Model choice and decision theory

Decision theory

Loss: $L(a,\theta)$ where a=action; θ=state of nature

Bayesian decision theory:

$$\min_a \left\{ \bar{L}(a) = E_{\theta|D}\left[L(a,\theta)\right] = \int L(a,\theta)p(\theta \mid D)d\theta \right\}$$

note separation of Loss function from posterior/likelihood!

Profit function is the natural loss for marketing applications!

Model Selection

We are often faced with the problem of selection from a set of models. The Bayes solution is to compute the posterior probability of each model.

For the set of models: $M_1,\ldots,M_k$

compute:

$$p(M_i|y) = \frac{p(y|M_i)p(M_i)}{p(y)}$$

Posterior odds Ratio

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \frac{p(M_1)}{p(M_2)}$$

$$= \text{Bayes Factor} \times \text{Prior Odds}$$

Model Probabilities cont.

For parametric models,

$$p(y|M_i) = \int p(y|\theta, M_i)\, p(\theta|M_i)\, d\theta$$

Depends on the prior! It should. One interpretation is that the model prob is the average of the "likelihood" wrt to the prior. Note the likelihood must include all normalizing constants!

$$\ell^*(y|M_i) = E_{\theta|M_i}\left[\ell(\theta|y, M_i)\right]$$

Bayes Factors

$$BF = \frac{\int p(y|\theta, M_1)p(\theta|M_1)d\theta}{\int p(y|\theta, M_2)p(\theta|M_2)d\theta} = \frac{p(y|M_1)}{p(y|M_2)}$$

BF is a ratio of predictive densities. There is no distinction between nested and non-nested models!

BF depends critically on prior. Must be carefully chosen. Diffusion matters!

As prior diffusion increases for one model, BF moves away from that model.

Improper Priors are dangerous!

Model Probabilities cont.

The marginal density of the data is also the normalizing constant for the posterior.

$$p\left(\theta \middle| y, M_i\right) = \frac{\ell\left(\theta \middle| y, M_i\right) p\left(\theta \middle| M_i\right)}{p\left(y \middle| M_i\right)}$$

The numerator above is the un-normalized posterior. This we can always evaluate. The marginal density of the data is not always easy!

$$p\left(y \middle| M_i\right) = \int \tilde{p}\left(\theta \middle| y, M_i\right) d\theta = \frac{\tilde{p}\left(\theta \middle| y, M_i\right)}{p\left(\theta \middle| y, M_i\right)}$$

Savage-Dickey Conjugate setting

$M_0 : \phi_1 = \phi_1^h, \quad M_1 : \text{unrestricted} \quad \text{where } \phi' = \left(\phi_1', \phi_2'\right).$

$$\text{BF}: \frac{p\left(y \middle| M_0\right)}{p\left(y \middle| M_1\right)} = \frac{\int \ell\left(\phi_2 \middle| y\right) p\left(\phi_2\right) d\phi_2}{\iint \ell\left(\phi_1, \phi_2 \middle| y\right) p\left(\phi_1, \phi_2\right) d\phi_1 d\phi_2}$$

$$\ell\left(\phi_2 \middle| y\right) = \ell\left(\phi_1, \phi_2 \middle| y\right)\Big|_{\phi_1 = \phi_1^h}$$

a “natural” choice for prior

$$p\left(\phi_2 \mid \phi_1 = \phi_1^h\right) = \frac{p\left(\phi_1, \phi_2\right)}{\int p\left(\phi_1, \phi_2\right) d\phi_2}\Bigg|_{\phi_1 = \phi_1^h}$$

Savage-Dickey Conjugate setting

$$\mathrm{BF} = \frac{\int \ell(\phi_1,\phi_2 \mid y)\left(p(\phi_1,\phi_2)/p(\phi_1)\right)d\phi_2\Big|_{\phi_1=\phi_1^h}}{\int\int \ell(\phi_1,\phi_2 \mid y)p(\phi_1,\phi_2)\,d\phi_1 d\phi_2}$$

$$= \frac{\int p(\phi_1,\phi_2 \mid y)\,d\phi_2}{p(\phi_1)}\Bigg|_{\phi_1=\phi_1^h}$$

$$= \frac{p(\phi_1 \mid y)}{p(\phi_1)}\Bigg|_{\phi_1=\phi_1^h} = \frac{\text{marginal posterior}}{\text{marginal prior}}$$

Asymptotic methods (Laplace)

$$p(y \mid M_i) = \int \exp\left(\Gamma(\theta)\right) d\theta \qquad \Gamma(\theta) = \log\left(\tilde{p}\left(\theta \middle| y\right)\right)$$

$$\approx \int \exp\left(\Gamma\left(\tilde{\theta}\right) - \frac{1}{2}\left(\theta - \tilde{\theta}\right)' H\left(\tilde{\theta}\right)\left(\theta - \tilde{\theta}\right) \right) d\theta$$

$$= \exp\left(\Gamma\left(\tilde{\theta}\right)\right)(2\pi)^{p_i/2}\left|H\left(\tilde{\theta}\right)\right|^{-1/2} \quad \text{where } H\left(\tilde{\theta}\right) = -\frac{\partial^2 \Gamma(\theta)}{\partial\theta\,\partial\theta'}\Big|_{\theta=\tilde{\theta}}$$

$$\frac{\partial^2 \Gamma(\theta)}{\partial\theta\,\partial\theta'} = \frac{\partial^2 \log\left(p(\theta)\right)}{\partial\theta\,\partial\theta'} + \frac{\partial^2 \log\left(p\left(y \middle|\ \theta\right)\right)}{\partial\theta\,\partial\theta'}$$

Asymptotic methods (Laplace)

Laplace approximation depends on:

prior evaluated @ posterior mode

prior curvature at mode

“highest” quality asymptotic approximation

Note: posterior can be easier to maximize than likelihood.

Tails on informative priors can deal with unidentified parameters and non-existence of MLEs.

Asymptotic methods (Laplace)

Laplace :

$$\mathrm{BF} \approx (2\pi)^{(p_1 - p_2)/2} \frac{p_1(\tilde{\theta})\, p_1(y|\tilde{\theta}_1)\, |H_1(\tilde{\theta})|^{-1/2}}{p_2(\tilde{\theta})\, p_2(y|\tilde{\theta}_2)\, |H_2(\tilde{\theta})|^{-1/2}}$$

number of parms in each model

Asymptotic methods (BIC)

Alternative(dumber): expand about the MLE and throw out terms that don't depend on n.

$$p(y \mid M_i) \approx \exp\left(\Gamma\left(\hat{\theta}_{mle}\right)\right)(2\pi)^{p_i/2}\left|Inf\left(\hat{\theta}_{mle}\right)\right|^{-1/2}$$

goes to constant matrix as n goes to infinity

$$= \exp\left(\Gamma\left(\hat{\theta}_{mle}\right)\right)(2\pi)^{p_i/2} n^{-p_i/2}\left|Inf_i\left(\hat{\theta}_{mle}\right)/n\right|^{-1/2}$$

$$\approx \tilde{p}\left(y \mid \hat{\theta}_{MLE}, M_i\right) n^{-p_i/2} \qquad \text{as } n \to \infty$$

Asymptotic methods (BIC)

$$\text{BIC}: \quad \log\left(p\left(M_i \middle| y\right)\right) \approx \log\left(\ell_i\left(\hat{\theta}\right)\right) - \frac{p_i}{2}\log(n)$$

or

$$\text{BF} \approx \log\left(LR_{1,2}\right) - \frac{\Delta p}{2}\log(n)$$

all that remains of the prior!

ridiculously inaccurate!

useful in two ways:

1). pick model with highest prob

2). establishes intuition that BF "automatically penalizes for "overfitting"

Computing Model Probs

For non-conjugate problems, there are three approaches:

1. Importance Sampling
2. Use of MCMC draws (NR)
3. Chib's Method (useful for latent var models)

BF using MCMC draws

$$\int \frac{q(\theta)}{\tilde{p}(\theta \mid y, M_i)} p(\theta \mid y, M_i)\, d\theta = \int \frac{q(\theta)}{p(\theta \mid M_i) p(y \mid \theta, M_i)} p(\theta \mid y, M_i)\, d\theta$$

$$= \frac{1}{p(y \mid M_i)} \int q(\theta)\, d\theta$$

$$= \frac{1}{p(y \mid M_i)}$$

$$E_{\theta \mid y, M_i}\left[\frac{q(\theta)}{\tilde{p}(\theta \mid y, M_i)}\right] = \frac{1}{p(y \mid M_i)}$$

Special case (Newton-Raftery)

If $q(\theta) = p(\theta \mid M_i)$,

$$p(y \mid M_i) = \frac{1}{E_{\theta \mid y, M_i}\left[\frac{1}{\ell(\theta \mid M_i)}\right]}$$

$$\hat{p}(y \mid M_i) = \frac{1}{\frac{1}{R}\sum_{r=1}^{R}\frac{1}{\ell(\theta^r \mid M_i)}}$$

`logMargDenNR`

Appeal: uses MCMC draws and likelihood

Problems: extreme sensitivity to outliers.
Distribution of likelihood values is non-uniform!

Marketing decisions and BDT

$$\max_{x} E\left[\pi\left(x \mid \Omega\right)\right]$$

There exists a range of marketing actions, x.

Actions are taken with respect to various information sets, Ω, reflecting what is known about model parameters.

Best action maximizes profits, π.

Loss/profit function

Canonical Example:

Sales Response Function

$$p(y|x,\theta)$$

Sales Marketing Mix

$$\max_x E_{y|x}\left[\pi(y|x,\theta)\right]$$

Full Bayes

Sources of Uncertainty:

- Distribution of sales given x, θ
- Uncertainty in θ (summarized in posterior)
  - note: posterior is a means to the end of computing predictive distribution of sales

$$\pi^{*}(x) = E_{\theta}\left[E_{y|x,\theta}\left[\pi(y|x,\theta)\right]\right]$$

$$\equiv E_{\theta}\left[\bar{\pi}(x|\theta)\right]$$

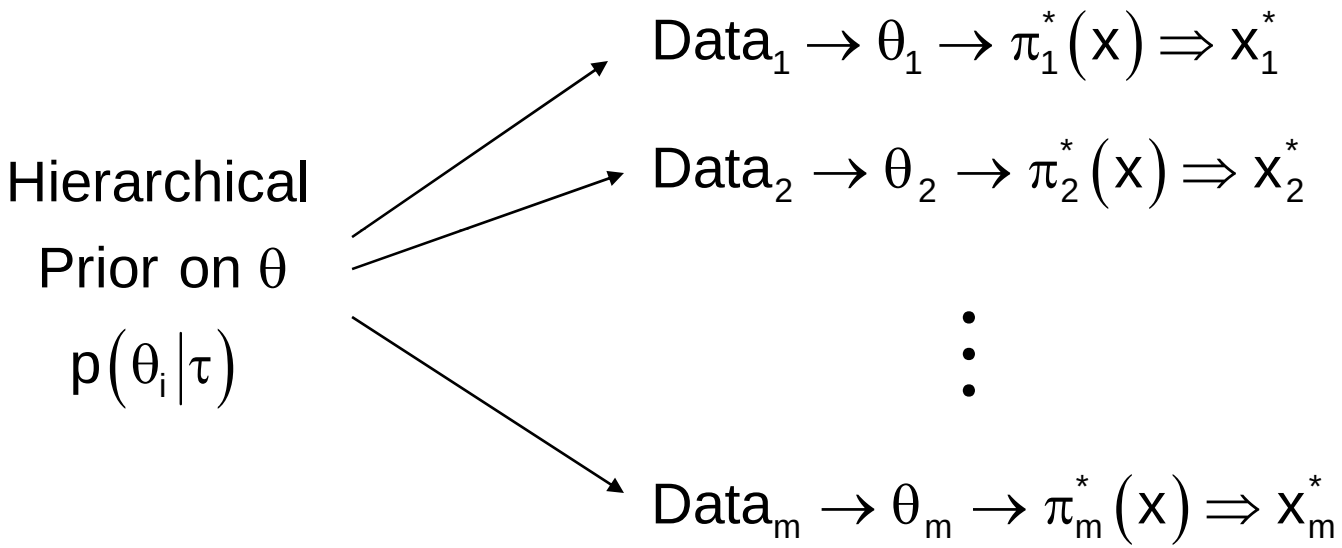
Why full Bayes?

Uncertainty in θ -- NO PLUG-In

Especially important with low amount of information – disaggregate data

$$\pi^{*}(x) = E_{\theta}\left[\overline{\pi}(x|\theta)\right] \neq \overline{\pi}(x|\theta)\Big|_{\theta=\hat{\theta}}$$

Disaggregate decisions



Information: 1. “Borrowed” via Prior 2. Data on Individual Units

Information/customization

Decisions customized to the individual units.

- Product design
- Price/promotion

Extent of Customization *should* depend on:

- Differences in units
- Uncertainty in θ at individual level
  - e.g. with very noisy data, you should be back at uniform decision!

Value of disaggregate information

Compare Profits based on disaggregate with aggregate decisions

Disaggregate Profits:

$$\Pi_{disagg} = \sum_i \pi_i(x_i^*)$$

Optimal Decision for ith unit

Aggregate Profits:

$$\Pi_{agg} = \sum_i \pi_i(x^*)$$

Optimal Uniform Decision

Value of disaggregate information

Valuation through comparison of:

- "Aggregate Profits" -- profits from "uniform" rule or same action for every unit
- "Disaggregate Profits" – profits from full customization allowing for optimal Bayes decisions at each unit

Shows the applied orientation of marketing as a decision and information oriented discipline

Ex: Value of HH Purchase Info

Rossi et al (1995).

What is the value of household level purchase information?

Requires a metric.

Use profits from customized couponing. Print coupons whose face value is customized to the consumer's preferences – to the extent we can infer these using limited data.

Unit-Level Model and Heterogeneity

Diagonal MNP

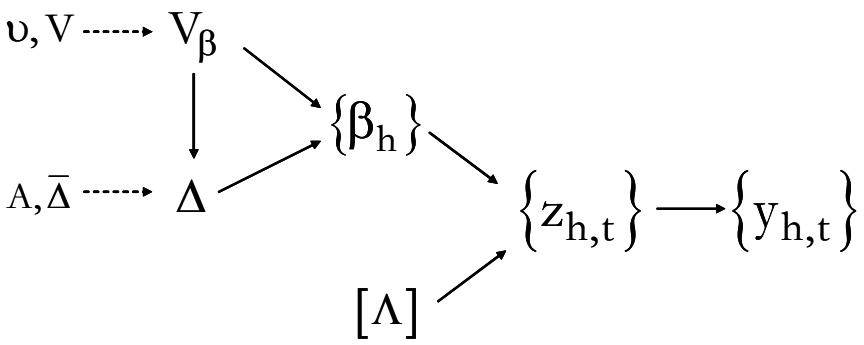
$$y_{h,t} = j \text{ if } \max\left(z_{h,t} = X_{h,t}\beta_h + \varepsilon_{h,t}\right) = z_{j,h,t}$$

$$\varepsilon_{h,t} \sim N(0, \Lambda)$$

Heterogeneity (observed demos and unobs)

$$\beta_h = \Delta' z_h + v_h$$

$$v_h \sim N(0, V_\beta)$$



Information Sets and Decisions

1. Full – purchase history information and demos

   customized coupons – a function of hh data

2. Demos Only – only demo info

   customized coupons – function of demos only

3. “Blanket” – no information about household but info on distribution of parameters across hhs.

   “blanket” or uniform coupon

Predictive Distributions

Full Info Set:

$$p\left(\beta_h \middle| \{y_1,\ldots,y_h,\ldots,y_H,X_1,\ldots,X_h,\ldots,X_H\},Z\right) =$$
$$\int p\left(\beta_h \middle| y_h,X_h,\Lambda,\Delta,V_\beta\right) p\left(\Lambda,\Delta,V_\beta \middle| \{y_1,\ldots,y_h,\ldots,y_H,X_1,\ldots,X_h,\ldots,X_H\},Z\right) d\Lambda d\Delta dV_\beta$$

Demos Only Set:

$$\int p\left(\beta_h \middle| z_h,\Delta,V_\beta\right) p\left(\Delta,V_\beta \middle| \text{Info}\right) d\Delta dV_\beta$$

from a sample of households

Blanket:

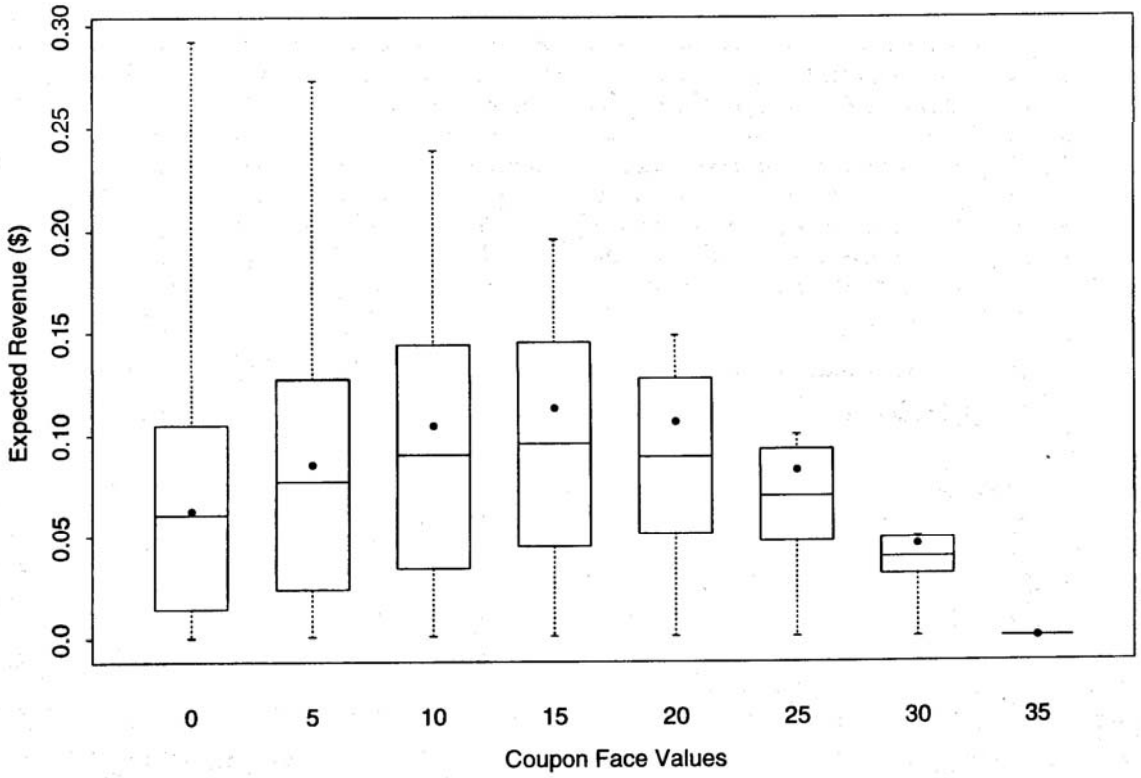
$$\int p\left(\beta_h \middle| z_h,\Delta,V_\beta\right) p(z_h)\, p\left(\Delta,V_\beta \middle| \text{Info}\right) dz_h d\Delta dV_\beta$$

integrate out demos!

Couponing Problem

$$\max_F \pi(F) = \int \Pr\left[i \middle| \beta_h, \Lambda, X(F)\right](M - F)\, p\left(\beta_h, \Lambda \middle| \Omega^*\right) d\beta_h d\Lambda$$

info set



Profit Results

Table 5 Relative Value of the Information Sets

Information Set	Net Revenue	Gain Relative to Blanket
Full	0.1570	2.55
Choices-Only	0.1529	1.93
One Obs	0.1500	1.56
Demos-Only	0.1467	1.12
Blanket	0.1459	1.0
No Coupon	0.1380	