Bayesian Inference and Decision Theory

Unit 9: Conclusion: Multinomial Distribution and Latent Groups

Learning Objectives for Unit 9

- Find the posterior distribution for the probability vector when observations are randomly sampled from a multinomial distribution and the prior distribution for the probability vector is a Dirichlet distribution.
- Define a latent variable and describe how Bayesians treat latent variables
- Describe the label switching problem in MCMC models with latent variables
- Summarize basic ideas of Bayesian approach to inference and decision-making

Running Example: Inventory Management and Sales Prediction

- We will review many of the fundamental ideas of this course through a simplified inventory management and sales prediction example
- We will apply the methods we learned in this course to a series of increasingly complex decision-making and prediction problems
  - Predict future sales of each product given data on past sales
  - Decide how much of each product to stock to maximize profits
  - Use hierarchical model to predict sales for sub-groups of customers
  - Use hierarchical model to discover clusters of customers with similar buying patterns
- To address this example we will first introduce a new conjugate pair to generalize the binomial / beta conjugate pair to problems with more than two categories

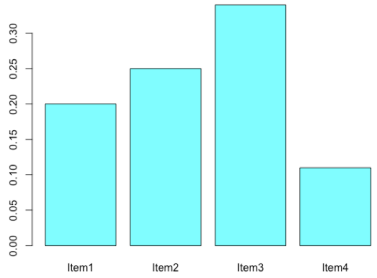
Categorical Data and the Multinomial Distribution

- Categorical data consists of observations falling into one of a finite number m of categories
  - Each patient has one of m diseases
  - Each customer purchase consists of one of m products
  - Each sampled organism belongs to one of m taxa
  - Each sampled word in a document is one of m possible words
- The *multinomial* distribution generalizes the binomial distribution to more than two categories
  - Parameters: probabilities $\Theta_1, \ldots, \Theta_m$, where $\sum_i \Theta_i$=1
  - Observation: $X_1, \ldots, X_m$ is a vector of counts of cases in each category, where $\sum_i X_i = n$ is the total count
  - Likelihood function $f(x_1, \ldots, x_m | \theta_1, \ldots, \theta_m) = \left(\frac{n!}{x_1! \cdots x_m!}\right) \theta_1^{x_1} \cdots \theta_m^{x_m}$

Dirichlet Distributions: A Conjugate Family to the Multinomial Family of Distributions

$(\Theta_1, \ldots, \Theta_m)$ has a Dirichlet distribution with shape parameters $\alpha_1, \ldots, \alpha_m$, all $\alpha_i > 0$:

- Sample space: Real positive numbers that sum to 1
- pdf: $\frac{\Gamma(\alpha_1+\cdots+\alpha_m)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_m)} \theta_1^{\alpha_1-1} \cdots \theta_m^{\alpha_m-1}$
- $E[\Theta_i|\alpha_1, \ldots, \alpha_m] = \hat{\theta}_i = \frac{\alpha_i}{\sum_i \alpha_i}$
- $\text{Var}[\Theta_i|\alpha_1, \ldots, \alpha_m] = \frac{\hat{\theta}_i(1-\hat{\theta}_i)}{(\sum_i \alpha_i+1)}$ $\quad\quad$ $\text{Cov}[\Theta_i, \Theta_j|\alpha_1, \ldots, \alpha_m] = \frac{-\hat{\theta}_i\hat{\theta}_j}{(\sum_i \alpha_i+1)}$



- Dirichlet distribution is a multivariate generalization of the Beta distribution
  - We call α_i the *virtual count* for category i
  - Marginal distribution for Θ_i is beta($\alpha_i, \sum_{j\neq i} \alpha_j$)
  - Dirichlet distribution with $(\alpha_1, \ldots, \alpha_m)$=(1,…,1) is the uniform distribution, putting equal density on all $\theta_1, \ldots, \theta_m$ with $\sum_i \theta_i$=1
    - If m>2, the uniform distribution on $(\Theta_1, \ldots, \Theta_m)$ is *not* uniform on any Θ_i
    - Example: for four categories, if $\theta_1, \ldots, \theta_4$ has uniform distribution, then θ_1 has Beta(1,3) distribution

Examples of Dirichlet distributions over $\mathbf{p} = (p_1, p_2, p_3)$ which can be plotted in 2D since $p_3 = 1 - p_1 - p_2$:

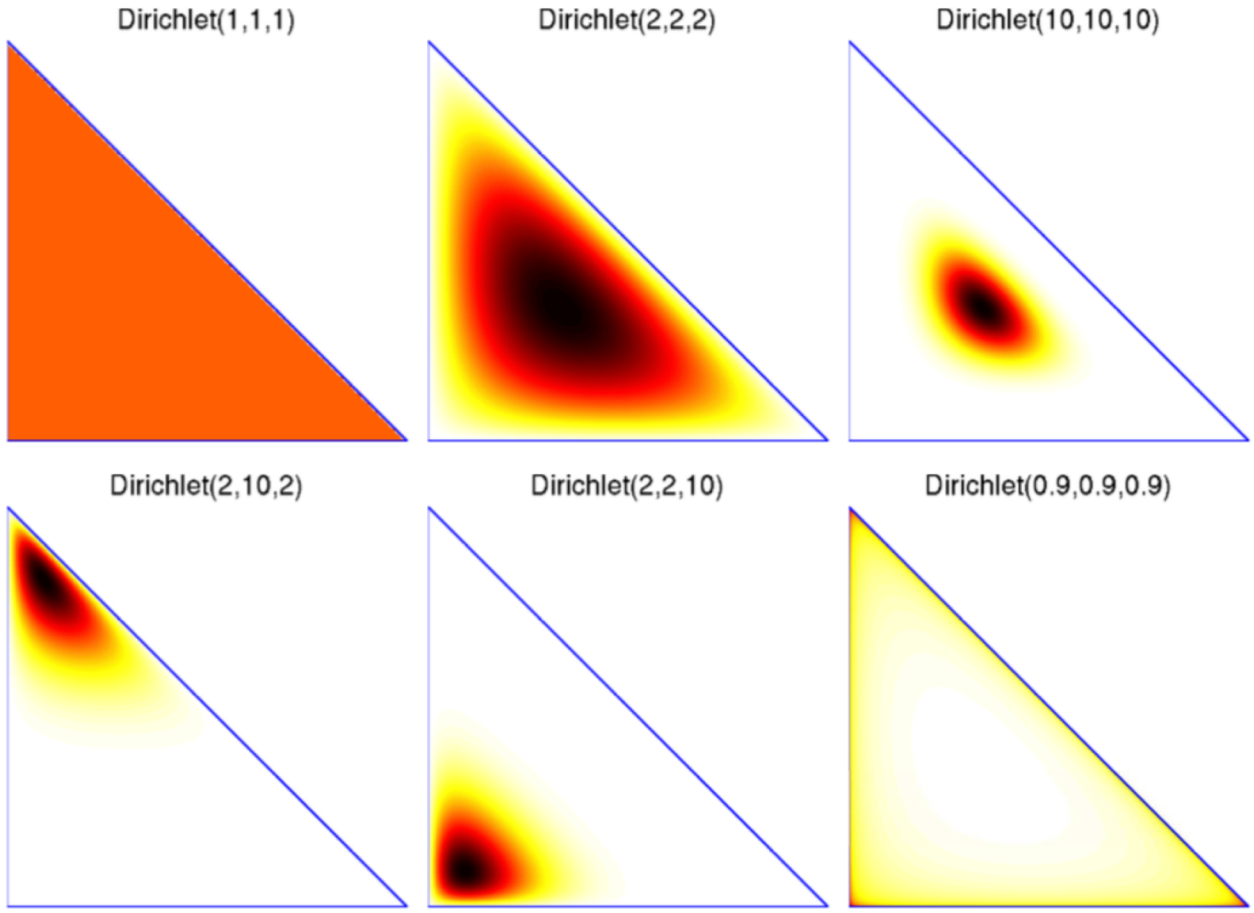


Figure taken from http://mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf

The Multinomial / Dirichlet Conjugate Pair

- The multinomial and Dirichlet families of distributions are a *conjugate pair*:

  IF Observations $\underline{X}_1, \ldots, \underline{X}_k = (X_{11}, \ldots, X_{1m}), \ldots, (X_{k1}, \ldots, X_{km})$ are a random sample of counts drawn from a multinomial distribution with probability vector $(\Theta_1, \ldots, \Theta_m)$ and the prior distribution for $\underline{\Theta}$ is Dirichlet$(\alpha_1, \ldots, \alpha_m)$

  THEN Posterior distribution for Θ is Dirichlet$(\alpha_1^* \cdots \alpha_m^*)$, another member of the conjugate family, where $\alpha_j^* = \alpha_j + \sum_{i=1}^{k} X_{ij}$

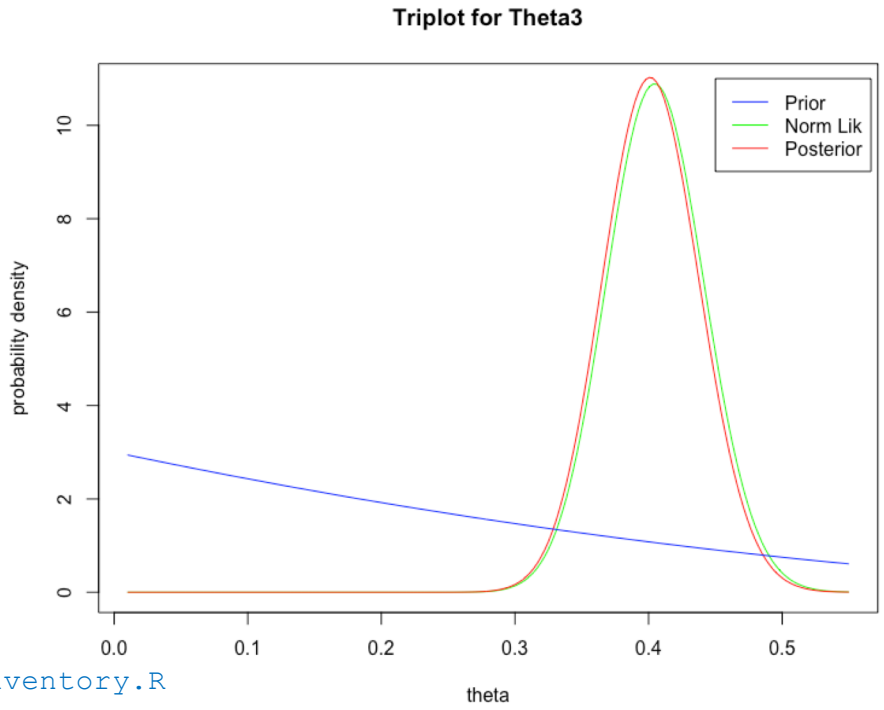
- The posterior virtual count for category j is the sum of the prior virtual count α_j and the k observed counts $(X_{1j}, \ldots, X_{kj})$ for category j

Example: Inventory Management

- A company needs to decide how much of each of four products to stock per time period
- The company's data analytics team models sales as follows:
  - The number of customers per time period has a Poisson distribution with rate $\lambda = 1000$ customers per period
  - Each customer orders one of the four products
  - Customer orders are are modeled as a multinomial distribution with probability $(\Theta_1, \Theta_2, \Theta_3, \Theta_4)$
  - If the product is in stock, the customer walks away with it; otherwise a rush order is placed for the item and it is delivered to the customer the next day
- The company's utility function is the sum of:
  - Profit of \$20 for each sale
  - Cost of \$2 for each item in inventory that is not sold
  - Cost of \$15 for each rush order
- The company assumes a uniform prior distribution on $(\Theta_1, \Theta_2, \Theta_3, \Theta_4)$
- Sales data has been collected for a random sample of 180 customers
- Given this model and the data, what is the optimal inventory for each product?

Inventory Management: Posterior Distribution for Item Choice Probabilities

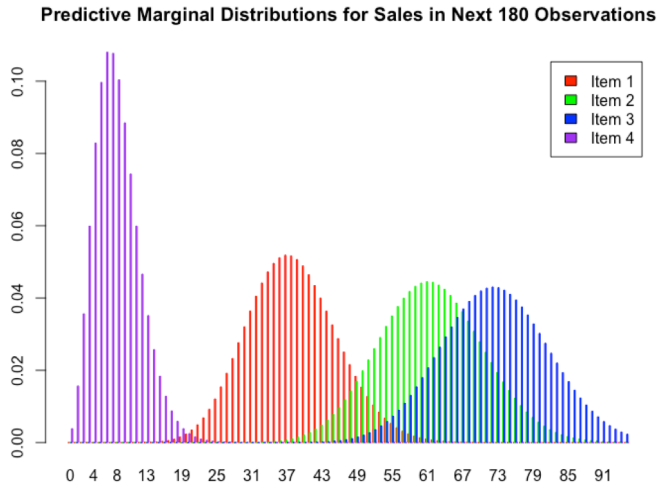
- Prior distribution for $(\Theta_1, \Theta_2, \Theta_3, \Theta_4)$ is Dirichlet(1,1,1,1)
- Observations (X_1, X_2, X_3, X_4) = (38, 62, 73, 7)
- Posterior distribution for $(\Theta_1, \Theta_2, \Theta_3, \Theta_4)$ is Dirichlet(39, 63, 74, 8)
  - Θ_i has a beta distribution with shape parameters X_i+1 and $\sum_{j \neq i} X_j$ +3
  - $E[(\Theta_1, \Theta_2, \Theta_3, \Theta_4)|(X_1, X_2, X_3, X_4)] = (0.212, 0.342, 0.402, 0.043)$
  - $SD[(\Theta_1, \Theta_2, \Theta_3, \Theta_4)|(X_1, X_2, X_3, X_4)] = (0.030, 0.035, 0.036, 0.015)$
  - 90% credible intervals for proportions:
    - Θ_1 : [0.164, 0.263] Θ_2 : [0.286, 0.401]
    - Θ_3 : [0.343, 0.462] Θ_4 : [0.022, 0.071]

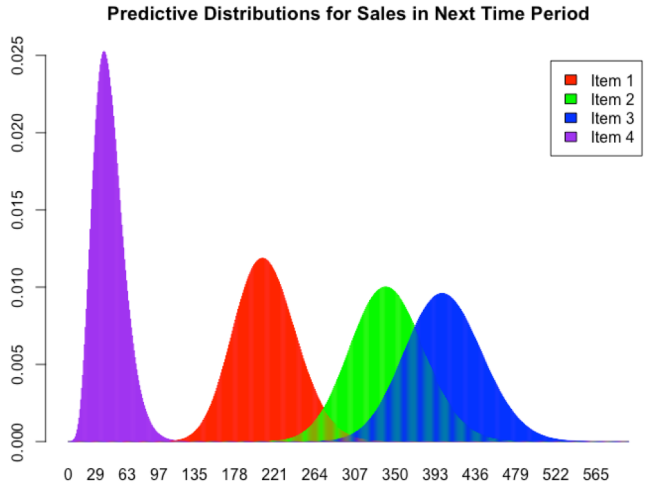


R code can be found in `DirichletExampleOptimizeInventory.R`

Predictive Distribution for Sales

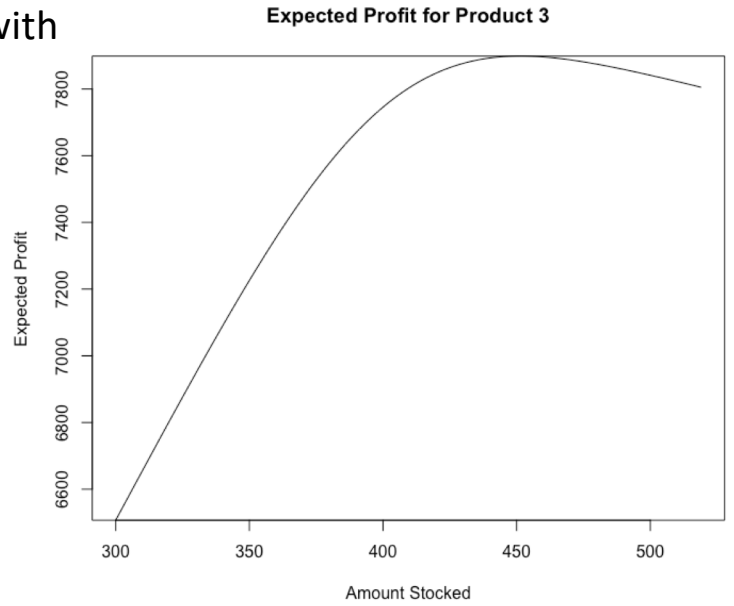
- Predictive distribution for another sample of 180:
  - Marginal distribution for number of sales in category i is beta-binomial with size 180, probability $\hat{\theta}_i = \frac{X_i+1}{184}$ and overdispersion $\sum_j \alpha_j^* = 184$
  - Joint distribution for all categories is Dirichlet-multinomial marginal likelihood
- To predict sales in each category in next time period
  - Total sales Y have Poisson distribution with mean 1000
  - Given total sales Y, sales X_i in category i are beta-binomial with size Y, probability $\hat{\theta}_i = \frac{X_i+1}{184}$ and overdispersion $\sum_j \alpha_j^* = 184$
  - Marginalize out Y to find predictive distribution for sales in each category





Optimal Inventory

- Step 1: Find predictive distribution of sales for product i:
  - Total sales Y have Poisson distribution with mean 1000
  - Given total sales Y, sales X_i in category i are beta-binomial with size Y, probability $\hat{\theta}_i = \frac{X_i+1}{184}$ and overdispersion $\sum_j \alpha_j^* = 184$
  - Marginalize out Y to find predictive distribution $f(x) = P(X_i = x)$ for sales of product i
- Step 2: Find expected utility for stocking r_i items in category i:
  - For each value x_i for X_i calculate net gain:
    - Gain of $20x_i$ from profit on sales
    - Loss of $15(x_i - r_i)$ from rush orders if $x_i > r_i$
    - Loss of $2(r_i - x_i)$ from excess inventory if if $x_i < r_i$
  - Multiply $(20x_i - 15(x_i - r_i)1_{[x_i > r_i]} - 2(r_i - x_i)1_{[x_i < r_i]})$ times predictive pmf $f(x_i)$ and sum all values
- Step 3: Choose r_i to maximize expected utility



Optimization Results

- Optimal inventory is (R_1, R_2, R_3, R_4) = (252, 390, 451,63)
- Expected sales are $E(X_1, X_2, X_3, X_4)$ = (211.9, 342.4, 402.1, 43.5)
  - Stock more items than expected sales to protect against cost of rush orders
  - Using binomial point estimates instead of beta-binomial predictive probabilities yields recommend inventory of (229, 364, 426, 51)
    - Less overstocking
    - Slightly sub-optimal solution
- Expected profit is 19535
  - Expected profit using best solution under binomial predictive distribution is 19422 (this is more than 99% of optimal profit)
  - Binomial model estimates profit at 19789 (about 3% overestimate)

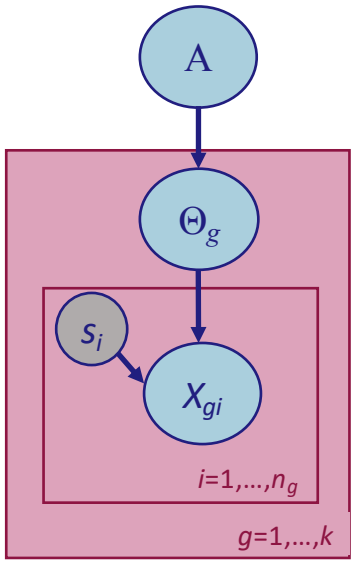
Extension: Hierarchical Multinomial Model

- There may be count data from multiple groups, each with its own sales distribution
  - Multinomial observations with group-dependent probabilities of purchasing different products
- Hierarchical model allows groups to share information
  - Top level is hyperparameter

    $(\alpha_1, \ldots, \alpha_m) \sim g(\underline{\alpha})$
  - Next level is parameters for groups

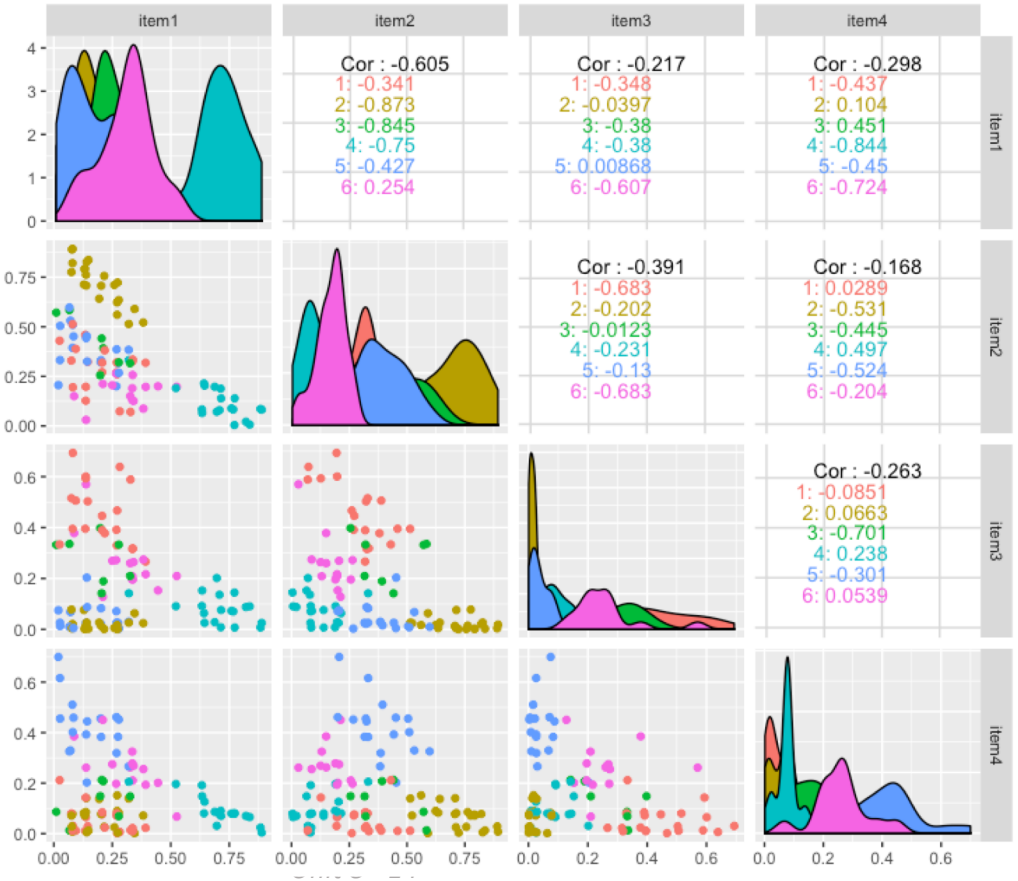
    $(\Theta_{g1}, \ldots, \Theta_{gm}) \sim_{\text{iid}} \text{Dirichlet}(\alpha_1, \ldots, \alpha_m)$
  - Bottom level are observations of sales to customers

    $(X_{g1i}, \ldots, X_{gmi}) \sim_{\text{iid}} \text{multinomial}(\Theta_{g1}, \ldots, \Theta_{gm}, s_i)$



A Sample Data Set

- 100 customers in six groups
- Each customer made 15 product choices
  - Selections were generated randomly according to group-specific probabilities
- Plot matrix shows frequencies of item selections for each customer color-coded by group



R code can be found in `DirichletExampleDataGen.R`

Fitting the Hierarchical Dirichlet-Multinomial Model in JAGS

The JAGS Model:

```
model{
  for (i in 1:numObs) {
    choice[i] ~ dcat(theta[grp[i],1:numItems]) # Counts of selections
  }
  for (i in 1:numGrp) {
    theta[i,1:numItems] ~ ddirch(vcounts[1:numItems]) # Dirichlet prior on choice prob
  }
  alphaitm <- rep(1,numItems)
  mu[1:numItems] ~ ddirch(alphaitm[1:numItems])  # uniform prior on item probabilities
  for (i in 1:numItems) {
    vcounts[i] <- mu[i]*conc      # prior on category virtual counts
  }
  conc ~ dgamma(1,0.1)             # gamma prior on total virtual count
  }
```

Data:

- `choice` – vector of product choices made by customers
- `grp` – vector of groups to which customers belong
- `numObs, numGrp, numItems` – total number of observations, number of groups, number of items customers can choose

Parameters:

- `mu` – grand mean of item probabilities
- `conc` – sum of virtual counts
- `theta` – matrix of group x item probabilities

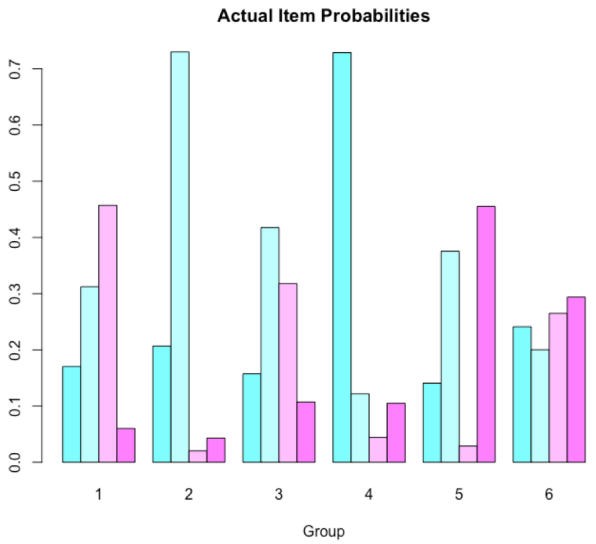
To Run The Model from R:

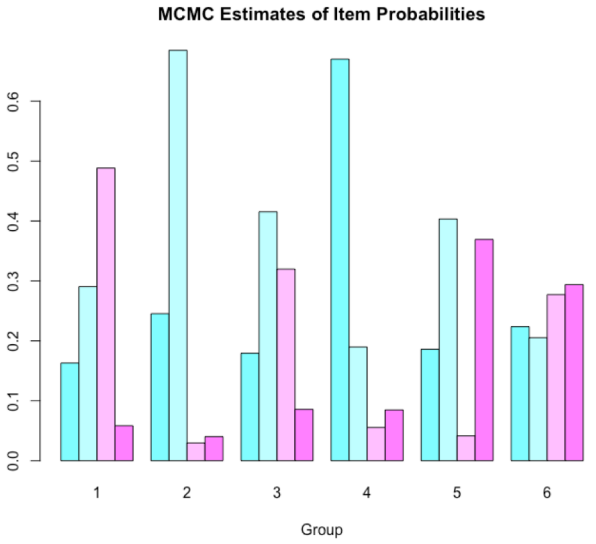
```
numSim=5000
numBurnin=floor(1000)

numObs=numUnits*numChoices
theta=array(dim=c(numGrp,numItems))   # Initialize theta as array
response.data=c("choice","grp","numObs","numGrp","numItems")
response.inits = function() {
  list( "mu"=array(1/numItems,numItems),
        "conc"=20, "pi"=array(1/numGrp,numGrp))
}
response.inits=NULL
response.params=c("mu","conc","theta")

response.fit <- jags(response.data,inits=response.inits,response.params,
                     model.file="Dirichlet.model.KnownGroups.jags",
                     n.iter=numSim,n.burnin=numBurnin,n.chains=2)
```

R code can be found in `DirichletExampleKnownGroups.R`

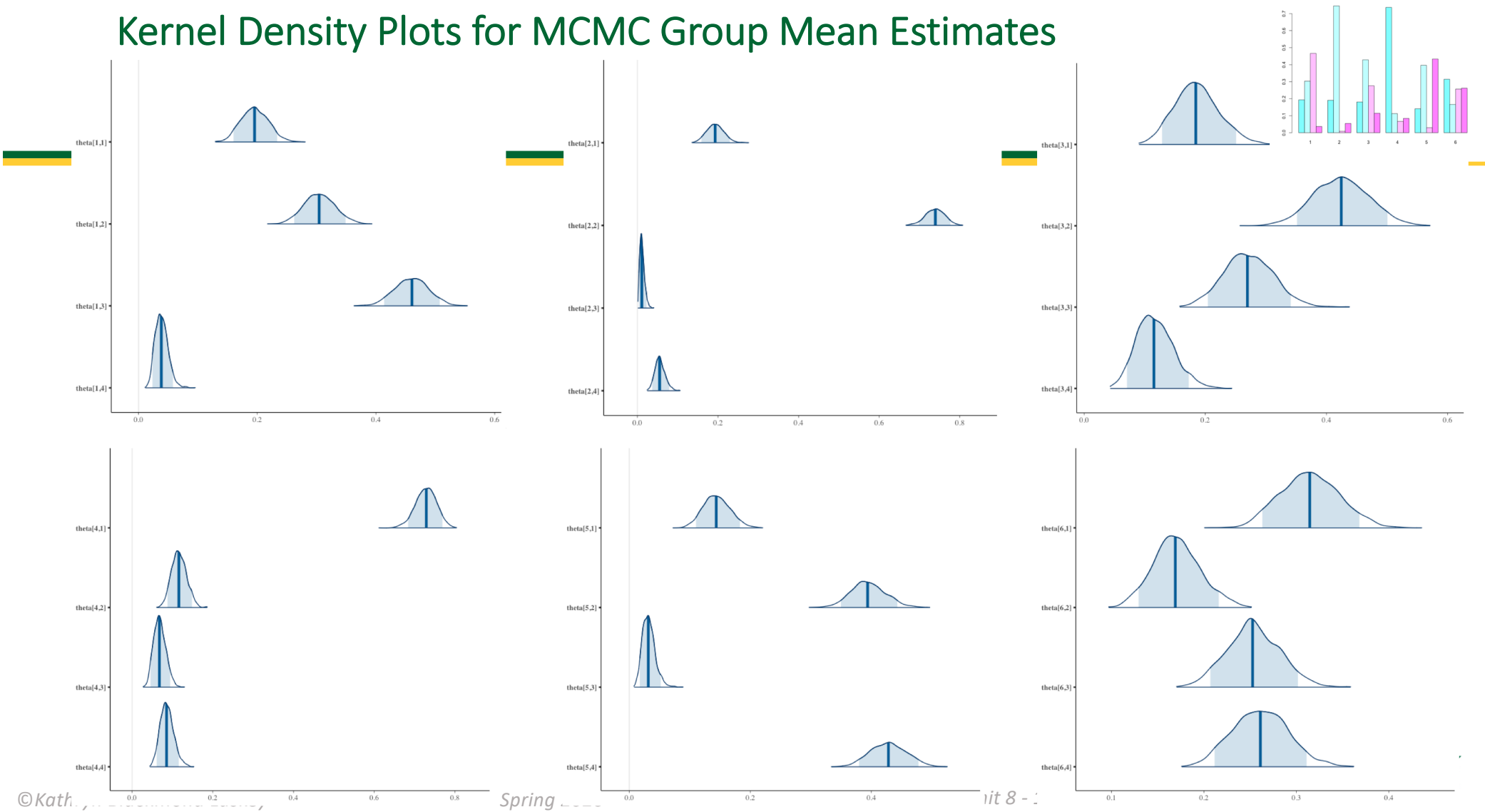
Results of Fitting the Model to Sample Data Set

- The data:
  - 100 customers in six groups
  - 15 choices by each customer according to group-specific probabilities
- Fitting the model:
  - 5000 iterations, 2 chains, burnin 1000
  - Used default thinning interval of 4
  - 1000 iterations saved per chain
  - Effective sample sizes of Θ_{gi} samples (according to `effectiveSize` function) range from 1738 to 2000
  - Trace plots look fairly stationary
- Results:
  - Estimates are fairly close to actual probabilities
  - MCMC estimate of total virtual count is 6.78
  - MCMC estimate of mean virtual count vector is (1.99, 2.47, 1.18, 1.15)



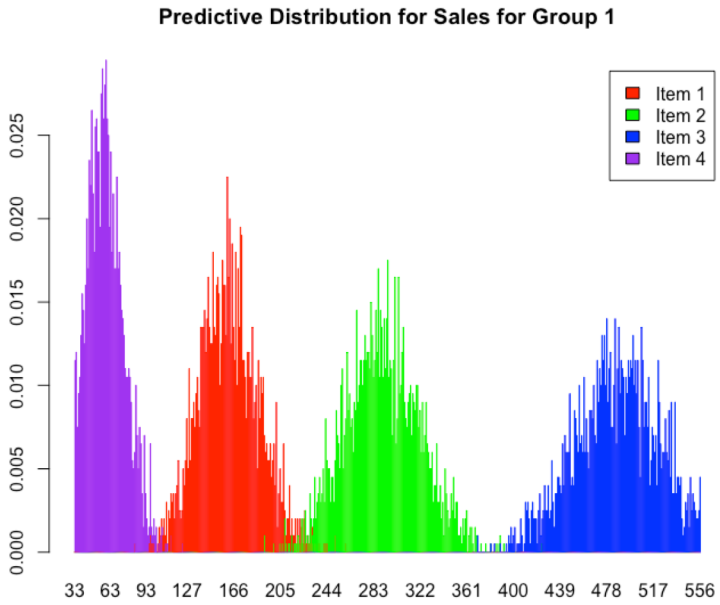


Kernel Density Plots for MCMC Group Mean Estimates



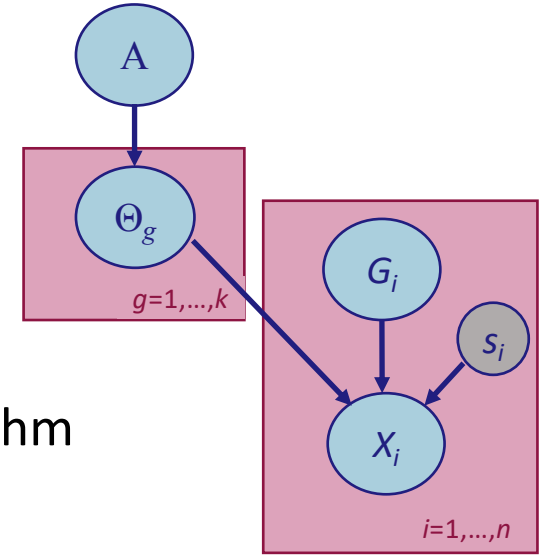
Predicting Sales by Group

- We can use our MCMC estimates to predict sales for each item in each group of customers
  - For each realization k of the MCMC sample
    - Generate random number n of purchases from customer arrival distribution
    - Use multinomial probabilities $\underline{\Theta}_g^{(k)}$ from group g and realization k to allocate the n purchases to items
  - Tally up total purchases of each item in each group
  - Normalize to sum to 1 to obtain predictive distribution for sales of that item in that group



Extension: Discovering Latent Groups

- In many applications, we are not given group labels for the observations, but would like to discover groups from the data
  - Medical applications: would like to design treatments that work for similar clusters of patients
  - Recommending systems: would like to tailor recommendations to clusters of similar customers
  - Applications in ecology: would like to discover clusters of similar organisms
- We will try removing the group labels and seeing if the algorithm can discover the groups
  - Group labels are *latent variables* – not directly observed but inferred through a model from other observed variables
  - To infer the group labels, we need to define a prior distribution for them
  - Then we use Bayesian inference to discover the group labels



Fitting the Latent Group Model in JAGS

The JAGS Model:

```
model{
  for (i in 1:numObs) {
    choice[i] ~ dcat(theta[grp[unit[i]],1:numItems]) # Counts of selections
  }
  for (u in 1:numUnits) {
    grp[u] ~ dcat(pi[1:numGrp])
  }
  for (g in 1:numGrp) {
    theta[g,1:numItems] ~ ddirch(vcounts[1:numItems]) # Dirichlet prior on choice prob
  }
  alphagrp <- rep(1,numGrp)
  pi[1:numGrp] ~ ddirch(alphagrp[1:numGrp])      # uniform prior on group membership probs
  alphaitm <- rep(1,numItems)
  mu[1:numItems] ~ ddirch(alphaitm[1:numItems])  # uniform prior on item probabilities
  for (i in 1:numItems) {
    vcounts[i] <- mu[i]*conc      # prior on category virtual counts
  }
  conc ~ dgamma(1,0.1)             # gamma prior on total virtual count
  }
```

Data:

- `choice` – vector of product choices made by customers
- `numObs, numGrp, numItems` – total number of observations, number of groups, number of items customers can choose
- `unit, numUnits` – customer ID and number of customers

Parameters:

- `mu` – grand mean of item probabilities
- `conc` – sum of virtual counts
- `theta` – matrix of group x item probabilities
- `grp` – latent group memberships
- `pi` – group membership probabilities

To Run The Model from R:

```
numSim=5000              # run simulation for 5000 itertions
numBurnin=floor(1000)    # discard 1000 samples for burnin

numObs=numUnits*numChoices
theta=array(dim=c(numGrp,numItems))   # Initialize theta as array
response.data=c("choice","unit","numObs","numGrp","numUnits","numItems")
response.inits = function() {
  list( "mu"=array(1/numItems,numItems),
        "conc"=20, "pi"=array(1/numGrp,numGrp))
}
response.inits=NULL
response.params=c("mu","conc","theta","grp","pi")

response.fit <- jags(response.data,inits=response.inits,response.params,
                     model.file="Dirichlet.model.UnknownGroups.jags",
                     n.iter=numSim,n.burnin=numBurnin,n.chains=2)
```

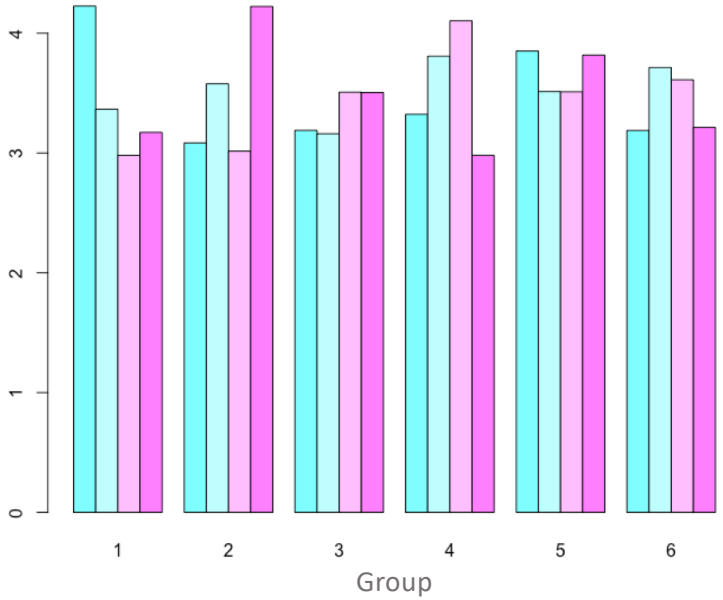
R code can be found in `DirichletExampleUnknownGroups.R`

Results of Fitting the Latent Groups Model to Example Data Set

Well, that didn't work!

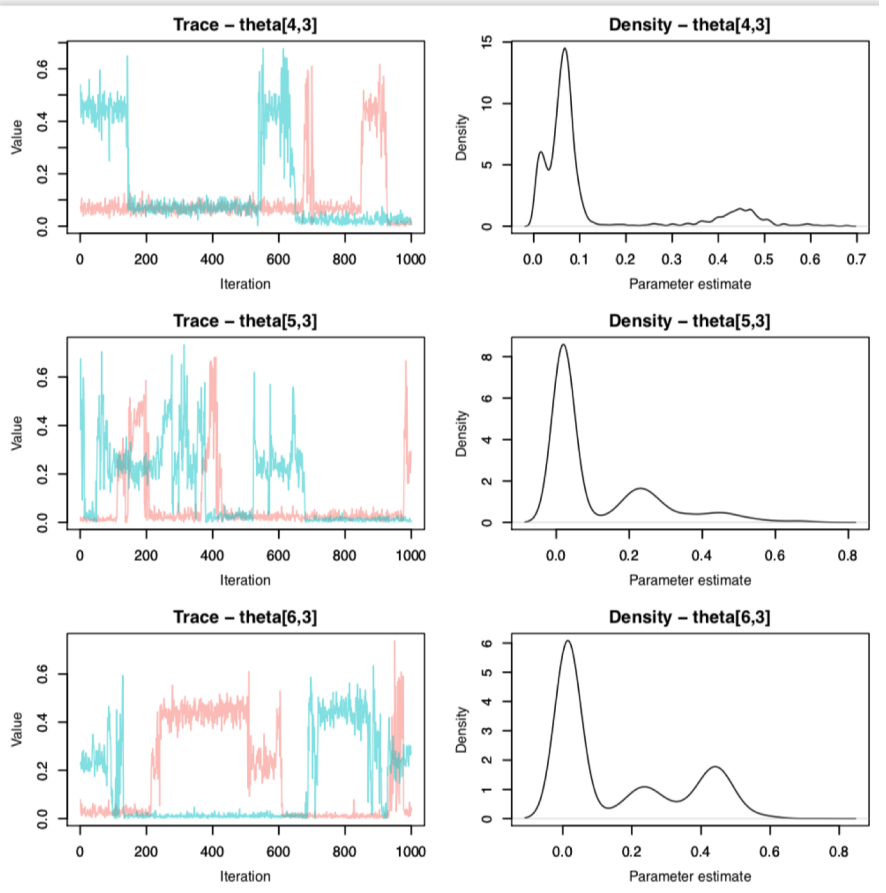
- 5000 iterations, 2 chains, burnin 1000
- Used default thinning interval of 4
- 1000 iterations saved per chain
- Effective sample sizes of Θ_{gi} samples (according to `effectiveSize` function) range from 6.1 to 151.0
- Trace plots of Θ_{gi} look very non-stationary
- Estimates of means of Θ_{gi} do not vary much across groups
- Kernel density estimates of Θ_{gi} are multi-modal

MCMC Point Estimates of Item Probabilities by Group (latent group model)



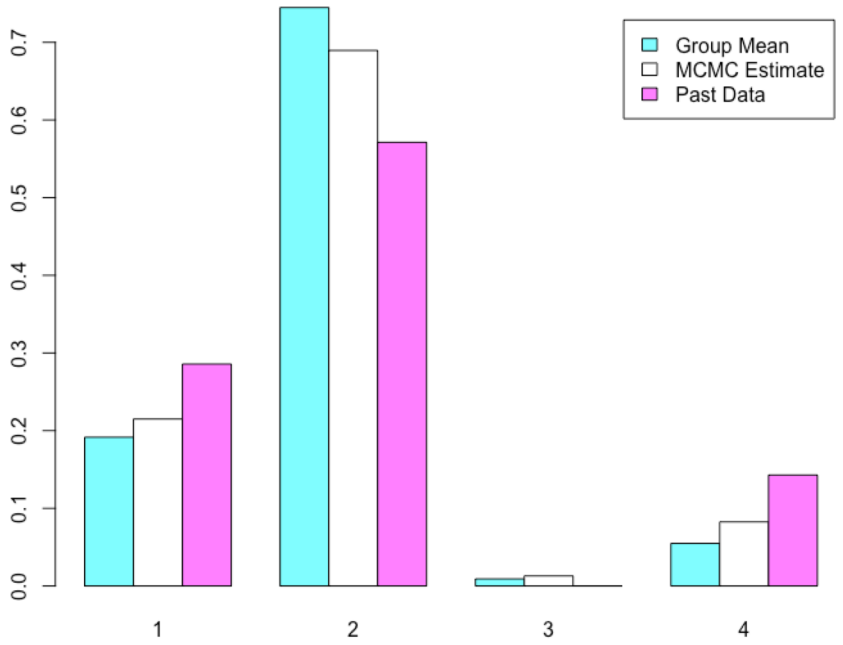
What Went Wrong?

- Example traceplots show a phenomenon known as *label switching*
  - Abrupt changes in mean; multi-modal density estimates
- From the model's point of view, the labels are interchangeable
  - Changing around the order of the labels leaves the likelihood (and the posterior distribution) unchanged
  - The same group label corresponds to different groups of observations on different iterations
  - The switches can clearly be seen on the traceplots
- Technically, labels for the groups are *not identifiable*
- Label switching is a well-known and much-discussed problem in latent variable models



Should we care about label switching?

- If group means are not of direct interest, then label-switching does not matter
- For example, the latent group model can predict a customer's future choices using past data from that customer
- Example:
  - Replace the last of 15 observations from a customer with NA (not available)
  - The JAGS model predicts the missing value using all other available information
  - We can record and tabulate the predicted value by defining it as a parameter in JAGS
- JAGS model shrinks mean of the remaining observations toward the group 2 mean
  - *The group labels were hidden from the model!*



What if the group means are of direct interest?

Correcting for Label Switching

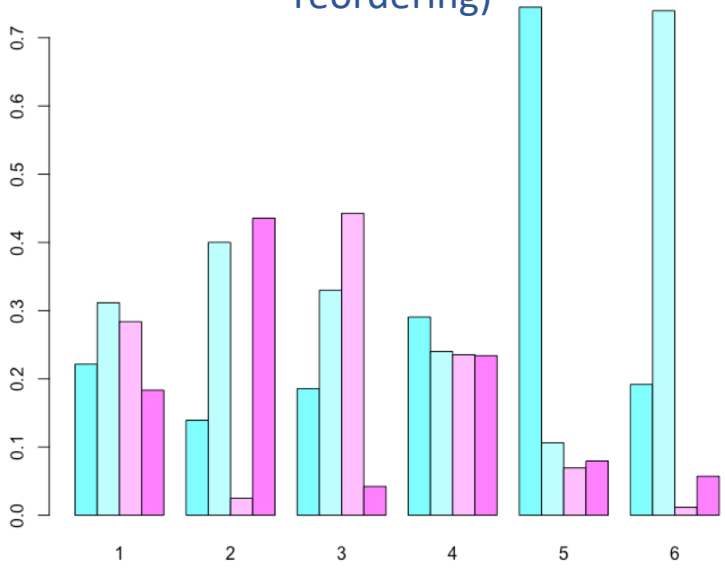
- Although the labels for the MCMC samples are not meaningful, the MCMC samples collectively contain good information on
  - Which groups of customers make similar choices to each other
  - Expected choice frequencies for similar clusters of customers
- MCMC samples can be post-processed to make the groups more homogeneous
- The `label.switching` package in R implements a several post-processing methods
- We try the `ecr.iterative.1` (first iterative version of Equivalence Classes Representatives) algorithm
  - Partition cluster assignments into equivalence classes that differ only by permutations of the class labels
  - Choose one of these equivalence classes

- Papastamoulis P. and Iliopoulos G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. Journal of Computational and Graphical Statistics, 19: 313-331.
- Rodriguez C.E. and Walker S. (2014). Label Switching in Bayesian Mixture Models: Deterministic relabeling strategies. Journal of Computational and Graphical Statistics. 23:1, 25-45

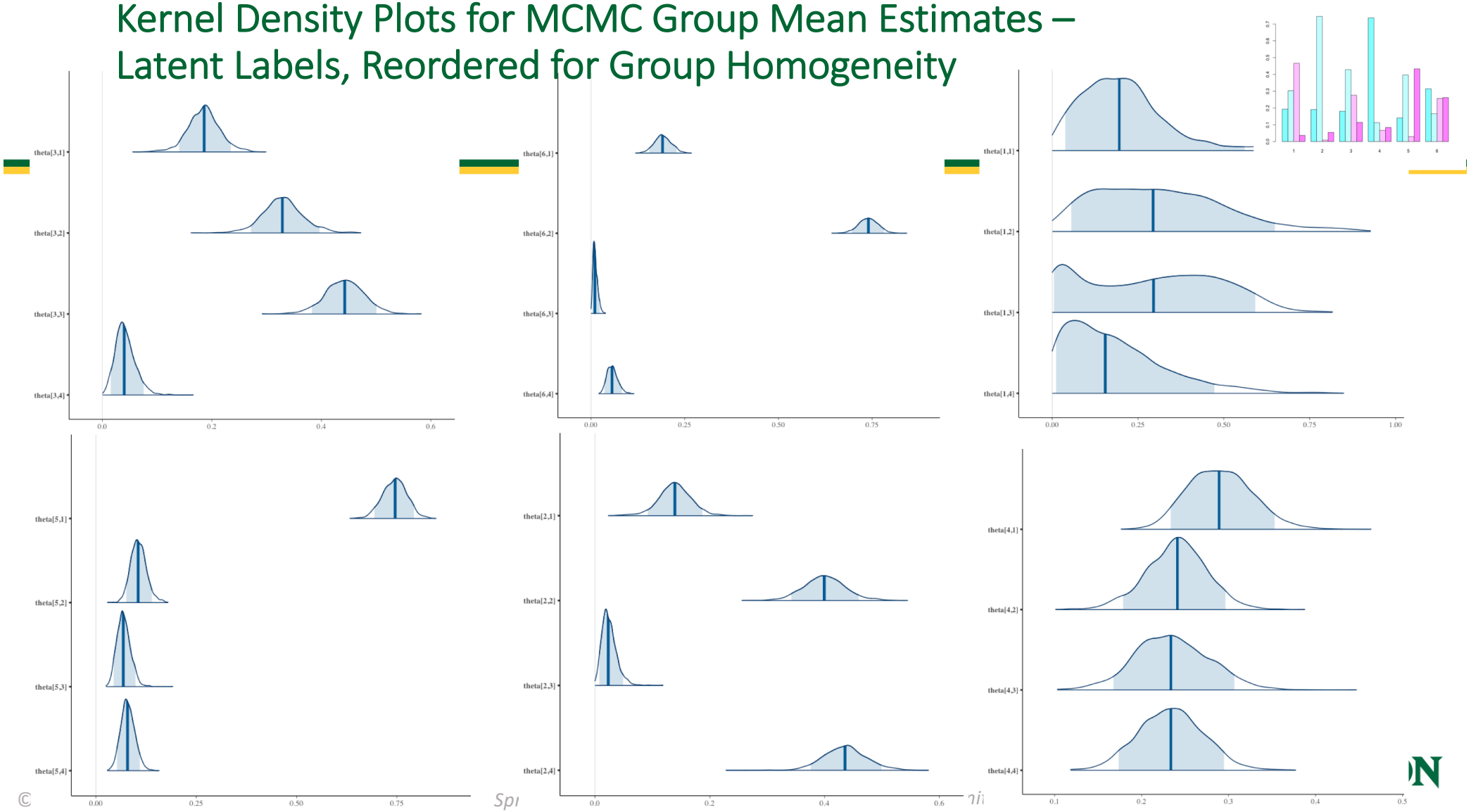
Results of Label Reordering

- 1000 iterations saved per chain
- Effective sample sizes of Θ_{gi} samples (according to `effectiveSize` function) range from 195.2 to 2025.0
- Trace plots of most Θ_{gi} look stationary
- Group means match rather well with means of a permutation of the original groups
- Some density estimates of Θ_{gi} are still multi-modal and have small effective sample sizes

MCMC Point Estimates of Item Probabilities by Group (latent group model with label reordering)

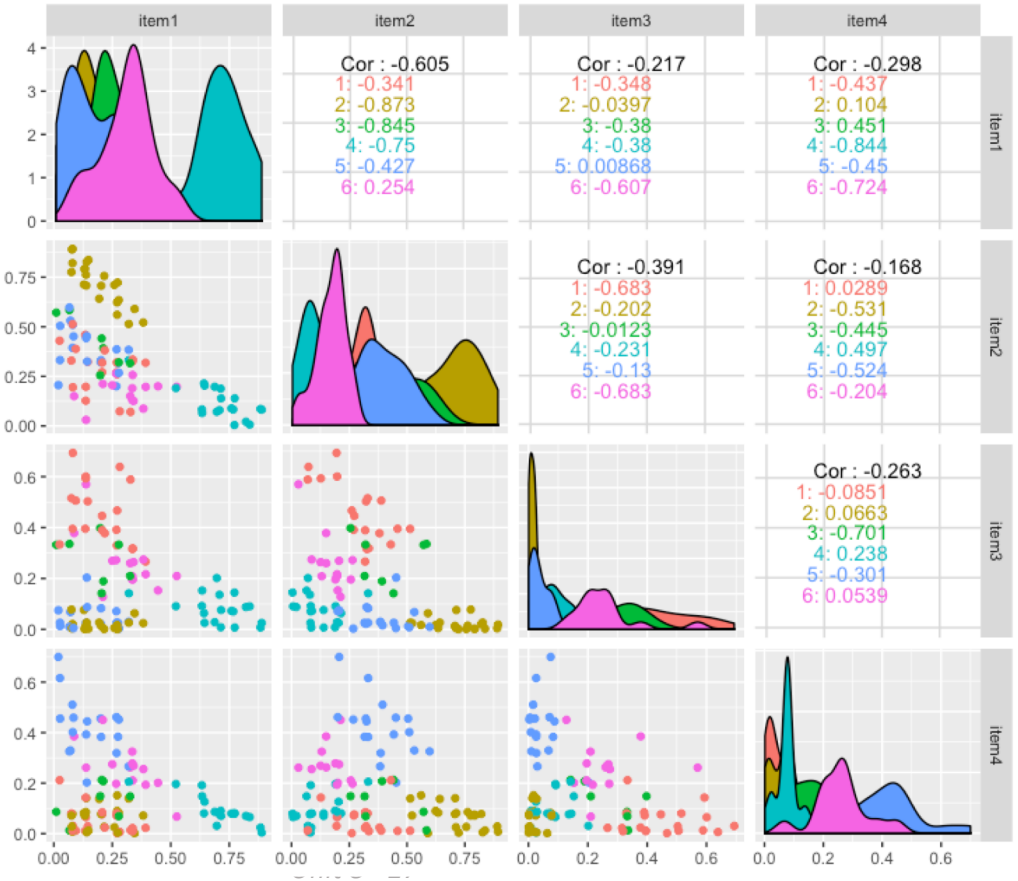


Kernel Density Plots for MCMC Group Mean Estimates – Latent Labels, Reordered for Group Homogeneity



Well-Separated Latent Groups are Easier to Discover

- Plot matrix shows that groups are fairly well separated except for group 3, which is inherently hard to distinguish from other groups
- Other than group 3 (group 1 in the reordered latent variable model) the lowest effective sample size for Θ_{gi} estimates was 973
- Group 3 effective sample sizes were all less than 500

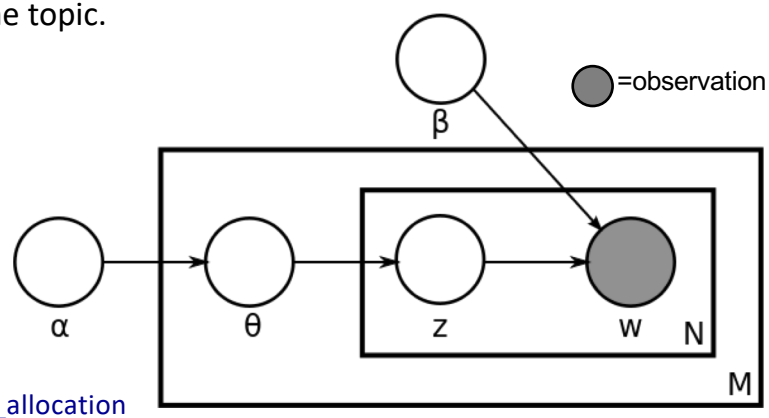


Summary: Discovering Latent Groups

- We considered an example in which the groups are not directly observable but are inferred from behavior of individuals
- The model gave good predictions for an individual's behavior given past behavior of all individuals
  - The model used behavior of similar individuals to adjust predictions for a given individual
- If group means are of direct interest, we need to do post-processing to correct for label switching
  - Inferences about group means are better when groups are well separated

Latent Dirichlet Allocation (LDA) Model

- Popular model used for natural language understanding and text retrieval (and has other applications such as finding bugs in software)
  - There are M documents
  - Each document has N words
  - The n^{th} word in the m^{th} document is W_{mn}
  - W_{mn} has an associated "topic" Z_{mn}
  - The topics Z_{mn} are independent draws from a K-dimensional multinomial distribution, where K is the number of topics. The parameter θ_m of this distribution depends on the document.
  - The words W_{mn} are independent draws from a L-dimensional multinomial distribution, where L is the number of words.
  - The parameter $\beta_{z_{mn}}$ of this distribution depends on the topic.
- The words are observed; the topics are discovered from the document corpus.
- A popular inference method is collapsed Gibbs sampling (marginalize out θ and sample Z from its marginal distribution)



Source: http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

LDA Topics from Enron Email Dataset

- About 500,000 emails generated by 500 people
- Made public by Federal Energy Regulatory Commission in wake of Enron collapse and scandal
- Widely used by machine learning researchers

https://www.cs.cmu.edu/~./enron/

commission	agreement	gas	davis	think	company	ticket	wine	var	click
pge	michelle	capacity	power	thing	firm	houston	date	area	free
customer	legal	pipeline	energy	going	fund	room	seller	random	offer
cpuc	employees	paso	stock	problem	partner	number	received	shaperect	online
comment	letter	natural_gas	consultant	look	ventures	reservation	cameron	color	price
decision	attached	california	governor	put	technology	confirmed	view	normal	receive
sce	plan	socal	public	lot	round	building	fax	error	link
filing	document	storage	calpine	believe	capital	hotel	pictures	engine	web
parties	review	market	company	big	services	street	suite	parentrandom	list
filed	company	demand	july	job	investor	seat	nov	diamond	save
final	emission	team	going	page	company	power	energy	market	power
schedule	environmental	group	think	court	stock	project	company	stock	edison
hour	air	process	weekend	employees	financial	india	power	earning	utility
hourahead	permit	review	hope	law	dynegy	government	market	report	billion
found	plant	sally	ill	labor	investor	company	trading	schwab	utilities
preferred	facility	operation	night	worker	shares	dpc	business	economic	pge
iep	unit	business	guy	union	billion	indian	natural_gas	index	california
sc_id	epa	global	thing	federal	credit	dabhol	companies	growth	states
mkt_type	water	office	friday	employer	rating	electricity	electricity	analyst	electricity
trans_date	station	meeting	sure	act	analyst	mseb	corp	nasdaq	bankruptcy
friend	meeting	lynn	company	contract	attached	bill	energy	travel	database
god	scheduled	gas	skilling	corp	report	dwr	california	hotel	message
gift	conference	contract	lay	review	file	puc	power	roundtrip	error
holiday	thursday	nng	business	material	document	access	electricity	fares	received
club	friday	point	stock	offer	comment	direct	plan	special	notify
love	monday	month	market	party	status	assembly	percent	offer	prohibited
children	attend	daily	profit	andor	draft	customer	davis	city	communication
family	office	meter	companies	message	review	committee	bush	visit	immediately
food	number	shipper	ceo	affiliates	version	senate	blackout	miles	delete
kid	tuesday	storage	world	basis	list	bond	bill	deal	operation
game	iso	robert	oil	american	product	project	service	texas	think
yard	market	michael	oil_gas	bush	business	request	services	longhorn	meeting
defense	load	david	industry	world	management	approval	fcc	team	forward
allowed	hour	mike	energy	attack	experience	resource	bill	game	discuss
fantasy	bid	steve	offshore	country	company	application	carrier	play	issues
point	nyiso	mark	drilling	government	sales	construction	telecom	brown	group
passing	prices	scott	field	president	marketing	manager	internet	season	jeff
rank	power	paul	company	war	services	transaction	commission	top	help
against	energy	richard	production	campaign	manager	site	local	true_orange	sure
team	pjm	james	houston	member	team	date	network	football	going
student	cost	updated	user	deal	california	energy	account	paper	market
program	customer	against	data	trading	ferc	conference	order	pulp	ferc
business	rate	play	access	risk	committee	program	buy	market	california
school	contract	free	password	position	refund	policy	online	houston	power
haas	rates	injury	center	transaction	power	event	free	ena	generation
university	pay	agent	server	book	senate	session	fund	prices	order
mba	amount	season	problem	power	price	member	investment	mill	transmission
berkeley	credit	start	address	product	document	research	cash	wind	utilities
class	period	expected	sap	credit	generator	group	receive	canada	price
interview	based	fantasy	outage	gas	billion	training	stock	story	price_cap

source: William Darling

Recap: The Bayesian Approach

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)}$$

Thomas Bayes (1702-1762)

- The Bayesian approach is:
  - A way of thinking about problems of inference and decision-making under uncertainty
  - A set of tools for applying this way of thinking to practical problems
- A Bayesian can answer the questions a decision-maker cares about:
  - What is the probability *this event* will happen?
  - What is the probability *this parameter* falls in a certain range?
  - What is my best decision in *these circumstances*?
- When we have a reasonable amount of data and weak prior information, we can give many standard statistical tools a Bayesian interpretation (at least approximately)
- Recent advances in computation have made Bayesian methods practical for many complex real-world problems

Fundamental Ideas of the Bayesian Approach [1 of 2]

- Probability expresses rational degrees of belief about uncertain phenomena
- Rational decision makers choose according to maximum expected utility
- Inference is belief dynamics
  - Prior beliefs are updated with evidence
  - Posterior beliefs at one stage become prior beliefs for next stage
  - Predict next stage using all knowledge up to present stage
  - Laplace: Probability is common sense reduced to calculation
- Conjugate prior / likelihood pairs simplify Bayesian inference
  - Posterior distribution and predictive distribution can be found exactly
  - Convenient and useful when a good model for the data

Fundamental Ideas of the Bayesian Approach [2 of 2]

- Approximation methods are important when exact results are unavailable
  - There is a rapidly growing literature in approximation methods for Bayesian inference
  - Markov Chain Monte Carlo (MCMC) is a general-purpose class of approximation methods that helped spark the Bayesian revolution
- Hierarchical models use structural assumptions to achieve better statistical power without sacrificing realism
  - Information sharing among related parameters
- Posterior predictive model evaluation can help assess whether model is adequate for the intended purpose
  - "All models are wrong but some models are useful" - Box

Bayesian and Frequentist Statistics

- Most statistics courses are taught from the frequentist perspective
- Frequentists
  - View probability as objective property of random processes
  - Assign probabilities to collectives but *not* individual events
  - Condition on parameters, treat data as probabilistic
- Subjectivists
  - View probability as rational degrees of belief about uncertain phenonmena
  - Assign probability to any unknown, including individual events
  - Condition on knowns, treat unknowns as probabilistic
- Frequentist analyses can often be given a Bayesian interpretation
  - Often good approximation if large sample and weak prior information

Unit 9: Summary and Synthesis

- We reviewed the fundamental principles of the Bayesian approach to inference and decision-making
- We introduced the Multinomial / Dirichlet conjugate pair
- We considered an example in inventory management and sales prediction
  - Predicted future sales of each product given data on past sales
  - Decided how much of each product to stock to maximize profits
  - Used hierarchical model to predict sales for sub-groups of customers
  - Used hierarchical model to discover latent clusters of customers with similar buying patterns
- We briefly introduced the latent Dirichlet allocation model to discover latent topics in collections of documents