Analyzing Marketing Data with an R-based Bayesian Approach

Peter Rossi
GSB/U of Chicago

based on work with Rob McCulloch, U of C, and Greg Allenby, OSU

Marketing Problems

Marketing is an applied field that seeks to optimize firm behavior with respect to a set of marketing actions, c.f.

- set prices optimally for a large number of items
- design products
- allocate marketing efforts – trade promotion budgets, sales force

Marketing Data

Survey Data: large number of respondents observed to choose between alternative products, rankings/ratings data. Multiple questions per respondent

Demand Data: data from point of sale optical scanning terminals. In US and Europe, all major retailers maintain large data warehouses with point of sale data.

- Items x Stores x Time >1000K.

Models and Methods of Inference

A great deal of disaggregate data

- panel structure (N large, T small)
- discrete response (mutually exclusive choices, multiple products consumed jointly)
- ordinal response (rankings)

Small amounts of information at the unit level

Requires Discrete Data models and a method of inference with a full accounting for uncertainty (only Bayes need apply)

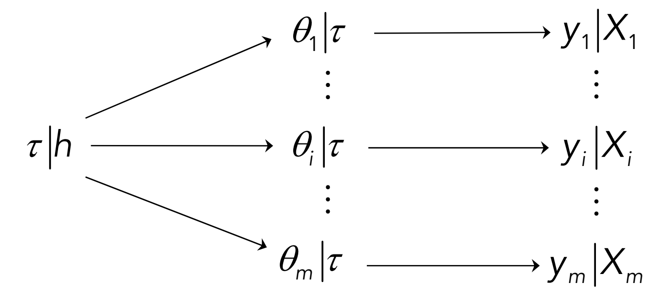
Hierarchical Models

A multi-level Model comprised of a set of conditional distributions:

- "unit-level" model – distribution of response given marketing variables
- first stage prior – specifies distribution of response parameters over units
- second stage prior – prior on parameters of first stage prior

Modular both conceptually and from a computational point of view.

A Graphical Review of Hierarchical Models



Second Stage Prior: Adaptive Shrinkage

First Stage Prior: Random Coef or Mixing Distribution

"Unit-Level" Likelihoods

Hierarchical Models and Bayesian Inference

Model to a Bayesian (Prior and Likelihood):

$$\prod_i p(\theta_i|\tau)p(\tau|h)\times\prod_i p(y_i|X_i,\theta_i)$$

Object of Interest for Inference (Posterior):

$$p(\theta_1,\ldots,\theta_m,\tau|y_1,\ldots,y_m)$$

Computational Method:

- MCMC (indirect simulation from joint posterior)

Implementation in R (*bayesm*)

Data Structures (all lists)

rxxxYyyZzz(Prior, Data, Mcmc)

Prior: list of hyperparms (defaults)

Data: list of lists for panel data

e.g. Data=list(regdata,Z)

regdata[[i]]=list(y,X)

Mcmc: Mcmc tuning parms

e.g. R (# draws), thining parm, Metropolis scaling (with def)

Implementation in R (*bayesm*)

Output:

- draws of model parameters:
- list of lists (e.g. normal components)
- 3 dim array (unit x coef x draw)

User Decisions:

- "burn-in" / convergence of the chain
  - run it longer!
  - Numerical Efficiency (`numEff`)
- how to summarize the joint distribution?

Coding

"Chambers" Philosophy – code in R, profile and rewrite only where necessary. Resulted in ~5000 lines of R code and 500 of C

As amateur R coders, we use only a tiny subset of R language. Code is numerically efficient but does not use many features such as classes

Moving toward more use of .Call to maximize use of R functions.

This maximizes readability of code.

We hope others will extend and modify.

Hierarchical Models considered in *bayesm*

rhierLinearModel	Normal Prior
rhierLinearMixed	Mixture of Normals
rhierMnlRwMixed	MNL with mixture of Normals
rhierMnlRwDP	MNL with Dirichlet Process Prior
rhierBinLogit	Binary logit with Normal prior
rhierNegBinRw	Neg Bin with Normal Prior
rscaleUsage	Ordinal Probit with Scale Usage
rnmixGibbs	Mixture of Normals density est
rDPGibbs	DP Prior density est

Hierarchical Linear Model- rhierLinearModel

Consider m regressions:

$$y_i = X_i\beta_i + \varepsilon_i \qquad \varepsilon_i \sim iidN\left(0,\sigma_i^2 I_{n_i}\right) \quad i = 1,\ldots,m$$

$$\beta_i = \bar{\beta} + v_i \quad v_i \sim iidN\left(0,V_\beta\right)$$

Priors :

$$\bar{\beta} \sim N\left(\bar{\bar{\beta}},A_\beta^{-1}\right);\ V_\beta \sim IW\left(\upsilon,\upsilon I\right)$$

Tie together via Prior

Adaptive Shrinkage

With fixed values of Δ, V_β, we have m independent Bayes regressions with informative priors.

In the hierarchical setting, we "learn" about the location and spread of the $\{\beta_i\}$.

The extent of shrinkage, for any one unit, depends on dispersion of betas across units and the amount of information available for that unit.

An Example – Key Account Data

y= log of sales of a "sliced cheese" product at a "key" account – market retailer combination

X: log(price)

display (dummy if on display in the store)

weekly data on 88 accounts. Average account has 65 weeks of data.

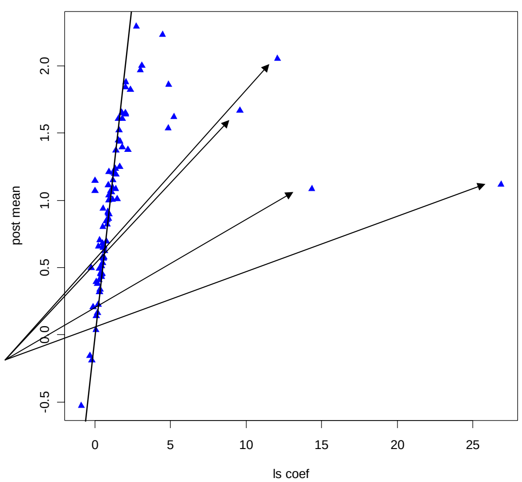
See data(cheese)

An Example – Key Account Data

Failure of Least Squares

some accounts have no displays!

some accounts have absurd coefs



Shrinkage

Prior on V_β is key.

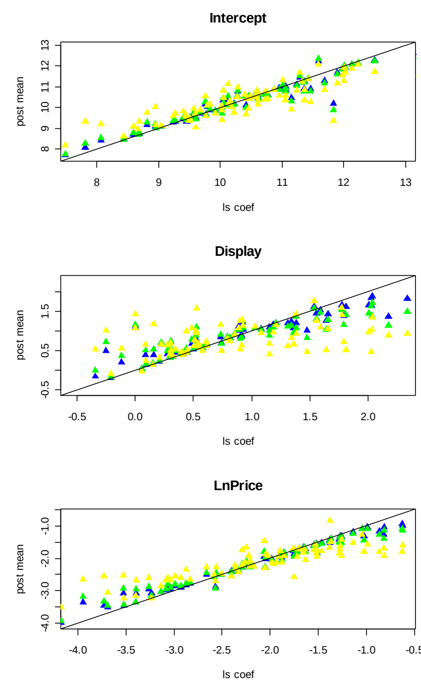
$V_\beta \sim IW(\upsilon, \upsilon I)$

blue : $\upsilon = k + 3$

green : $\upsilon = k + .5n$

yellow : $\upsilon = k + 2n$

Greatest Shrinkage for Display, least for intercepts



Heterogeneous logit model

Assume T_h observations per respondent

$$\Pr(y_{jth} = i) = \frac{\exp[x_{it}'\beta_h]}{\sum_j \exp[x_{jt}'\beta_h]}$$

The posterior:

$$p(\{\beta_h\}, \bar{\beta}, V_\beta \mid Data) \propto \prod_{h=1}^{H} \left(\prod_{t=1}^{T_h} p\left(y_{iht} \mid X_{ht}, \beta_h\right) \right) p\left(\beta_h \mid \tau\right) p\left(\tau\right)$$

logit model — normal heterogeneity — prior

Random effects with regressors

$$\beta_h = \Delta' z_h + v_i \quad v_h \sim iidN\left(0, V_\beta\right)$$

or

$$B = Z\Delta + U$$

Priors :

$$\delta = vec\left(\Delta\right) \sim N\left(\bar{\delta}, A_\beta^{-1}\right);\ V_\beta \sim IW\left(\upsilon, \upsilon I\right)$$

Δ is a matrix of regression coefficients related covariates (Z) to mean of random-effects distribution.

z_h are covariates for respondent h

data(bank)

Pairs of proto-type credit cards were offered to respondents. The respondents were asked to choose between cards as defined by "attributes."

Each respondent made between 13 and 17 paired comparisons.

Sample Attributes (14 in all):

Interest rate, annual fee, grace period, out-of-state or in-state bank, …

data(bank)

Not all possible combinations of attributes were offered to each respondent. Logit structure (independence of irrelevant alternatives makes this possible).

14,799 comparisons made by 946 respondents.

$$\Pr(card\ 1\ chosen) = \frac{\exp[x_{h,i,1}'\beta_h]}{\exp[x_{h,i,1}'\beta_h] + \exp[x_{h,i,2}'\beta_h]}$$

$$= \frac{\exp[(x_{h,i,1} - x_{h,i,2})'\beta_h]}{1 + \exp[(x_{h,i,1} - x_{h,i,2})'\beta_h]}$$

differences in attributes is all that matters

Sample observations

respondent 1 choose first card on first pair. Card chosen had attribute 1 on. Card not chosen had attribute 4 on.

id	choice	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14
1	1	1	0	0	-1	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	1	-1	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	1	-1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0	-1	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0	1	-1	0	0	0	0
1	1	0	0	0	-1	0	0	0	0	0	0	1	-1	0	0
1	1	0	0	0	0	0	0	0	0	-1	0	0	0	-1	0
1	0	0	0	0	0	0	0	0	0	1	0	0	0	-1	0
2	1	1	0	0	-1	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	1	-1	0	0	0	0	0	0	0	0	0

Sample demographics (Z)

id	age	income	gender
1	60	20	1
2	40	40	1
3	75	30	0
4	40	40	0
6	30	30	0
7	30	60	0
8	50	50	1
9	50	100	0
10	50	50	0
11	40	40	0
12	30	30	0
13	60	70	0
14	75	50	0

rhierBinLogit

```
z=read.table("bank.dat",header=TRUE)
d=read.table("bank demo.dat",header=TRUE)

# center demo data so that mean of random-effects
# distribution can be interpretted as the average respondents
d[,1]=rep(1,nrow(d))
d[,2]=d[,2]-mean(d[,2])
d[,3]=d[,3]-mean(d[,3])
d[,4]=d[,4]-mean(d[,4])
hh=levels(factor(z$id))
nhh=length(hh)

Dat=NULL

for (i in 1:nhh) {
   y=z[z[,1]==hh[i],2]
   nobs=length(y)
   X=as.matrix(z[z[,1]==hh[i],c(3:16)])
   Dat[[i]]=list(y=y,X=X)
        }
```


Running rhierBinLogit (continued)

```
Data=list(Dat=Dat,Demo=d)

nxvar=14
ndvar=4
nu=nxvar+5
Prior=list(nu=nu,V0=nu*diag(rep(1,nxvar)),
      deltabar=matrix(rep(0,nxvar*ndvar),
   ncol=nxvar),
      Adelta=.01*diag(rep(1,ndvar)))

Mcmc=list(R=20000,sbeta=0.2,keep=20)

out=rhierBinLogit(Prior=Prior,Data=Data,Mcmc=Mcmc)
```

Running rhierBinLogit (continued)

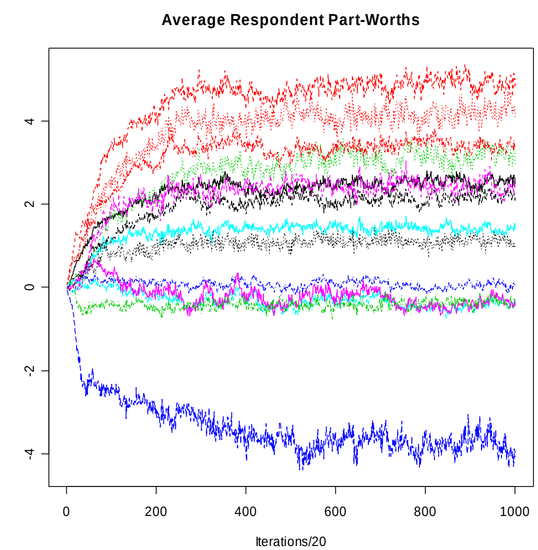
```
Attempting MCMC Inference for Hierarchical Binary Logit:
   14  variables in X
   4  variables in Z
   for  946  cross-sectional units

Prior Parms:
nu = 17
V
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
 [1,]   17    0    0    0    0    0    0    0    0     0     0     0     0     0
 [2,]    0   17    0    0    0    0    0    0    0     0     0     0     0     0
 [3,]    0    0   17    0    0    0    0    0    0     0     0     0     0     0
 [4,]    0    0    0   17    0    0    0    0    0     0     0     0     0     0
 [5,]    0    0    0    0   17    0    0    0    0     0     0     0     0     0
 [6,]    0    0    0    0    0   17    0    0    0     0     0     0     0     0
 [7,]    0    0    0    0    0    0   17    0    0     0     0     0     0     0
 [8,]    0    0    0    0    0    0    0   17    0     0     0     0     0     0
 [9,]    0    0    0    0    0    0    0    0   17     0     0     0     0     0
[10,]    0    0    0    0    0    0    0    0    0    17     0     0     0     0
[11,]    0    0    0    0    0    0    0    0    0     0    17     0     0     0
[12,]    0    0    0    0    0    0    0    0    0     0     0    17     0     0
[13,]    0    0    0    0    0    0    0    0    0     0     0     0    17     0
[14,]    0    0    0    0    0    0    0    0    0     0     0     0     0    17
Deltabar
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    0    0    0    0    0    0    0    0    0     0     0     0     0     0
[2,]    0    0    0    0    0    0    0    0    0     0     0     0     0     0
[3,]    0    0    0    0    0    0    0    0    0     0     0     0     0     0
[4,]    0    0    0    0    0    0    0    0    0     0     0     0     0     0
ADelta
     [,1] [,2] [,3] [,4]
[1,] 0.01 0.00 0.00 0.00
[2,] 0.00 0.01 0.00 0.00
[3,] 0.00 0.00 0.01 0.00
[4,] 0.00 0.00 0.00 0.01

MCMC Parms:
sbeta= 0.2  R=  20000  keep=  20

MCMC Iteration (est time to end - min)
  100  ( 153.6 )
```


Running rhierBinLogit (continued)

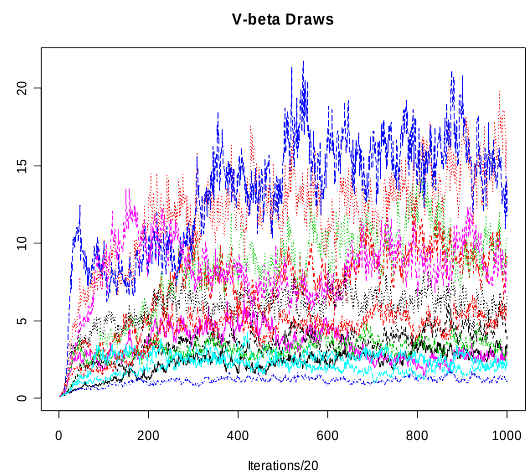
```
 19900  ( 0.8 )
  20000  ( 0 )
 Total Time Elapsed:  154.33
> str(out)
List of 5
 $ betadraw : num [1:946, 1:14, 1:1000]  0.4868  0.1015 -0.2833 -0.3313  0.0549 ...
 $ Vbetadraw: num [1:1000, 1:196] 0.0651 0.0880 0.0973 0.1332 0.1204 ...
 $ Deltadraw: num [1:1000, 1:56] -0.00758 -0.00291  0.00996  0.03392  0.03758 ...
 $ llike    : num [, 1:1000] -9744 -9592 -9372 -9262 -8997 ...
 $ reject   : num [, 1:1000] 0.607 0.593 0.598 0.653 0.607 ...
```

We now must summarize these numbers:

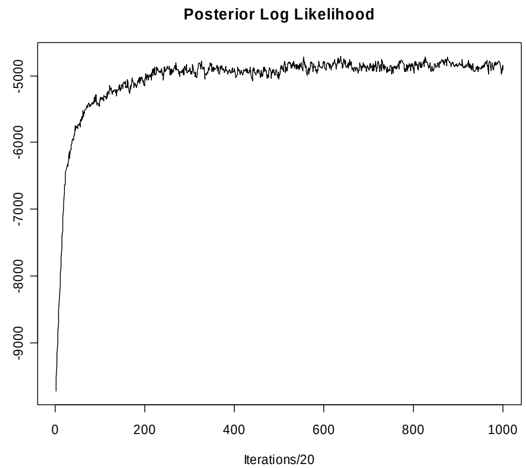
1. Convergence of chain (trace plots)
2. Marginal distribution of various model parameters



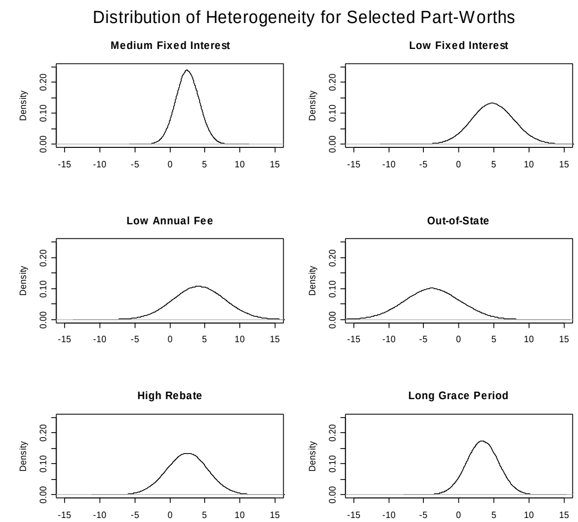
Elements of out$Deltadraw



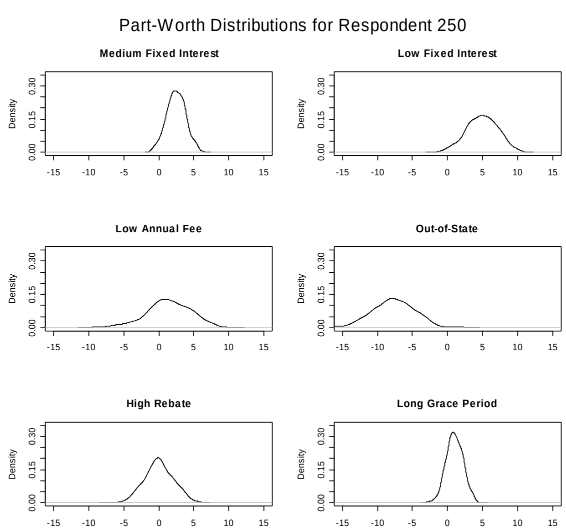
Elements of out$Vbetadraw



out$llike



Smoothed density estimate of out$betadraws (after burn-in)



Non-normal Priors (mixture of normals)

$$\beta_h = \Delta' z_h + v_h \quad v_h \sim iidN\left(\mu_{ind_h}, \Sigma_{ind_h}\right)$$

$$ind_h \sim multinomial(pvec)$$

Priors:

$$pvec \sim Dirichlet(a)$$

$$(\mu_k, \Sigma_k) \sim iid \;\; Natural\ Conjugate\ k = 1, \ldots, \dim(pvec)$$

An Application to Scanner Panel Data

Observe a panel of 347 households selecting from 5 brands of tub margarine.

No reason to believe that coefficients of the multinomial logit are normally distributed over households.

For example, some households may be willing to pay a premium for certain brands.

Included covariates: brand intercepts, log-price, "loyalty" variable

RhierMnlRwMixture

Implements an unconstrained Gibbs Sampler for a mixture of normals distribution as the first stage prior.

Combined with Metropolis algorithm to draw logit coefficient vectors for each panelist.

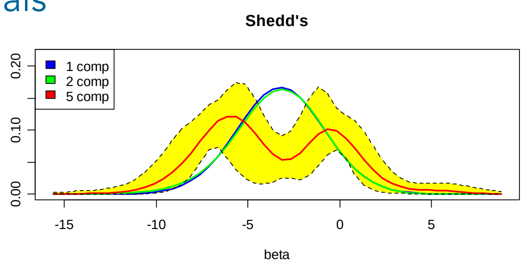
Returns draws of each component in normal mixture. Estimate the density at a point:

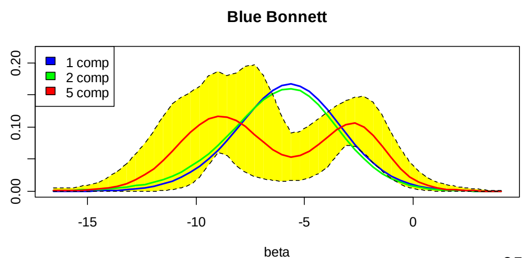
$$\hat{p}(\beta) = \frac{1}{R}\sum_{r}\sum_{k} pvec_k^r \times \varphi\left(\beta \middle| \mu_k^r, \Sigma_k^r\right)$$

eMixMargDen(grid, probdraw,compdraw)

Mixture of Normals

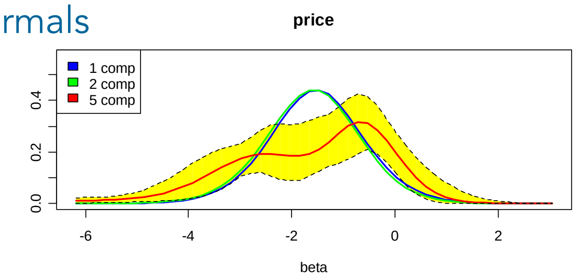
Brand Intercepts

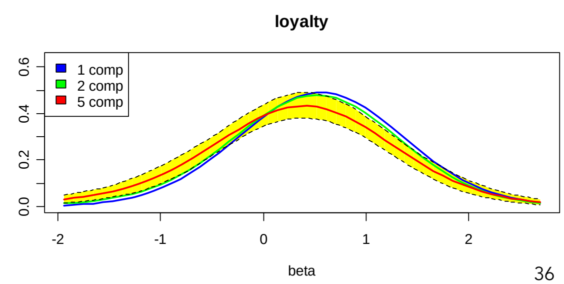




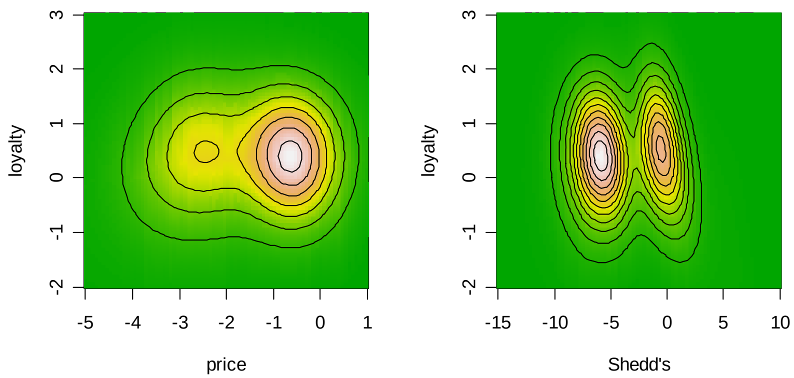
Mixture of Normals

loyalty distribution pretty normal but everything else non-normal!





Mixture of Normals



mixDenBi(i,j,gridi,gridj,probdraw,compdraw)

Scale Usage Heterogeneity

Survey questions involving a rating scale for satisfaction/purchase intention/happiness are commonplace

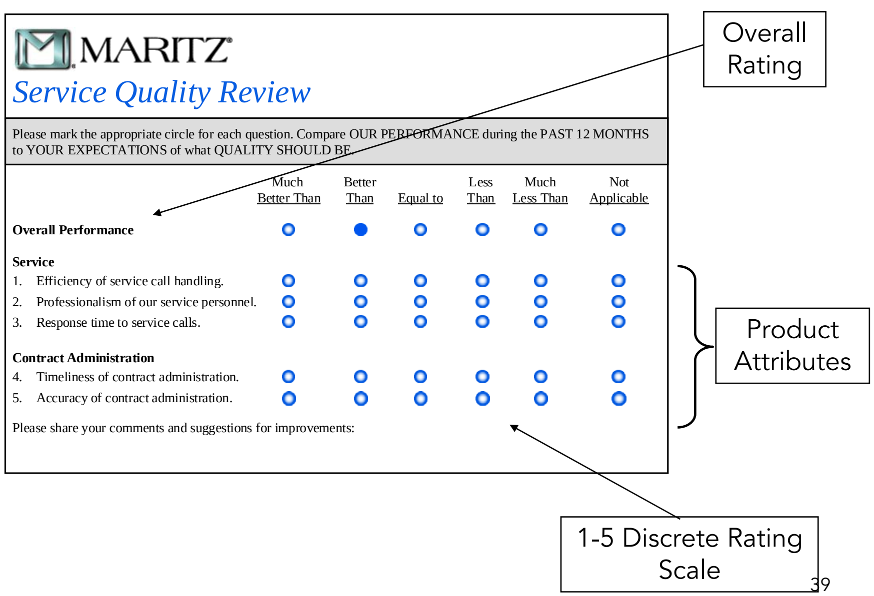
Typically, respondents rate products (overall) and attributes on a ordinal (5/7/9) point scale

Respondents exhibit scale usage heterogeneity. Some use only upper or lower end of the scale.

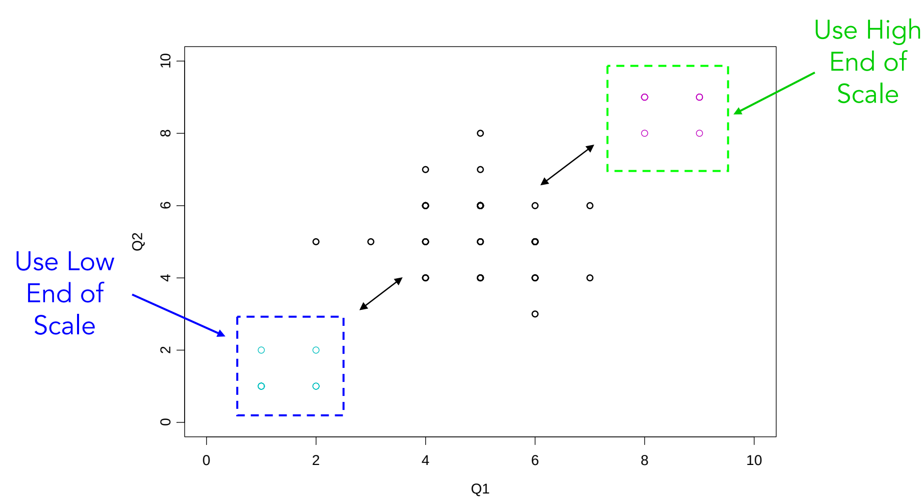
What biases are caused by this?

Can we make anything more than ordinal statements?

Example of CSM Questionnaire



+ve Covariance Bias



Model

Latent Variable Formulation:

We observe a vector x_i (M x 1) of discrete/ordered responses:

No. of Survey Questions

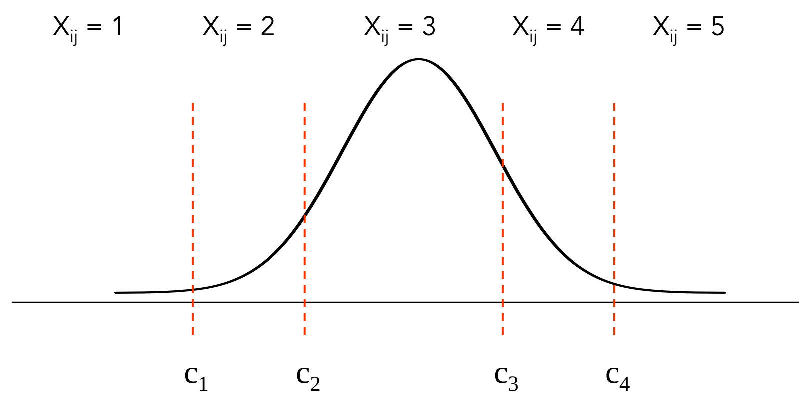
$x_{ij} = \{1, \ldots, K\}$; $i = 1, \ldots, N$

Pts in the scale

$$y_{ij} < c_1 \qquad x_{ij} = 1$$
$$c_1 < y_{ij} < c_2 \qquad x_{ij} = 2$$
$$\vdots$$
$$y_{ij} > c_{K-1} \qquad x_{ij} = K$$

$$y_i \sim iidN\left(\mu_i^*, \Sigma_i^*\right)$$

Model: Example with 5 point scale



Model: Scale Usage Heterogeneity

We incorporate scale usage heterogeneity using location-scale shift at the latent variable level

Location shift

$$y_i = \mu + \tau_i \iota + \sigma_i z_i$$

$$z_i \sim N(0, \Sigma)$$

Scale shift

For example:

top end of scale -- large value of τ and small σ

Hierarchical Model rscaleUsage

We use non-standard hierarchical (random effects) formulation: $\begin{bmatrix} \tau_i \\ \ln \sigma_i \end{bmatrix} \sim N(\varphi, \Lambda)$

$$\varphi, \Lambda | h \rightarrow (\tau_1, \sigma_1) | \varphi, \Lambda \rightarrow x_1 | y_1 \qquad y_1 | \tau_1, \sigma_1, \Sigma$$
$$\vdots \qquad\qquad \vdots$$
$$\varphi, \Lambda | h \rightarrow (\tau_i, \sigma_i) | \varphi, \Lambda \rightarrow x_i | y_i \qquad y_i | \tau_i, \sigma_i, \Sigma$$
$$\vdots \qquad\qquad \vdots$$
$$\varphi, \Lambda | h \rightarrow (\tau_N, \sigma_N) | \varphi, \Lambda \rightarrow x_N | y_N \qquad y_N | \tau_N, \sigma_N, \Sigma$$

Some Real Data: data(customerSat)

Customer Survey in Business-to-Business Context

Product is a form of Business Advertising

10 Qs -- 10 pt scale (10 is “excellent,” 1 is “poor”)

N=1810/M=10/K=10

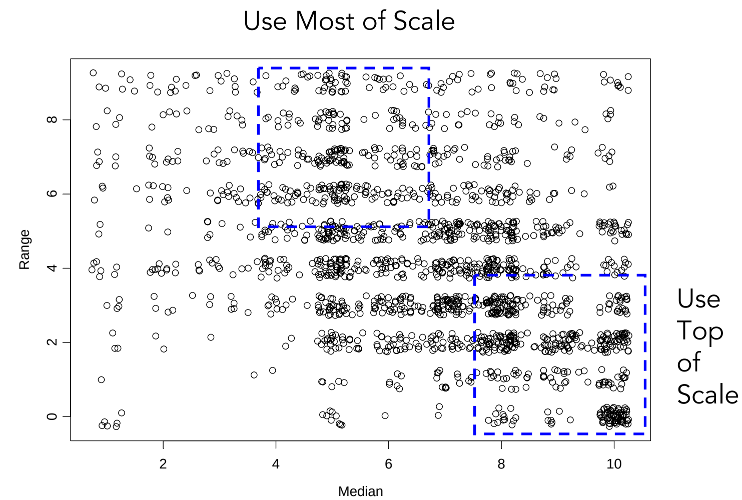
Q1: Overall Value

Q2-Q4: Price

Q5-Q10: Effectiveness

reach/geographic area/attracting customers/evaluation of effectiveness

Evidence of Scale Usage Heterogeneity



Correlation Structure: Raw Data

High Correlations between each Q2-Q10 and Q1.

Positive correlations Q2-Q10

Q.	Mean	Covariance*Correlation* Matrix									
1	6.06	6.50	*0.65*	*0.62*	*0.78*	*0.65*	*0.74*	*0.59*	*0.56*	*0.44*	*0.45*
2	5.88	4.38	7.00	*0.77*	*0.76*	*0.55*	*0.49*	*0.42*	*0.43*	*0.35*	*0.35*
3	6.27	4.16	5.45	7.06	*0.72*	*0.52*	*0.46*	*0.43*	*0.46*	*0.38*	*0.40*
4	5.55	5.36	5.43	5.16	7.37	*0.64*	*0.67*	*0.52*	*0.52*	*0.41*	*0.40*
5	6.13	4.35	3.83	3.62	4.53	6.84	*0.69*	*0.58*	*0.59*	*0.49*	*0.46*
6	6.05	4.82	3.29	3.15	4.61	4.61	6.49	*0.59*	*0.59*	*0.45*	*0.44*
7	7.25	3.64	2.70	2.73	3.42	3.68	3.66	5.85	*0.65*	*0.62*	*0.60*
8	7.46	3.28	2.61	2.79	3.23	3.51	3.41	3.61	5.21	*0.62*	*0.62*
9	7.89	2.41	1.99	2.18	2.39	2.72	2.47	3.20	3.02	4.57	*0.75*
10	7.77	2.55	2.06	2.33	2.42	2.67	2.51	3.21	2.95	3.54	4.89

Correlation Structure: Standardized Data

Correlations are attenuated -- some -ve

Q.	Mean	Covariance*Correlation* Matrix									
1	-0.29	0.66	*-0.07*	*-0.13*	*0.03*	*-0.14*	*0.06*	*-0.11*	*-0.16*	*-0.24*	*-0.21*
2	-0.42	-0.05	0.82	*0.35*	*0.20*	*-0.19*	*-0.36*	*-0.32*	*-0.25*	*-0.26*	*-0.27*
3	-0.18	-0.10	0.31	0.93	*0.14*	*-0.21*	*-0.33*	*-0.33*	*-0.24*	*-0.24*	*-0.22*
4	-0.60	0.02	0.14	0.11	0.62	*-0.23*	*-0.17*	*-0.24*	*-0.20*	*-0.26*	*-0.28*
5	-0.28	-0.09	-0.15	-0.18	-0.16	0.76	*0.04*	*-0.07*	*-0.01*	*-0.10*	*-0.11*
6	-0.32	0.04	-0.28	-0.27	-0.12	0.03	0.74	*0.03*	*0.03*	*-0.12*	*-0.14*
7	0.33	-0.08	-0.23	-0.26	-0.16	-0.05	0.02	0.67	*0.01*	*0.06*	*0.05*
8	0.46	-0.09	-0.16	-0.17	-0.12	-0.01	0.02	0.01	0.56	*0.01*	*-0.04*
9	0.68	-0.14	-0.17	-0.18	-0.16	-0.07	-0.08	0.04	0.00	0.58	*0.31*
10	0.61	-0.14	-0.20	-0.18	-0.18	-0.08	-0.10	0.03	-0.02	0.19	0.67

Correlation Structure of Latent Variables

Not all strongly related to overall

Q.	Mean (μ)	Covariance*Correlation* Matrix (Σ)									
1	6.43 (.08)	4.13 (.73)	*.31*	*.25*	*.55*	*.29*	*.39*	*.15*	*.05*	*-.15*	*-.12*
2	6.16 (.08)	1.50 (.65)	5.7 (.77)	*.65*	*.61*	*.16*	*-.09*	*-.11*	*-.11*	*-.25*	*-.23*
3	6.47 (.08)	1.33 (.67)	4.07 (.73)	6.93 (.86)	*.53*	*.13*	*-.08*	*-.05*	*-.03*	*-.13*	*-.09*
4	6.00 (.08)	2.79 (.70)	3.70 (.74)	3.49 (.76)	6.34 (.86)	*.31*	*.29*	*.07*	*.05*	*-.15*	*-.14*
5	6.46 (.08)	1.36 (.65)	0.87 (.65)	0.82 (.67)	1.81 (.70)	5.44 (.78)	*.38*	*.22*	*.21*	*.02*	*.02*
6	7.39 (.08)	1.55 (.63)	-0.42 (.60)	-.39 (.62)	1.42 (.66)	1.73 (.74)	3.89 (.69)	*.20*	*.12*	*-.13*	*-.13*
7	7.50 (.08)	0.77 (.60)	-0.67 (.59)	-0.34 (.63)	0.43 (.64)	1.31 (.62)	1.00 (.59)	6.49 (.78)	*.49*	*.49*	*.46*
8	7.50 (.08)	0.24 (.57)	-0.60 (.57)	-0.15 (.61)	0.26 (.60)	1.10 (.60)	0.56 (.56)	2.84 (.65)	5.29 (.73)	*.47*	*.43*
9	7.84 (.08)	-0.75 (.58)	-1.45 (.57)	-0.82 (.64)	-0.96 (.60)	0.11 (.61)	-0.65 (.56)	3.07 (.71)	2.68 (.69)	6.13 (.87)	*.71*
10	7.76 (.08)	-0.60 (.59)	-1.38 (.59)	-0.58 (.65)	-.91 (.62)	0.10 (.62)	-0.64 (.57)	2.97 (.71)	2.48 (.69)	4.41 (.80)	6.36 (.89)

-ve between price and reach

External Validation

Survey contains some information on intention to increase expenditure next year as well as past years expenditures.

Sort by overall measures, and compare cumulative expenditure % change to average % change ("lift")

Quantile	Raw	Centered	Row Mean	τ_i	Latent
Top 5%	.69	.66	-.076	-.30	3.59
Top 10%	1.39	1.28	.25	.78	2.35
Top 25%	1.76	1.38	1.59	1.18	1.98
Top 50%	1.29	.95	1.051	1.11	1.62

Summary

Analysis of Marketing Data requires models appropriate for discrete, panel data.

Bayesian methods are the only computationally feasible methods for many of these models.

User discretion and judgement is required for any sensible analysis.

R-based implementations are possible and provide useable solutions even for large datasets.