Minimally Informative Nonparametric Bayesian Analysis

Christopher Bush, Novartis

Juhee Lee, Steven MacEachern, Ohio State University

Outline

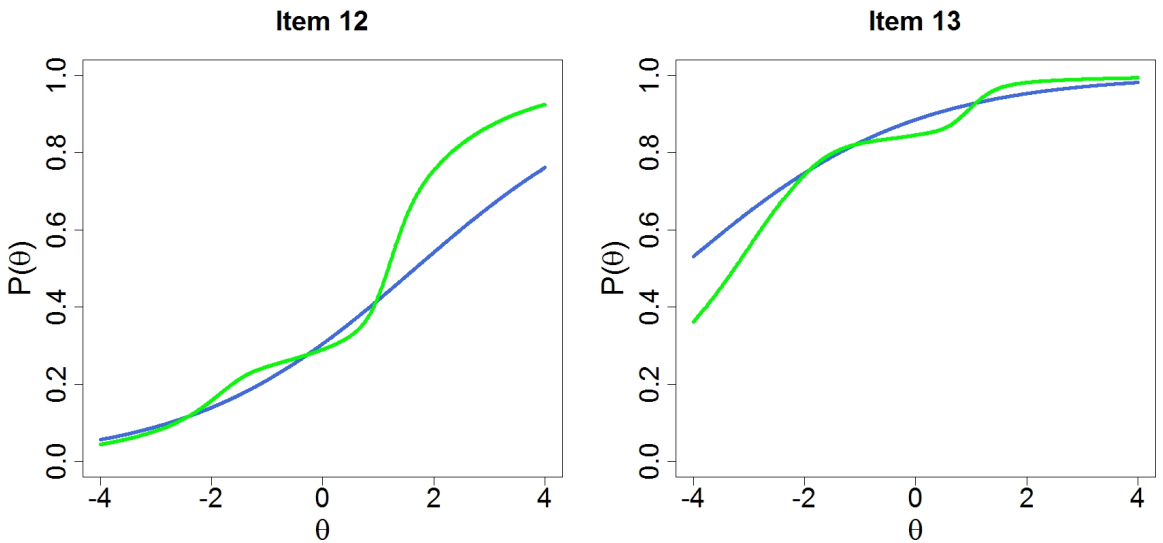
- Motivation
- Noninformative np Bayes
- Local mass
- Consistency issues
- A few pictures
- Kernel density estimates
- Current exploration

Motivation

- **Many successes for nonparametric Bayesian methods**
  Success judged by
  - **Goodness of fit vs. parametric model**
  - **Predictive performance vs. parametric model**
  - **Bayes factor vs. parametric model**
  - **Compelling argument as to why a np model is "right"**
- **Two reasons for success**
  - **Better "mean" function**
  - **Better residual distribution**

Np Bayes Item Response Theory - Qin, Duncan

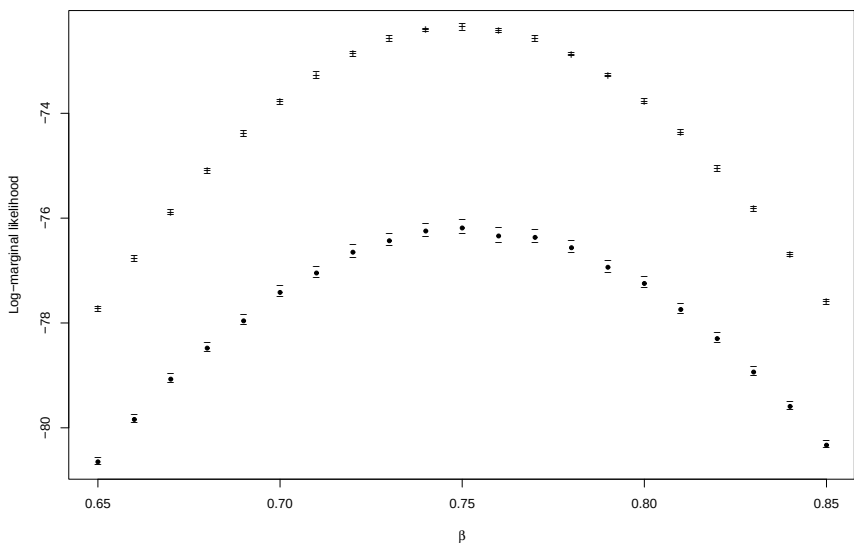
- Two curves showing a difference between the parametric and non-parametric models



- The npBayes model leads to a smaller lack of fit
- npBayes yields better predictions

Regression with semiparametric error dist'n - Guha

- npBayes curve higher, npBayes curve greater curvature implies tighter posterior on slope



- Distribution of covariate plays strong role in determining whether parametric or nonparametric model has tighter posterior

Rarer motivations

- **Less common is contrast between np Bayes models**
  - **Form in which covariates enter model**
  - **Choice of underlying process**
  - **DDP's and many variants**
- **Even less common is to stretch npBayes procedures toward parametric procedures**

Bayesian one-way anova

- **Traditional ANOVA model**

$$\begin{aligned} X_{i1}, \ldots, X_{in_i} | \theta_i, \sigma_i^2 &\sim N(\theta_i, \sigma_i^2) \\ \sigma_i &= \sigma^2 \\ G &\sim \mathbf{Dir}(\alpha) \\ \theta_1, \ldots, \theta_k &\sim G \end{aligned}$$

- **Decompose base measure into mass, base cdf**
- **Take G_0 to be $N(\cdot, \tau^2)$**
- **Simplest model assumes known σ^2. More useful models place a prior on σ^2**
- **Posterior inference over treatment means. All partitions receive positive posterior probability**

Noninformative DP?

- **Tradition in npBayes follows that of other Bayes settings - a non-informative analysis matches the results from some targeted classical analysis**
- **In the context of np Bayes, Ferguson: Target for a distribution function is the empirical cdf**

$$G \sim \mathbf{Dir}(MG_0)$$

$$\hat{G}_B(t) = \frac{MG_0(t) + n\hat{G}(t)}{M+n}$$

  - **When $M \to 0$, recover the empirical cdf**
  - **Susarla and van Ryzin: same holds for survival analysis; exact and right censored observations lead to Kaplan-Meier**
- **Same sort of result holds for typical mixture model - countable, finite with large enough number of components, etc.**

$M \to 0$

- **Use s_i to denote a partition of the treatments**
- **Look at posterior via posterior odds of partitions**

$$\begin{aligned} \frac{\pi(s_i|x)}{\pi(s_j|x)} &= \frac{\pi(s_i)f(x|s_i)}{\pi(s_j)f(x|s_j)} \\ &= \frac{\pi(s_i)}{\pi(s_j)}c \end{aligned}$$

If s_i has more components than s_j

$$\to \ 0$$

If s_i, s_j have the same number of components

$$\to \quad \textbf{in } (0, \infty)$$

- **Limiting posterior concentrates all mass on $\theta_1 = \ldots = \theta_k$. This does not depend on the data!**

$\tau^2 \to \infty$

- **Posterior odds of partitions**

$$\begin{aligned} \frac{\pi(s_i|x)}{\pi(s_j|x)} &= \frac{\pi(s_i)f(x|s_i)}{\pi(s_j)f(x|s_j)} \\ &= c\frac{f(x|s_i)}{f(x|s_j)} \end{aligned}$$

If s_i has more components than s_j

$$\to 0$$

This follows from the difference in dimensionality of s_i and s_j

- **Again, limiting posterior concentrates all mass on $\theta_1 = \ldots = \theta_k$. This does not depend on the data!**

Local mass

- **Define** $M_t = M\sqrt{2\pi}\tau_t$
- **Let** τ_t **tend to** ∞

$$lim_{t\to\infty} M_t([a,b]) = M[b-a]$$

- **Implications for posterior across partitions**

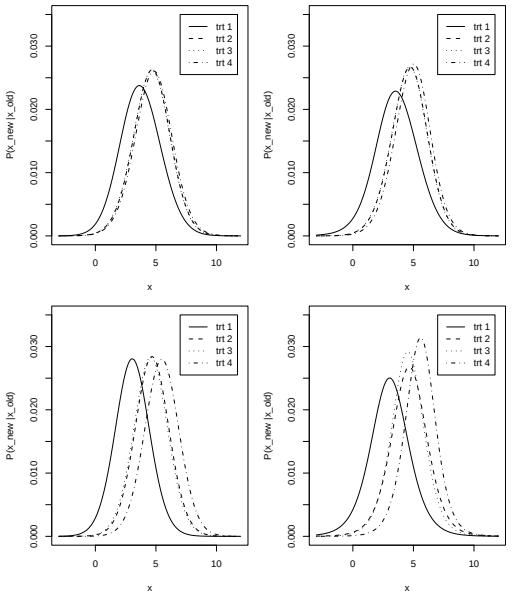
$$\frac{\pi(s_i|x)}{\pi(s_j|x)} \in (0,\infty)$$

- **All partitions receive positive posterior probability**
- **Result follows from formal calculation**
- **Alternatively, result follows from conditional used for Gibbs sampler**
  - **Incremental update: all changes considered have positive limiting conditional posterior probability**
  - **Finite state space Markov chain**
  - **Limiting distribution of Markov chain assigns positive probability to all states**

Consistency issues

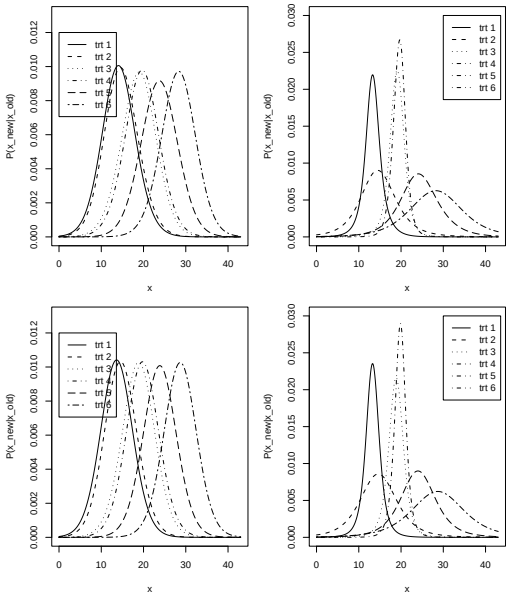
- **Consistency in this setting for**
  - **treatment means, variances**
  - **equality of treatment means and/or variances**
  - **probability statements about** $X_{1\cdot}$
- **Conditions are loosely**
  - **sample sizes large enough to guarantee all partition specific posteriors are proper**
  - **finite means and variances for all treatments**
  - **normality of distributions**
- **Interestingly, under some conditions, one may be consistent for the treatment means, but not for equality of the treatment means**

Sucrose data, $n_i = 8$



- Left: equal variances; Right: variances differ

Rhizobium data, $n_i = 5$



- Left: equal variances; Right: variances differ

Suggestions

- **Structure of prior distribution**
  - **Change $\pi(\tau^2)\pi(M)$ to $\pi(\tau^2)\pi(\text{local mass})$**
  - **Sampling is still doable, though with some Metropolis steps**
- **Inferential problem**
  - **Careful description of inference, may matter under wrong-model asymptotics**
- **Overdispersed priors**
  - **Often used in parametric Bayesian analysis**
    preference for underweighting rather than overweighting prior
  - **Often used in np Bayesian analysis**
  - **Mass parameter isn't really mass parameter when coupled with overdispersed prior**

Recovering the kde

- **Traditional kernel density estimate has several features**
  - **No shrinkage: kernels centered at data points implies $\tau^2 = \infty$**
  - **No clustering: one kernel per data point; data points not grouped implies $M = \infty$, a big ∞, as local mass must be infinite**
  - **Kernels of some shape (play role of likelihood)**
- **Posterior concentrates on "all distinct" event**
- **Inference conditional on new case clustering with existing case leads to traditional kde**
- **Still no reason to change bandwidth with n, much less to change according to "optimal rate"**

Current explorations

- As part of her dissertation, Juhee is looking at connection between multiple shrinkage estimation and np Bayes
  - James-Stein: shrink a set of normal means toward common value
  - Choose a set of subspaces; a shrinkage estimator for each
  - Combine the estimators "with" Bayes theorem
- np Bayes "shrinks" by clustering data values
  allows shrinkage to all subspaces defined by possible partitions
- np Bayes decides point of shrinkage through prior
  minimally informative technique allows data driven choice of this point
- Variant on usual models also allows shrinkage toward subspace
- Early results look promising - beat James-Stein, beat best of multiple shrinkage estimators on empirical examples