



Libraries and Learning Services

University of Auckland Research Repository, ResearchSpace

Suggested Reference

Fernando, M. A. C. S., Curran, J. M., & Meyer, R. (2016, November 21). Performance and limitations of likelihood based information criteria and leave-one-out cross-validation approximation methods. In *18th International Conference on Statistics and Analysis*. Singapore.

Copyright

Items in ResearchSpace are protected by copyright, with all rights reserved, unless otherwise indicated. Previously published items are made available in accordance with the copyright policy of the publisher.

For more information, see [General copyright](#).

Performance and Limitations of Likelihood based Information Criteria and Leave-one-out Cross-validation Approximation Methods

M.A.C.S.Sampath Fernando
James M. Curran
Renate Meyer

Department of Statistics
University of Auckland
New Zealand

sampathf73@gmail.com

November 21, 2016

Overview

1 Introduction

- Modelling the behaviour data
- Bayesian statistical models
- Electropherogram (EPG)

2 Statistical models for stutter ratio (SR)

- Mean and variance of SR
- Different Models for SR

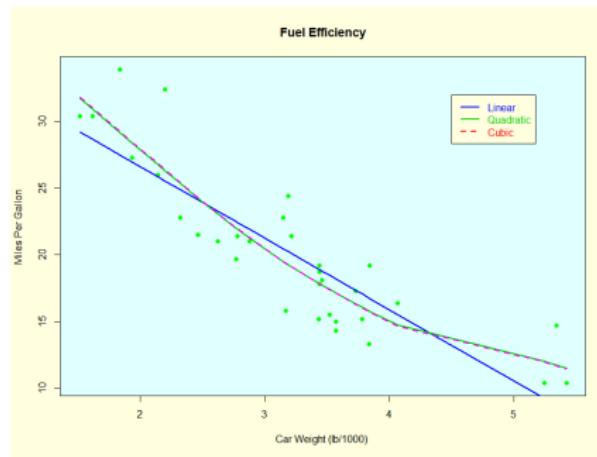
3 Bayesian model comparisons

- Information criteria
- Limitations of information criteria
- LOO-CV and LOO-CV approximations

4 Results and findings

Introduction

- What is a statistical model?
 - A probabilistic system
 - A finite/infinite mixture of probability distributions



- All models are approximations
- “All models are wrong, but some are useful” - George Box (1976)

Modelling the Behaviour of Data

- If we wish to perform statistical inference, or use our model to probabilistically evaluate the behavior of new observations, then we need three steps:
 - ① Assume that the data are generated from some statistical distribution
 - ② Write down equations for the parameters of the assumed distribution, e.g. the mean and the standard deviation
 - ③ Use standard techniques to estimate the unknown parameters in 2
- Steps 1-3 should be repeated as often as possible to get the "best" model
- Most model building consists of steps 2 and 3
- Classical and Bayesian approaches

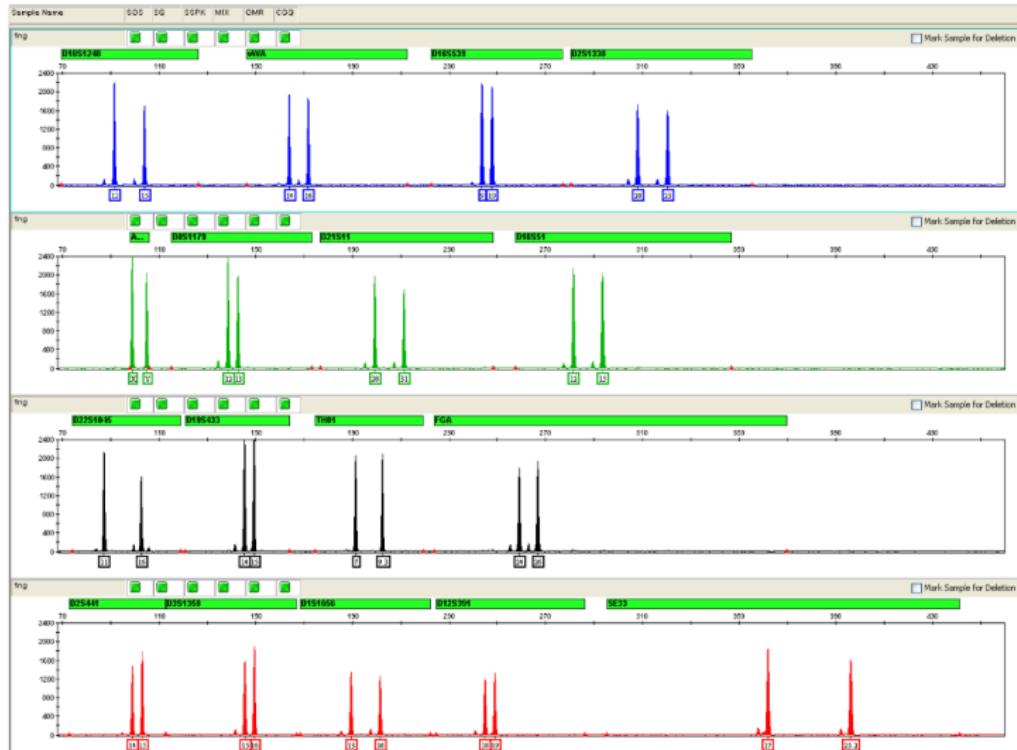
Bayesian Statistical Models

- Distribution of data (X) depends on unknown parameter θ
- Inference on θ
- Consists of a parametric statistical model(s) $f(x|\theta)$
- Prior distribution(s) of parameter θ

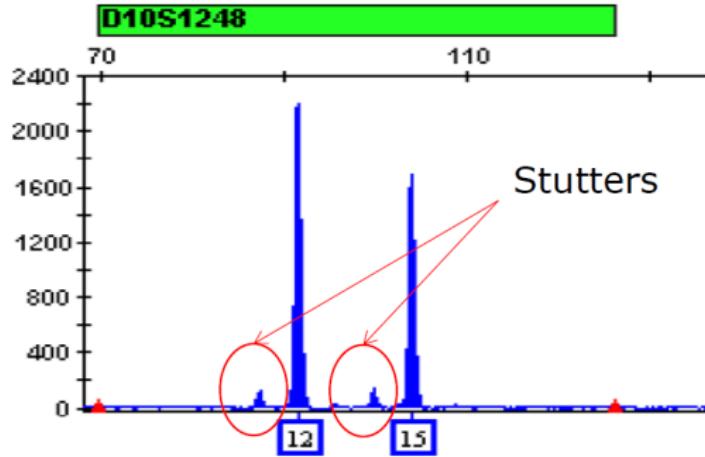
Different types of Bayesian models

- Non-hierarchical models
- Hierarchical models
- Mixture models

Electropherogram (EPG)



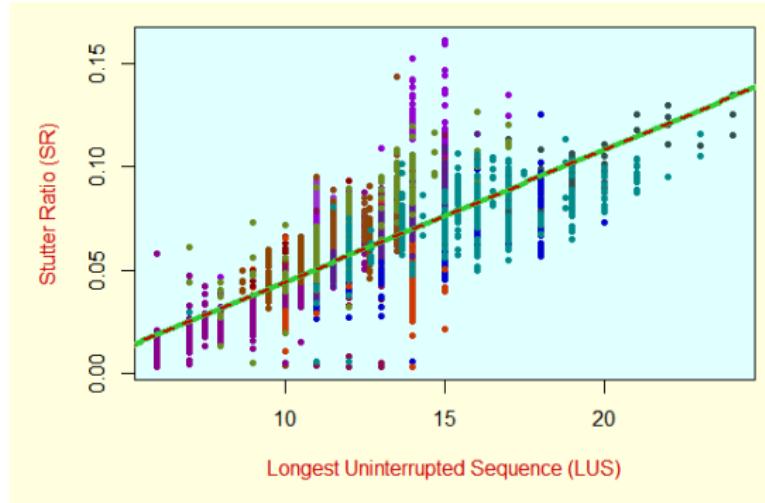
Stutters in an EPG



We are interested in modelling the stutter ratio (SR)

$$SR = \frac{\text{Observed height of the stutter peak}}{\text{The height of the parent allelic peak}}$$

Behaviour of SR



- SR is affected by the longest uninterrupted sequence of the allele, LUS
- SR is more variable for smaller values of observed allele height

Mean and Variance of Stutter Ratio

- Mean stutter ratio

$$\mu_{li} = E(SR_{li}) = \beta_{0l} + \beta_{1l} LUS_{li}$$

- Variance of stutter ratio is inversely proportional to the allele height
 - A common variance for all the loci - models with profile wide variances

$$\sigma_i^2 = \frac{\sigma^2}{O_{ai}}$$

- Locus specific variance

$$\sigma_{li}^2 = \frac{\sigma_l^2}{O_{ali}}$$

Models proposed by Bright et al.(2013)

Model	Distribution	Mean	Variance
LN ₀	$\ln(SR_{li}) \sim N(\mu_{li}, \sigma^2_i)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma^2_i = \frac{\sigma^2}{O_{ai}}$
LN ₁	$\ln(SR_{li}) \sim N(\mu_{li}, \sigma^2_{li})$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma^2_{li} = \frac{\sigma^2_l}{O_{ali}}$
G ₀	$SR_{li} \sim Gamma(\alpha_{li}, \theta_{li})$	$\mu_{li} = e^{(\beta_{0li} + \beta_{1li} LUS_{li})}$	$\sigma^2_i = \frac{\sigma^2}{O_{ai}}$
G ₁	$SR_{li} \sim Gamma(\alpha_{li}, \theta_{li})$	$\mu_{li} = e^{(\beta_{0li} + \beta_{1li} LUS_{li})}$	$\sigma^2_{li} = \frac{\sigma^2_l}{O_{ali}}$
MLN ₁	$\ln(SR_{li}) \sim \pi N(\mu_{li}, \sigma^2_{0li}) + (1 - \pi) N(\mu_{li}, \sigma^2_{1li})$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma^2_{0li} = \frac{\sigma^2_{0l}}{O_{ali}}$ $\sigma^2_{1li} = \frac{\sigma^2_{0l} + \sigma^2_{1l}}{O_{ali}}$

Description of the Proposed Models

Model	Distribution	Mean	Variance
N ₀	$SR_{li} \sim N(\mu_{li}, \sigma_i^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_i^2 = \frac{\sigma^2}{O_{aj}}$
N ₁	$SR_{li} \sim N(\mu_{li}, \sigma_{li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma^2}{O_{ali}}$
T ₀	$SR_{li} \sim t(\mu_{li}, \sigma_i^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_i^2 = \frac{\sigma^2}{O_{aj}}$
T ₀	$SR_{li} \sim t(\mu_{li}, \sigma_{li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{li}^2 = \frac{\sigma^2}{O_{ali}}$
MN ₁	$SR_{li} \sim \pi N(\mu_{li}, \sigma_{0li}^2) + (1 - \pi) N(\mu_{li}, \sigma_{1li}^2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma^2}{O_{aj}}$ $\sigma_{1li}^2 = \frac{\sigma_{0li}^2 + \sigma_{1li}^2}{O_{ali}}$
MT ₁	$SR_{li} \sim \pi t(\mu_{li}, \sigma_{0li}^2, \nu_1) + (1 - \pi) t(\mu_{li}, \sigma_{1li}^2, \nu_2)$	$\mu_{li} = \beta_{0li} + \beta_{1li} LUS_{li}$	$\sigma_{0li}^2 = \frac{\sigma^2}{O_{aj}}$ $\sigma_{1li}^2 = \frac{\sigma_{0li}^2 + \sigma_{1li}^2}{O_{ali}}$

Information Criteria

- Akaike information criterion (AIC)

$$\text{AIC} = -2 \log p(\mathbf{y} | \hat{\theta}_{mle}) + 2k$$

- Bayesian information criterion (BIC)

$$\text{BIC} = -2 \log p(\mathbf{y} | \hat{\theta}_{mle}) + k \log n$$

- Deviance information criterion (DIC)

$$\text{DIC} = -2 \log p(\mathbf{y} | \hat{\theta}_{Bayes}) - p_{\text{DIC}}$$

$$\widehat{p}_{\text{DIC}} = 2 \left[\log p(\mathbf{y} | \hat{\theta}_{Bayes}) - \frac{1}{S} \sum_{s=1}^S \log p(\mathbf{y} | \theta_s) \right]$$

$$p_{\text{DIC alt}} = 2 \text{Var}_{\text{post}} [\log p(\mathbf{y} | \theta)]$$

Widely Available or Watanabe-Akaike Information Criterion (WAIC)

- Log point wise predictive density (**lppd**)

$$\text{lppd} = \sum_{i=1}^n \log p_{\text{post}}(y_i) = \sum_{i=1}^n \log \int_{\theta} p(y_i|\theta) p_{\text{post}}(\theta) d\theta$$

- Estimated (computed) log point wise predictive density (**clppd**)

$$\text{clppd} = \sum_{i=1}^n \log \left[\frac{1}{S} \sum_{s=1}^S p(y_i|\theta_s) \right] \quad \text{WAIC} = -2 (\text{clppd} - p_{\text{WAIC}})$$

$$\hat{p}_{\text{WAIC}} = 2 \sum_{i=1}^n \left\{ \log \left[\frac{1}{S} \sum_{s=1}^S p(y_i|\theta_s) \right] - \frac{1}{S} \sum_{s=1}^S \log p(y_i|\theta) \right\}$$

$$p_{\text{WAIC alt}} = \sum_{i=1}^n \text{Var}_{\text{post}} [\log p(y_i|\theta)]$$

Limitations of Information Criteria

- AIC & BIC
 - MLE
 - Cannot use with hierarchical models
 - Not recommended for singular models
- DIC
 - Cannot use with mixture models (posterior estimates of means are quite delicate)
- WAIC
 - Valid if $\text{Var}_{\text{post}}[\log p(y_i|\theta)] \leq 0.4$
 - If $\text{Var}_{\text{post}}[\log p(y_i|\theta)] > 0.4$
Then use leave-one-out cross-validation (LOO-CV)

LOO-CV and LOO-CV Approximations

- Leave-one-out cross-validated lppd

$$\text{lppd}_{\text{loo-cv}} = \sum_{i=1}^n \log p_{\text{post}(-i)}(y_i) \quad \text{lppd}_{\text{loo-cv}} = \sum_{i=1}^n \log \left[\frac{1}{S} \sum_{s=1}^S p(y_i | \theta_{is}) \right]$$

- Importance sampling LOO-CV (IS-LOO)

$$r_{is} = \frac{1}{p(y_i | \theta_{is})} \propto \frac{p(\theta_{is} | y_{-i})}{p(\theta_{is} | y)}$$

- Truncated importance sampling LOO-CV (TIS-LOO)

$$w_{is} = \min(r_{is}, \sqrt{S} \bar{r}_i)$$

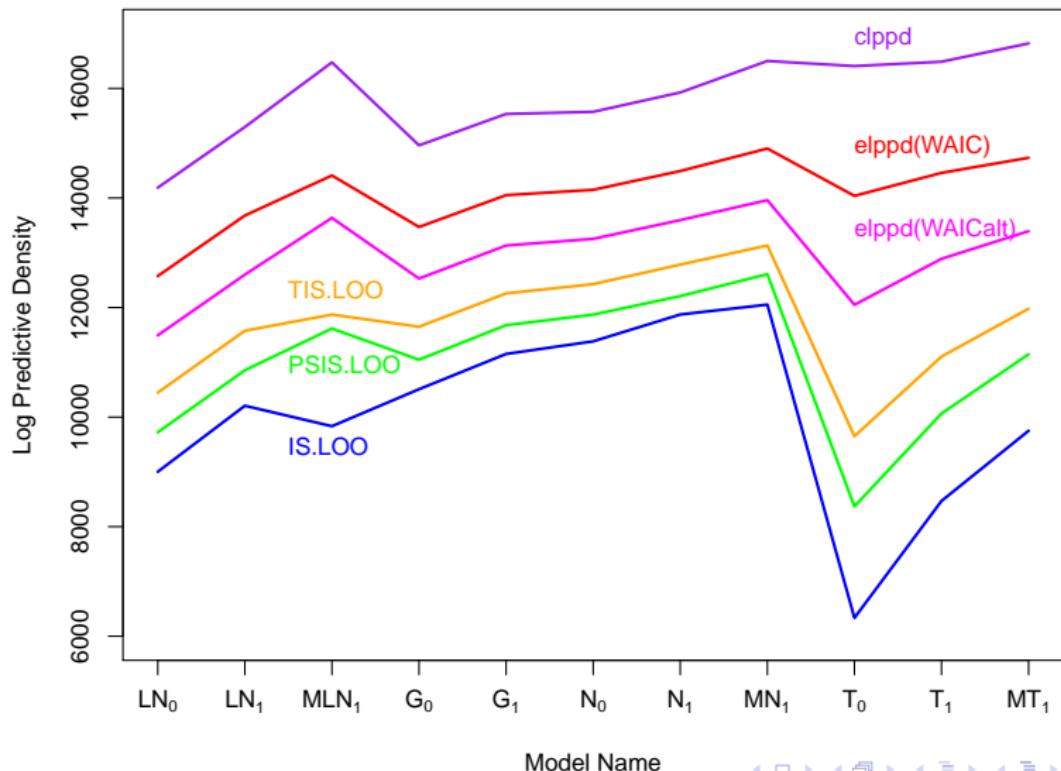
- Pareto-smoothed importance sampling LOO-CV (PSIS-LOO)

$$\tilde{w}_{is} = \min(m_{is}, S^{\frac{3}{4}} \bar{m}_i)$$

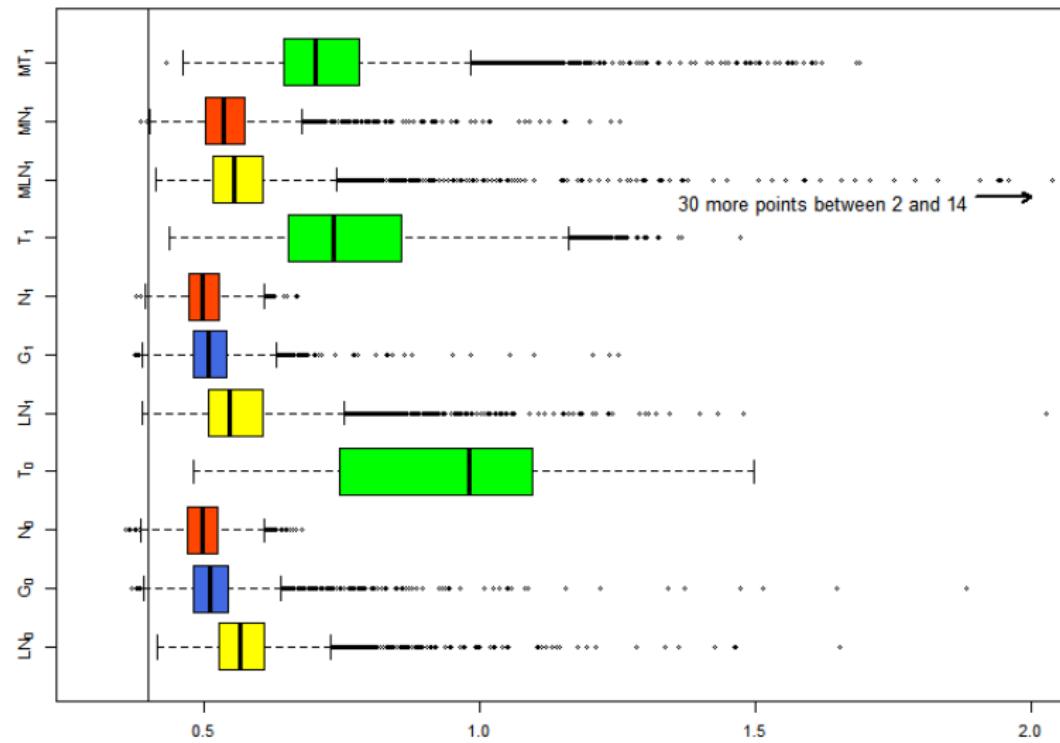
Log Predictive Densities (NGM SElectTM Data)

Model	IS	TIS	PSIS	clppd	WAIC	WAICalt
LN ₀	9002	10447	9722	14188	12575	11492
LN ₁	10207	11576	10855	15295	13680	12600
MLN ₁	9834	11871	11617	16475	14411	13636
G ₀	10509	11649	11048	14961	13470	12529
G ₁	11153	12259	11677	15533	14052	13132
N ₀	11383	12427	11870	15573	14149	13253
N ₁	11873	12784	12208	15926	14492	13596
MN ₁	12053	13133	12609	16501	14903	13961
T ₀	6334	9650	8370	16408	14038	12050
T ₁	8469	11105	10065	16488	14458	12889
MT ₁	9751	11979	11147	16821	14734	13395

Log Predictive Density Profiles (NGM SElectTM Data)



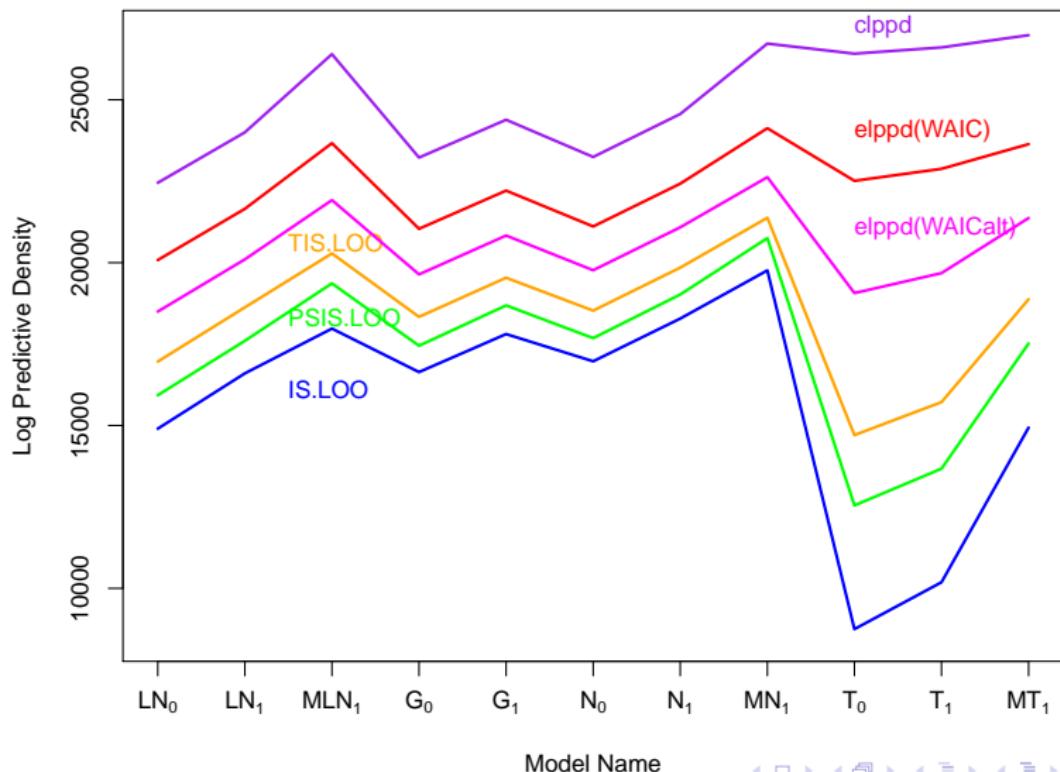
Posterior Variances of lppds (NGM SElectTM Data)



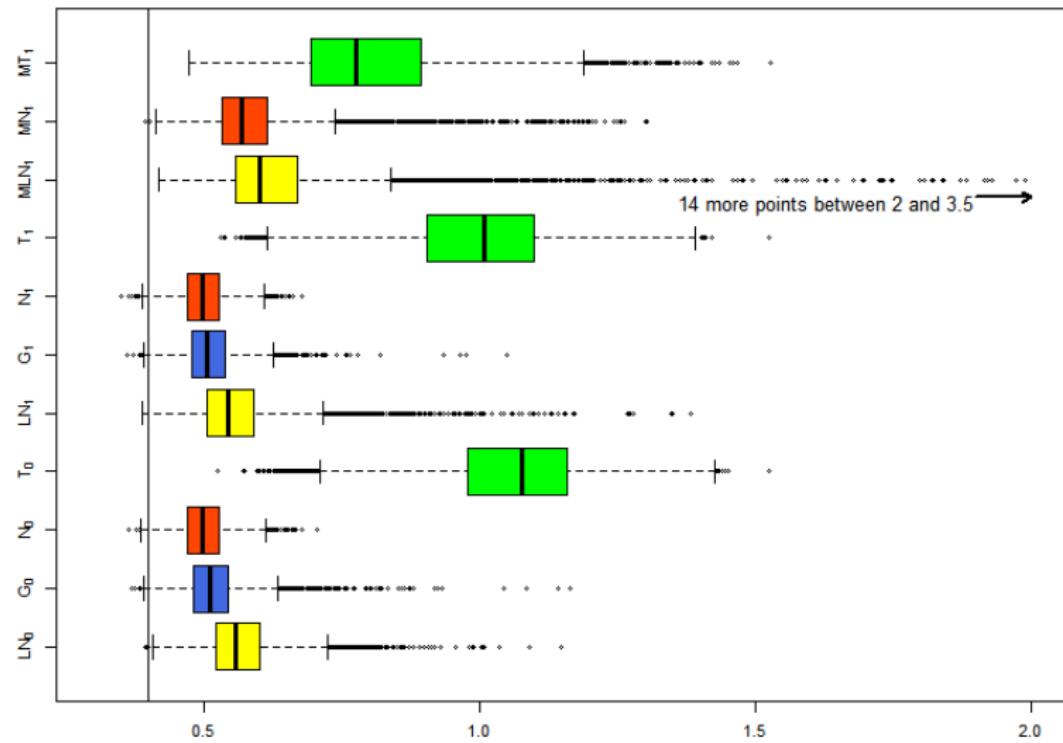
Log Predictive Densities (IdentifilerTM Data)

Model	IS	TIS	PSIS	clppd	WAIC	WAICalt
LN ₀	14902	16962	15929	22452	20079	18498
LN ₁	16601	18627	17600	24000	21652	20101
MLN ₁	17974	20283	19363	26400	23668	21920
G ₀	16640	18335	17448	23226	21038	19639
G ₁	17807	19536	18687	24386	22211	20832
N ₀	16971	18527	17682	23244	21111	19767
N ₁	18283	19846	19022	24563	22425	21085
MN ₁	19761	21378	20753	26725	24127	22623
T ₀	8749	14706	12547	26416	22512	19065
T ₁	10187	15718	13675	26607	22880	19679
MT ₁	14935	18876	17518	26984	23638	21369

Log Predictive Density Profiles (Identifier™ Data)



Posterior Variances of lppds (Identifier™ Data)



Thank You!