.119
0.363	-0.119
0.363	0.146
-0.681	0.146
-0.681	-1.551

MACQUARIE UNIVERSITY

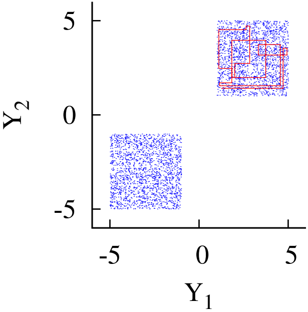
A non-ergodic Gibbs sampler

$$\mathrm{P}(Y_1, Y_2) = \begin{cases} c & \text{if } 1 < Y_1, Y_2 < 5 \text{ or } -5 < Y_1, Y_2 < -1 \\ 0 & \text{otherwise} \end{cases}$$

- The Gibbs sampler for $\mathrm{P}(Y_1, Y_2)$, initialized at (2,2), samples repeatedly from:

$$\begin{aligned} \mathrm{P}(Y_2 | Y_1) &= \textsc{Uniform}(1,5) \\ \mathrm{P}(Y_1 | Y_2) &= \textsc{Uniform}(1,5) \end{aligned}$$

I.e., *never visits the negative values of Y_1, Y_2*



Sample run

Y_1	Y_2
2	2
2	2.72
2.84	2.72
2.84	4.71
2.63	4.71

MACQUARIE UNIVERSITY

Why does the Gibbs sampler work?

- The Gibbs sampler tpm is $P = \prod_{j=1}^{m} P^{(j)}$, where $P^{(j)}$ replaces y_j with a sample from $\mathrm{P}(Y_j|\mathbf{Y}_{-j} = \mathbf{y}_{-j})$ to produce y'
- But if $\mathbf{y}$ is a sample from $\mathrm{P}(\mathbf{Y})$, then so is $\mathbf{y}'$, since $\mathbf{y}'$ differs from $\mathbf{y}$ only by replacing y_j with a sample from $\mathrm{P}(Y_j|\mathbf{Y}_{-j} = \mathbf{y}_{-j})$
- Since $P^{(j)}$ maps samples from $\mathrm{P}(\mathbf{Y})$ to samples from $\mathrm{P}(\mathbf{Y})$, so does P

$\Rightarrow$ $\mathrm{P}(\mathbf{Y})$ is a stationary distribution for P

- If P is ergodic, then $\mathrm{P}(\mathbf{Y})$ is the unique stationary distribution for P, i.e., the sampler converges to $\mathrm{P}(\mathbf{Y})$

Summary

- Dirichlet-multinomial distributions can be handled largely analytically
- Complex models often don't have analytic solutions
- Approximate inference can be used on many such models
- Monte Carlo Markov chain methods produce samples from (an approximation to) the posterior distribution
- Gibbs sampling is an MCMC procedure that resamples each variable conditioned on the values of the other variables
- If you can sample from the conditional distribution of each hidden variable in a Bayes net, you can use Gibbs sampling to sample from the joint posterior distribution