Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression

John R. Lewis   Steven N. MacEachern   Yoonkyung Lee

Department of Statistics, The Ohio State University

Feb. 9, 2022

Introduction

- Formal Bayesian inference relies on the prior, likelihood, loss
- Each require assumptions that should be questioned
- We focus this paper on imperfections in the Likeihood:
    - Start with a full model as if it is correct
    - Sumarise the data with a summary statistic
    - The prior is updated with the summary statistic rather than the complete data

Resticted Likelihood Examples

- Outliers
  - Known subset of bad data are removed prior to analysis.

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\prod_{i=1}^{n-k} f(y_i|\boldsymbol{\theta})\right)\left(\prod_{i=n-k+1}^{n} f(y_i|\boldsymbol{\theta})\right)$$

Resticted Likelihood Examples

- Censoring
  - Reaction Time Experiments (e.g., Ratcliff, 1993)
  - Some reactions are too fast or too slow to be believable
- $c(y_i) = t_1$ if $y_i \leq t_1$, $c(y_i) = t_2$ if $y_i \geq t_2$, and $c(y_i) = y_i$ otherwise.

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\prod_{i=1}^{n} g(c(y_i)|\boldsymbol{\theta})\right)\left(\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}, c(y_i)).\right)$$

Resticted Likelihood Generalization

- Conditioning statistic $T(\boldsymbol{y})$
  - Example 1: $T(\boldsymbol{y}) = (y_1, \ldots, y_{n-k})$
  - Example 2: $T(\boldsymbol{y}) = (c(y_1), \ldots, c(y_n))$

$$L(\boldsymbol{\theta}|\boldsymbol{y}) \quad = \quad f(T(\boldsymbol{y})|\boldsymbol{\theta}) \; f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}))$$

- Restricted Likelihood Posterior

$$\pi(\boldsymbol{\theta}|T(\boldsymbol{y})) \quad = \quad \frac{\pi(\boldsymbol{\theta})f(T(\boldsymbol{y})|\boldsymbol{\theta})}{m(T(\boldsymbol{y}))}$$

- Predictive Density

$$f(y_{n+1}|T(\boldsymbol{y})) = \int f(y_{n+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|T(\boldsymbol{y})) \; d\boldsymbol{\theta}$$

Literature Review

- Rank Likelihoods: Savage (1969), Pettitt (1983, 1982), Hoff et al. (2013).
- Order Statistics: Lewis et al. (2012)
- Asymptotics: Doksum and Lo (1990), Clarke and Ghosh (1995), Yuan and Clarke (2004), and Hwang et al. (2005)
  - Often, posterior distribution resembles the asymptotic sampling distribution of the conditioning statistic
- Approximate sufficiency of mean/sd: Pratt (1965)

Literature Review: Approximate Bayesian Computation

- Posterior approximation with success in many applications: (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002; Marjoram et al., 2003; Fearnhead and Prangle, 2012; Drovandi et al., 2015).
- $L(\boldsymbol{\theta}|\mathcal{B}(\boldsymbol{y}))$, where $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{y}^*|\rho(T(\boldsymbol{y}), T(\boldsymbol{y}^*)) \leq \epsilon\}$
  - metric ρ, tolerance level ϵ
- Goal is often to approximate the full posterior
- Choose an approximately sufficient (Joyce and Marjoram, 2008) $T(\boldsymbol{y})$ and small ϵ
- Sampling Methods: Standard reject (Pritchard et al., 1999), Extensions: Beaumont et al. (2009); Turner and Van Zandt (2012, 2014)

Application to the Linear Model

$$
\begin{aligned}
\boldsymbol{\theta} &= (\boldsymbol{\beta}, \sigma^2) \sim \pi(\boldsymbol{\theta}) \\
y_i &= x_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ for } i = 1, \ldots, n
\end{aligned}
$$

- $T(\boldsymbol{y}) = (\boldsymbol{b}(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$
  - $\boldsymbol{b}(X, \boldsymbol{y}) = (b_1(X, \boldsymbol{y}), \ldots, b_p(X, \boldsymbol{y}))^\top \in \mathbb{R}^p$
  - $s(X, \boldsymbol{y}) \in \{0\} \cup \mathbb{R}^+$
- E.g. M-estimators Huber (1964), Least Median Squares, Least Trimmed Squares

Computational Strategy

- Direct Sampling (Small Dimensions): Relies on KDE of $L(\theta|T(y))$ Lewis (2014)
- MCMC: Data augmented Gibbs Sampler targeting $f(\boldsymbol{\theta}, \boldsymbol{y}|T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$
  1. $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})) = \pi(\boldsymbol{\theta}|\boldsymbol{y})$ (full poserior)
  2. $f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$
- For 2, prosose a Metropolis-Hastings sampler.
- accept/reject a sample full data $\boldsymbol{y} \in \mathcal{A} := \{\boldsymbol{y} \in \mathbb{R}^n | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})\}$ from a well defined distribution with support $\mathcal{A}$.

Computational Strategy

- Difficult to sample from $\mathcal{A}$ directly
- Can sample $\boldsymbol{z}^* \in \mathbb{R}^n$ and transform:

$$\boldsymbol{y} = h(\boldsymbol{z}^*) := \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*)\right)$$

- Under certain conditions on regression/scale estimators (C3-C8 in the paper), $T(y) = T(y_{obs})$
- Idea:
  - sample $\boldsymbol{z}^* \sim p(\boldsymbol{z}^*)$, transform via h
  - proposal $p(\boldsymbol{y}|\boldsymbol{\theta})$ is then a change-of-variables adjustment on $p(\boldsymbol{z}^*)$.

Computational Strategy

- h is not 1-1/onto. (Change of variables difficult)
- Can restrict sample space of z^*, so that it is
  - $\mathcal{A}$ is an $n - p - 1$ space
  - Sample space for z^*: $\mathbb{S} := \{\boldsymbol{z}^* \in \mathcal{C}^\perp(X) \mid ||\boldsymbol{z}^*|| = 1\}$
    - i.e. the unit space in the orthogonal complement of the column space of the design matrix.
  - $h : \mathbb{S} \to \mathcal{A}$ is then 1-1/onto.
  - easier to figure out the Jacobian of the transformation from $p(\boldsymbol{z}^*)$ to $p(\boldsymbol{y}|\boldsymbol{\theta})$

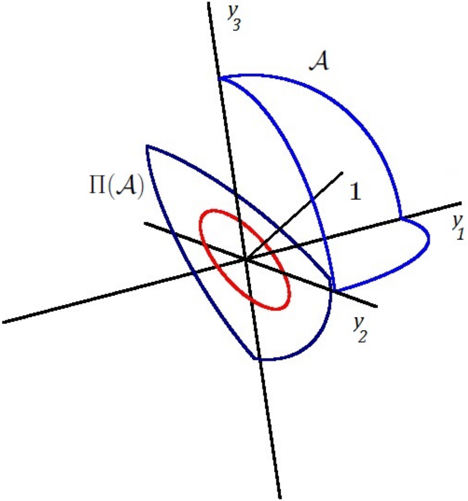
Computational Strategy

For the proposal distribrution.

1. Sample $\boldsymbol{z}^*$ from a distribution with known density whose support is the entirety of $\mathbb{S}$.
2. Set $\boldsymbol{y} = h(\boldsymbol{z}^*)$
3. Compute the Jacobian (think about it in 2 steps)
   - $\boldsymbol{z} = \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*$. Scale from $\mathbb{S}$ to the set $\Pi(\mathcal{A}) := \{\boldsymbol{z} \in \mathbb{R}^n | \ \exists \ \boldsymbol{y} \in \mathcal{A} \ s.t. \ \boldsymbol{z} = Q\boldsymbol{y}\}$ with $Q = I - XX^\top$.
   - $\boldsymbol{y} = \boldsymbol{z} + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \boldsymbol{z})\right)$. Shift from $\Pi(\mathcal{A})$ to $\mathcal{A}$ along $\mathcal{C}(X)$

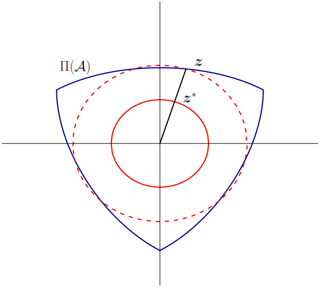
Computational Strategy (Visualization, $n = 3, p = 1$)

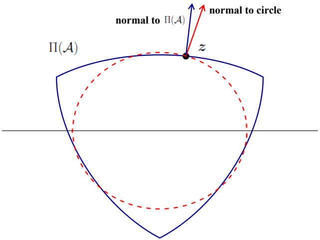
$$T(\boldsymbol{y}) = (\min(\boldsymbol{y}), \sum(y_i - \min(\boldsymbol{y}))^2), \; T(\boldsymbol{y}_{obs}) = (0, 1)$$



Computational Strategy (Visualization, $n = 3, p = 1$)

Scaling step





- resize sphere: $r^{-(n-p-1)}$
- deformation onto $\Pi(\mathcal{A})$: $\cos(\gamma) = \dfrac{\nabla s(X,\boldsymbol{y})^\top \boldsymbol{z}}{\|\nabla s(X,\boldsymbol{y})\| \|\boldsymbol{z}\|}$

Computational Strategy (Visualization, $n = 3, p = 1$)

Shifting step of $\boldsymbol{z}$ to $\boldsymbol{y}$ along the column space of X

- Contribution is the ratio of the infinitesimal volumes along $\Pi(\mathcal{A})$ at $\boldsymbol{z}$ to the corresponding volume along $\mathcal{A}$ at $\boldsymbol{y}$.
- $\mathrm{Vol}(P) := \sqrt{\det(P^\top P)} = \prod_{i=1}^{r} \sigma_i$
  - $P = QA$,
  - columns of A form an orthonormal basis for the tangent space to $\mathcal{A}$ at $\boldsymbol{y}$. Can be found from $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$
  - σ_i are the singular values of P

Full Jacobian: $p(\boldsymbol{y}) = p(\boldsymbol{z}^*) r^{-(n-p-1)} \cos(\gamma) \mathrm{Vol}(P)$

Simulation Example 1

- hierarchical setting with outliers.

$$\theta_i \sim N(0,1),\ i = 1, 2, \ldots, 90$$
$$y_{ij} \sim (1 - p_i)N(\theta_i, 4) + p_i N(\theta_i, 4m_i),\ j = 1, 2, ..., n_i$$

- $p_i = .1, .2, .3,\ m_i = 9, 25,$ and $n_i = 25, 50, 100$
- 5 groups for each combination, 90 groups total
- Base model for fitting:

$$\theta_i \sim N(\mu, \tau^2),\ \sigma_i^2 \sim IG(a_s, b_s),\ i = 1, 2, \ldots, 90,$$
$$y_{ij} \sim N(\theta_i, \sigma_i^2),\ j = 1, 2, \ldots, n_i.$$

- Restricted likihood fit: Robust M-estimators for each group: $T_i(y_{i1}, \ldots, y_{in_i}) = (\hat{\theta}_i, \hat{\sigma}_i^2), i = 1, 2, ..., 90.$

K = 30 simulations, M indexes the method, MSE

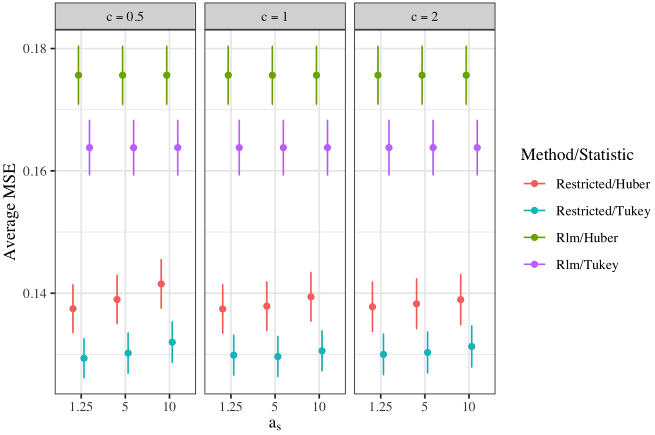


Figure: Average MSE plus/minus one standard error for each value of a_s and c. Smaller values represent better fits. The panels correspond to $c = 0.5$ (left), $c = 1$ (middle), and $c = 2$ (right), with the values of a_s on the horizontal axis. The average MSE for the normal theory model ranges from 0.24 to 0.25 and is left out of the figure.

Simulation Example 2

Data: several correlated covariates, only a few govern data

- $y = \boldsymbol{\beta}^\top x + \epsilon$
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$
- $\epsilon \sim N(0, \sigma^2)$ with probability 0.8 and
  $\epsilon \sim$ Half-Normal$(0, 25\sigma^2)$ with probability 0.2
- $x_1 \sim N(0, 1)$ and $x_j = x_1 + \eta_j$ with $\eta_j \sim N(0, 4)$ for $j = 2, 3$
- additional covariates: x^* 27 additional covariates
- 21 generated independently, 6 are x_1, x_2, and x_3 with random noise.

Model used for fitting:

- $y = \beta^\top x + \beta^{*\top} x^* + \epsilon$
- $\boldsymbol{\beta}_{all} \sim N_{20}(\mathbf{0}, \sigma_\beta^2 I)$ with $\boldsymbol{\beta}_{all} = (\boldsymbol{\beta}, \boldsymbol{\beta}^*)^\top$ and $\sigma^2 \sim IG(5, 8)$

$K = 30$ simulations, n $= 500$, MSE

$$MSE = (||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2 + ||\hat{\boldsymbol{\beta}}^*||^2)/30$$

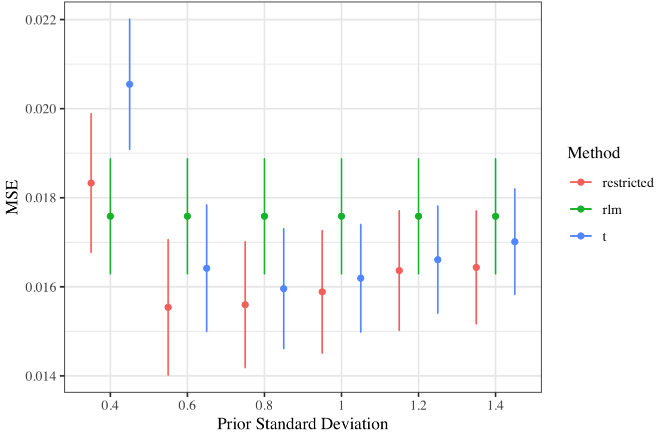


Figure: Average MSE plus/minus one standard error over the $K = 30$ simulations for each value of the prior standard deviation (σ_β) and each of the fitting methods.

Prediction of non-outlying data

- $MNLL = -\frac{1}{N}\sum \log f(y_i|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^*, \hat{\sigma})$, f is the assumed likelihood, average over non-outlying points

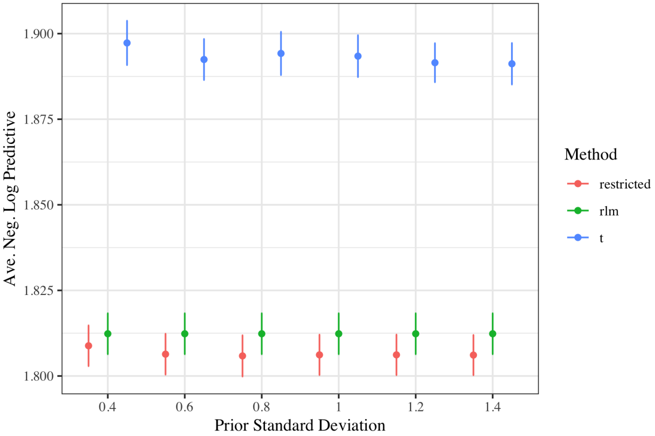


Figure: Average MNLL plus/minus one standard error for each value of the prior standard deviation (σ_β)

Real Data: Insurance Data

- Interested in future performance of agencies that have varying contractual agreements

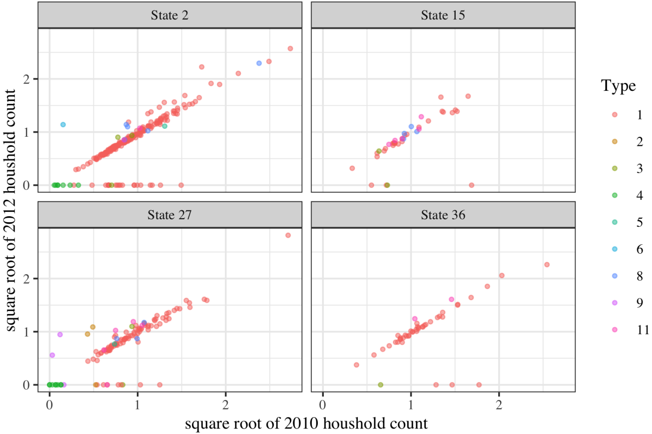


Figure: The square root of (scaled) count in 2012 versus that in 2010 for four states.

Real Data: State Level Regressions

Regression fit seperately within each state

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \Sigma_0); \quad \sigma^2 \sim IG(a_0, b_0); \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

- Covariates: square root of household count in 2010, two different size/experience measures
- response: square root of household count in 2012
- hyper-parameters fixed via regression of data in time-period two years before.
- model misspecification: omitting contract type, closure information
- causes many cases to appear "outlying"

Method of Model Comparison

- both training and validation data will contain outliers
- timming approach with log predictive density (Jung et al., 2014)
- case i in holdout set: $log(f(y_i))$
- Scoring a procedure:
  - Choose a base method (e.g. Student-t model) and trimming fractions α
  - Order holdout sample by $log(f_b(y_i))$
  - Denote ordering by: $y^b_{(1)}, y^b_{(2)}, \ldots, y^b_{(M)}$
  - score each method with "mean trimmed log margainal psuedo likelihood"

$$TLM_b(A) = (M - [\alpha M])^{-1} \sum_{i=[\alpha M]+1}^{M} \log(f_A(y^b_{(i)})),$$

- f_A - predictive distribution under the method "A" being scored.

Predictive Performance: 50 training/holdouts sets

Evaluation of 'Type 1' agencies (of special interest to the company)

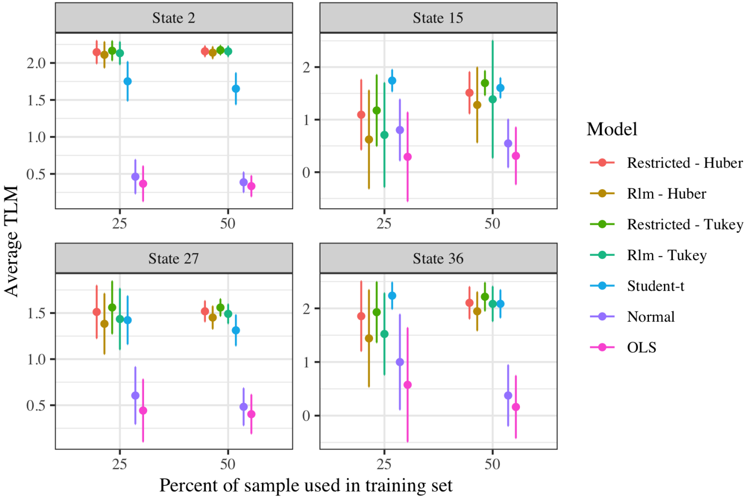
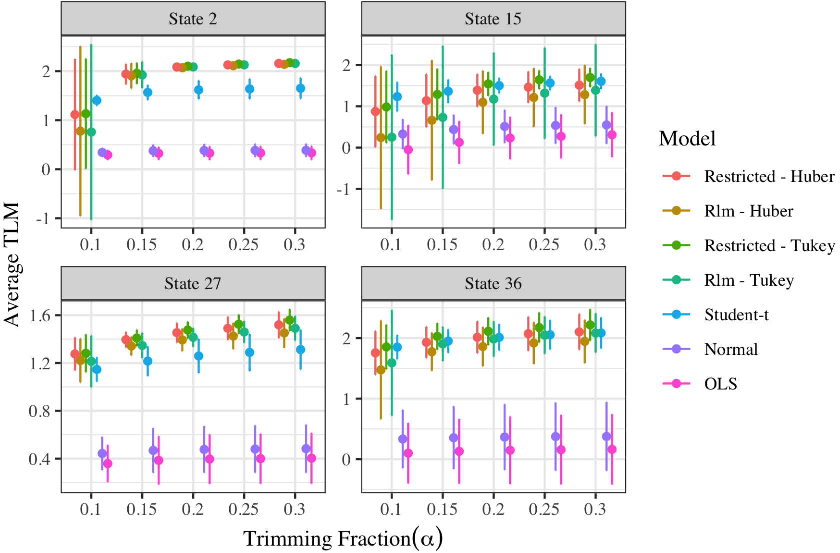


Figure: $\alpha = 0.3$. States 2, 15, 27, and 36, have $n = 222, 40, 117$, and 46

Predictive Performance: changing trimming fraction

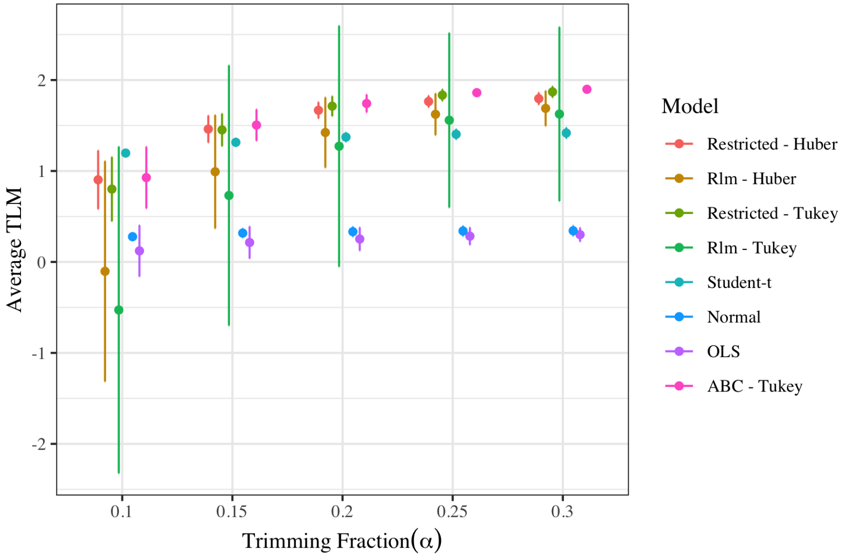


Real Data: Hierarchical Regression

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \overset{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0);$$

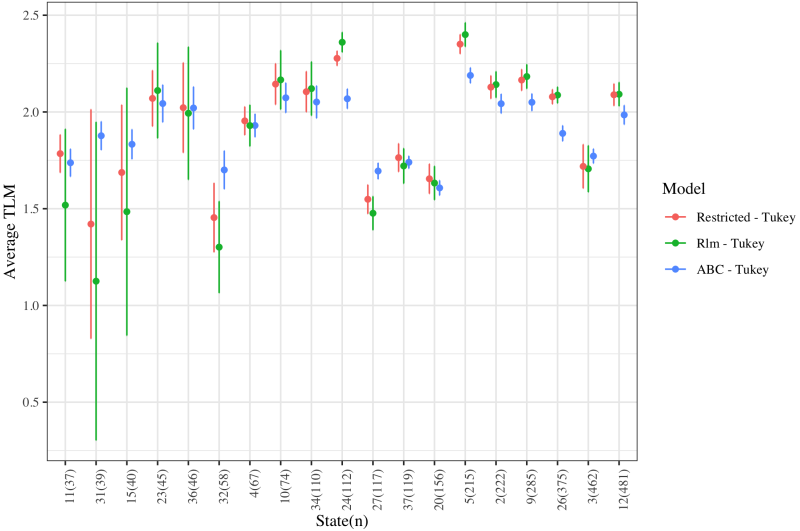
$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2), \; i = 1, \ldots, n_j, \; j = 1, \ldots, J$$

Predictive Performance



Predictive Performance by State

- restricted likelihood average TLM is larger than ABC in 14 of the 20 states, median difference of 0.04



References I

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. Biometrika, 96(4): 983–990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. Genetics, 162:2025–2035.

Clarke, B. and Ghosh, J. K. (1995). Posterior convergence given the mean. The Annals of Statistics, 23:2116–2144.

Doksum, K. A. and Lo, A. Y. (1990). Consistent and robust Bayes procedures for location based on partial information. The Annals of Statistics, 18:443–453.

Drovandi, C., Pettitt, A., and Lee, A. (2015). Bayesian indirect inference using a parametric auxiliary model. Statistical Science, 30: 72–95.

References II

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society: Series B, 74:419–474.

Hoff, P., Fosdick, B., Volfovsky, A., and Stovel, K. (2013). Likelihoods for fixed rank nomination networks. Network Science, 1:253–277.

Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1): 73–101.

Hwang, H., So, B., and Kim, Y. (2005). On limiting posterior distributions. Test, 14:567–580.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biology, 7(1).

References III

Jung, Y., MacEachern, S., and Lee, Y. (2014). Cross-validation via outlier trimming. In preparation.

Lewis, J. (2014). Bayesian Restricted Likelihood Methods. PhD thesis, The Ohio State University.

Lewis, J., Lee, Y., and MacEachern, S. (2012). Robust inference via the blended paradigm. In JSM Proceedings, Section on Bayesian Statistical Science, pages 1773–1786. American Statistical Association.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America, 100:15324–15328.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. Journal of the Royal Statistical Society. Series B, 44:234–243.

References IV

Pettitt, A. N. (1983). Likelihood based inference using signed ranks for matched pairs. Journal of the Royal Statistical Society. Series B, 45:287–296.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. Journal of the Royal Statistical Society. Series B, 27:169–203.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of y chromosome microsatellites. Molecular Biology and Evolution, 16:1791–1798.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychological Bulletin, 114:510.

Savage, I. R. (1969). Nonparametric statistics: A personal review. Sankhya: The Indian Journal of Statistics, Series A (1961-2002), 31:107–144.

References V

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. Genetics, 145:505–518.

Turner, B. M. and Van Zandt, T. (2012). A tutorial on approximate bayesian computation. Journal of Mathematical Psychology, 56(2):69–85.

Turner, B. M. and Van Zandt, T. (2014). Hierarchical approximate bayesian computation. Psychometrika, 79(2):185–209.

Yuan, A. and Clarke, B. (2004). Asymptotic normality of the posterior given a statistic. The Canadian Journal of Statistics, 32:119–137.