THE BASICS
OF TALKING
BAYESIAN STATISTICS

Hal Stern
Department of Statistics
Iowa State University

Introduction

- Bayesian methods have become popular and their use continues to grow

  Science, 19 November 1999 -
  A 236-year-old approach to statistics is making a comeback, as its ability to factor in hunches as well as hard data finds applications from pharmaceuticals to fisheries

- Bayesian statisticians need to communicate with

  - policymakers and decisionmakers

  - researchers in other fields

  - classically trained statisticians

- Today: comments based on my experiences

  - what do Bayesian methods offer

  - some things that are hard to explain

  - illustrate via several examples

  - summary/conclusions

Terminology

- Bayesian methods

  - treat all unknowns as random variables

  - condition on observed data

  - a.k.a. inverse probability

  - 'Bayesian' is a relatively recent term

- Non-Bayesian methods

  - frequentist? .... but that refers to an interpretation of probability or an approach for evaluating procedures

  - classical? .... but Bayes published first

  - traditional? .... but whose tradition!

What do Bayesian methods have to offer

- They make explicit use of probability

- They allow "flexibility" in inference

- Developing/extending models is a natural process

- They provide a way to incorporate prior information

- Computational benefits

Ex. 1: linear models in animal breeding

- Animal breeding research

  - estimate/predict genetic potential of animals
    for one or more trait

  - use information from each animal and
    its relatives

  - select the best animals for breeding

  - example today: flour beetles
    (Wright, Stern and Berger, JABES 2000)

Ex. 1: linear models in animal breeding

- The model

$$Y = X\beta + Z_1 u + Z_2 c + e$$

  - $Y =$ vector of n trait measurements

  - $X =$ matrix of "fixed" effects (e.g., sex)
  - $\beta =$ vector of p fixed effect coefficients

  - $Z_1 =$ incidence matrix for animal effects
  - $u =$ vector of N animal random effects

  - $Z_2 =$ incidence matrix for family (env.) effects
  - $c =$ vector of q family random effects

  - $e =$ vector of random errors

  - our example (a small problem):
    $n = 324, N = 397, q = 54, p = 6$

Ex. 1: linear models in animal breeding

$$Y = X\beta + Z_1 u + Z_2 c + e$$

- Distributional assumptions

  - $u \sim N(0, \ \sigma_a^2 A)$

    A is known $N \times N$ relationship matrix
  - $c \sim N(0, \ \sigma_c^2 I)$

  - $e \sim N(0, \ \sigma_e^2 I)$

- Normality for u is usually justified by thinking of "infinitesimal" model

- Normality for e may be achieved by transformation

- Possible problems/opportunities

  - major genes

  - outliers (preferential treatments)

Ex. 1: linear models in animal breeding

- Classical approach (REML/BLUP)

  - inference for variance components using restricted maximum likelihood (REML)

  - inference for random effects given variance components

    * solve Henderson's equations

    * equivalent to maximizing $p(u, c, \beta | Y, \sigma^2)$

    $$\begin{pmatrix} X'X & X'Z_1 & X'Z_2 \\ Z_1'X & Z_1'Z_1 + A^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z_1'Z_2 \\ Z_2'X & Z_2'Z_1 & Z_2'Z_2 + I_q\frac{\sigma_c^2}{\sigma_a^2} \end{pmatrix} \begin{pmatrix} \beta \\ u \\ c \end{pmatrix} = \begin{pmatrix} X'Y \\ Z_1'Y \\ Z_2'Y \end{pmatrix}$$

    * yields best linear unbiased predictions (BLUP)

    * may require solving large system of equations (can be thousands of rows)

Ex. 1: linear models in animal breeding

- Bayesian approach

  - probability model

    * $p(y|u, c, \beta, \sigma^2) = N(y|X\beta + Z_1 u + Z_2 c, \ \sigma_e^2 I_n)$

    * $p(u|\sigma_a^2) = N(u|0, \ \sigma_a^2 A)$

    * $p(c|\sigma_c^2) = N(c|0, \ \sigma_c^2 I_q)$

    * $p(\beta) \propto 1$

    * $p(\sigma_a^2, \sigma_c^2, \sigma_e^2) \propto 1/\sigma_e^2$

    * last two are noninformative(?) and improper
      prior distributions

  - Markov chain Monte Carlo sampling
    can be used to study posterior
    (Gibbs sampling is easy to
    implement for this model)

Ex. 1: linear models in animal breeding

- Triboliun castaneum data (flour beetles)

  - $n = 324$ animals

    * one generation out of 16 in study

  - $N = 397$ animals (includes 73 ancestors)

  - $p = 6$ "fixed effects"
    (intercept, gender, date, interactions)

  - $q = 54$ environments

    * 6 beetles taken from each mating and raised in common environment

Ex. 1: linear models in animal breeding

- Variance components

| Parameter | Posterior distribution | | | Poster. | |
	2.5%	50%	97.5%	mean	REML
σ_a^2	757	6301	22919	7781	2253
σ_c^2	290	3881	11396	4371	4426
σ_e^2	30354	40834	50202	40682	43405
heritability	0.014	0.121	0.398	0.145	0.045

Ex. 1: linear models in animal breeding

- Animal effects are of great interest

    - BLUP estimates are posterior means conditional on REML variance estimates

    - full Bayesian posterior means accommodate uncertainty about variance parameters

    - both are motivated by squared error loss

    - a key point is that other summaries of the posterior distn are possible

Ex. 1: linear models in animal breeding

- What do we learn

  - flexible inferences (e.g., ranks)

  - incorporate prior information at some levels of
    the model (i.e., family structure) ... but not at
    others

  - computationally convenient

- What questions come up

  - defining optimal point estimates
    (post. mean, post. median, post. mode)

  - role of prior distributions for variances

Ex. 2: estimation of soil texture quantiles

- Soil texture or composition variables
  (e.g., % clay) are important for land management

- Goal: estimate quantiles of distn of soil texture
  variables for 48 inch profile

- Data available

  - horizon = layer of soil

  - S = surface points (field measurements only
    on surface horizon, 6-12 in)

  - F = field points (field measurements on all
    horizons up to 48 in)

  - L = lab points (field and lab data on all horizons
    up to 48 in)

  - lots of missing data
    (e.g., lab measurements on S and F sites)

- Data are from National Cooperative Soil Survey
  (NCSS), a USDA/state agencies production

- Ph.D. thesis of P. J. Abbitt (ISU, 1999)

Ex. 2: estimation of soil texture quantiles

- Calibration/imputation approach

  - calibration

    * linear model used to calibrate field and lab measurements

    * lab measurements modeled as a function of field measurements (and other variables) using sites with both types of data

    * "predict" lab measurements for sites in S/F

    * different models used for different types of soils

Ex. 2: estimation of soil texture quantiles

- Calibration/imputation approach (cont'd)

    - imputation

        * after calibration have complete "lab" data except for sites in S

        * next impute calibrated lab data for full profile for sites in S

        * regression model to impute inch i's data using inch 1's data as predictor

        * different models used for different types of soils

    - quantile function estimator

    - variance estimation via jackknife

Ex. 2: estimation of soil texture quantiles

- Hierarchical probability model

  - normal linear model relates field measurements to lab measurements (similar to calibration step)

  - lab measurements model
    * normal distribution

    * site random effect

    * mean and variance differ by horizon type (A, B, C)

  - horizon profile is a Markov model (six states at any inch: continue A, start A, continue B, start B, continue C, start C)

- Bayesian inference via MCMC (priors are vague distns centered at values based on science)

- Quantile estimates derived from a mixture of normal (lab measurement) distns with mixture weights a function of the Markov model probs

Ex. 2: estimation of soil texture quantiles

- Bayesian analysis vs calibration/imputation

  - computation is more difficult for the Bayesian approach

  - calibration approach is series of separate steps

  - more distributional assumptions for Bayesian approach (requires more model checking)

  - Bayesian approach does not rely on asymptotics

  - Bayesian approach provides flexible inference (more/better output)

  - as an example we consider:
    quantiles for % clay in Old Alluvium soils

Things that can be hard to explain

- Interpretation of probability

- The role of prior distns
  (and noninformative prior distns)

- What is the right point estimate?

- How do we test hypotheses?

- Computational difficulties

Interpretation of probability

- Most people have been (and continue to be) trained with the frequentist view of probability

- A common claim is that people relate more naturally to the "Bayesian" (subjective?) view, e.g., in confidence intervals

- This has not always been the case in my consulting/collaborative experiences

- The better-trained the person the more likely this is to be a problem

The role of prior distributions

- Many want to use the likelihood function as a distn on parameters (e.g., Fisher's fiducial inference) but are unwilling to specify a prior distn

- This has always been the most problematic part for people

- Subjective prior distns often worry people

- The fact that the prior distn doesn't matter much for large sample size helps ....
  but not always true in hierarchical models

- In practice people are drawn to noninformative prior distns but ...
  - "noninformative" is not well defined
  - "flat" and "vague" are clear concepts but don't always yield noninformative prior distns
  - this subject is very confusing to people

- Recommendation might be to use vague proper prior distns and perform a sensitivity analysis

Point estimates

- MLE is the dominant point estimation technique

- To get a unique Bayesian point estimate one must specify a loss function

- Example: animal breeding

Parameter	Posterior distribution 2.5%	50%	97.5%	Poster. mean	REML
σ_a^2	757	6301	22919	7781	2253
σ_c^2	290	3881	11396	4371	4426
σ_e^2	30354	40834	50202	40682	43405
heritability	0.014	0.121	0.398	0.145	0.045

  – REML estimates differ from posterior mean/median

  – Why? ML is more like a mode

Testing Hypotheses

- Applied science research is still often dominated by .05 significance level tests

- Can sometimes be addressed in a Bayesian analysis by examining 95% posterior interval
  (but not always, e.g., variance components)

- Bayes factors are an option

  - can be difficult to compute

  - sensitive to prior distns

  - interpretation?

- The best solution is to shift emphasis to point estimation and posterior intervals

Ex. 3: binary regr with random effects

- Natural selection study

- 244 turtles in 31 families

- Response: survival (0/1)

- Predictors: birthweight, clutch/family effects

- Data collected by Janzen et al. (Ecology, 2000)

- It turns out that birthweight is a significant factor

- Question: are clutch effects important?

Ex. 3: binary regr with random effects

- Probit model with random effects (Sinharay and Stern, 2000 technical report)

- Posterior inference
  - coefficient of birthweight
    * post. mean = .38, post. s. d. = .10
    * 95% int. = (.18, .59)

  - variance component (σ^2)
    * post. mean = .32, post. s. d. = .12
    * post. mode = .26
    * 95% int. = (.16, .61)

- Bayes factor (simple probit model vs probit with random effects): 3.25

- Bayes factor is sensitive to the prior on σ^2

- Conclude variance parameter is not significant

Summary/Conclusions

- Need to emphasize the difference between procedure development and procedure evaluation

- Outlook for increased use of Bayesian procedures is good

  - flexible inference

  - natural to develop/extend models (e.g., missing data)

  - success stories continue to develop

- Cautions

  - need to educate as we work

  - more work to be done on model checking/sensitivity analysis

  - these are generally not black-box procedures, we need to work with scientists