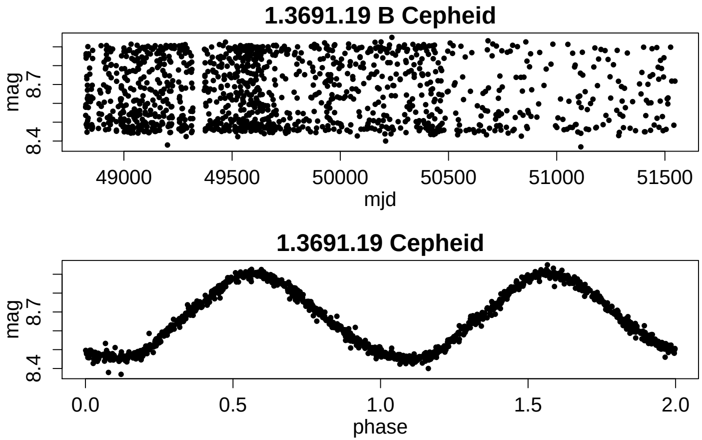
Designing Test Information and Test Information in Design

David Jones
Joint work with Xiao-Li Meng

Harvard University

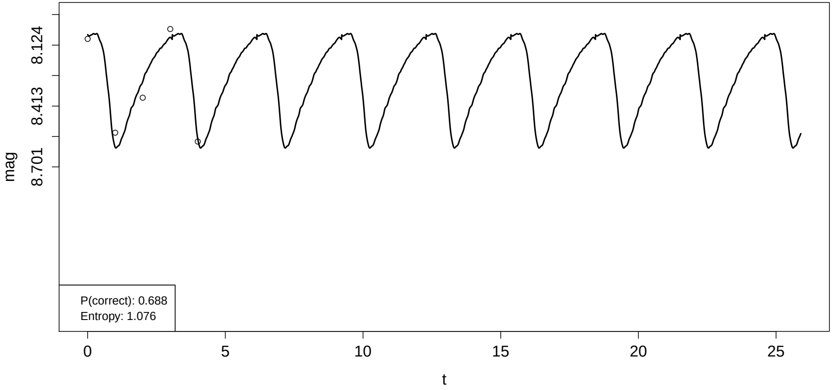
October 13, 2015

Light curve classification (earlier Dan Cervone's project)

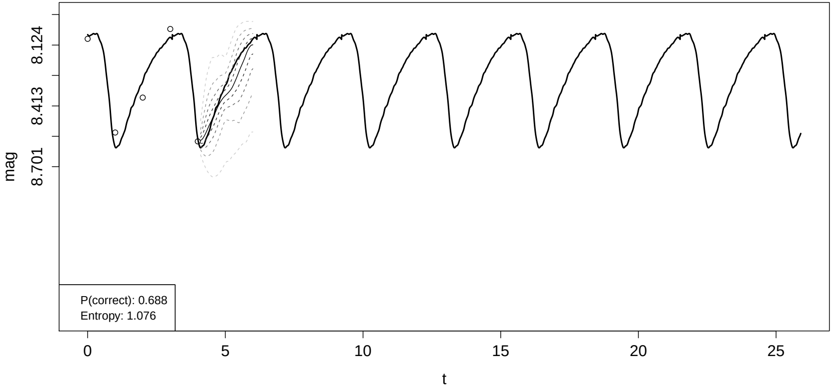


- Data from the MACHO light curve catalog
- Nine types of sources
- All light curves are assumed to follow a Gaussian Process
- The priors for the Gaussian Process parameters are class specific

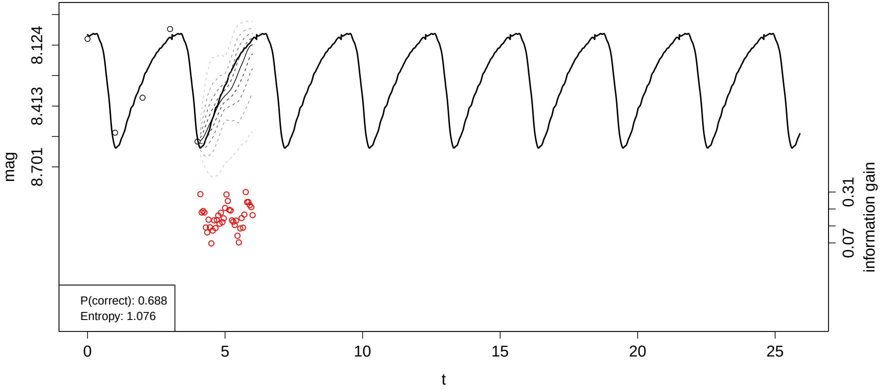
Light curve classification



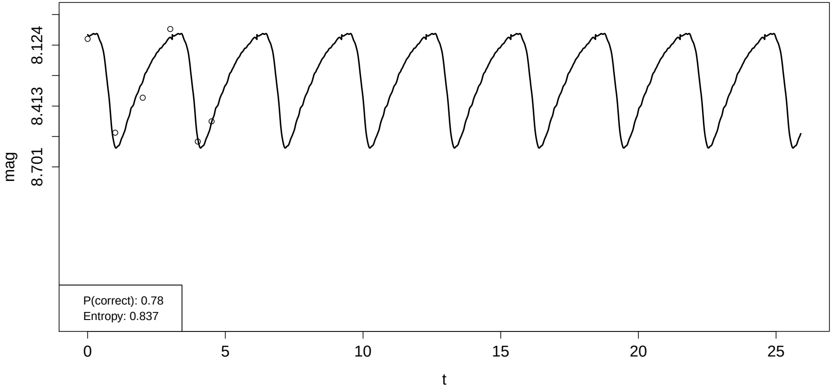
Light curve classification



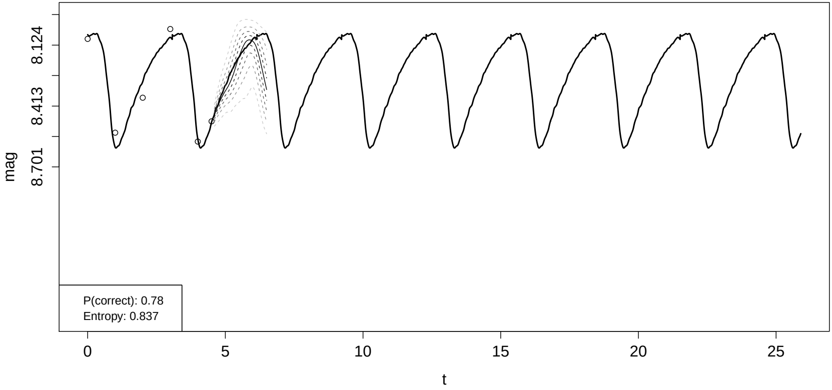
Light curve classification



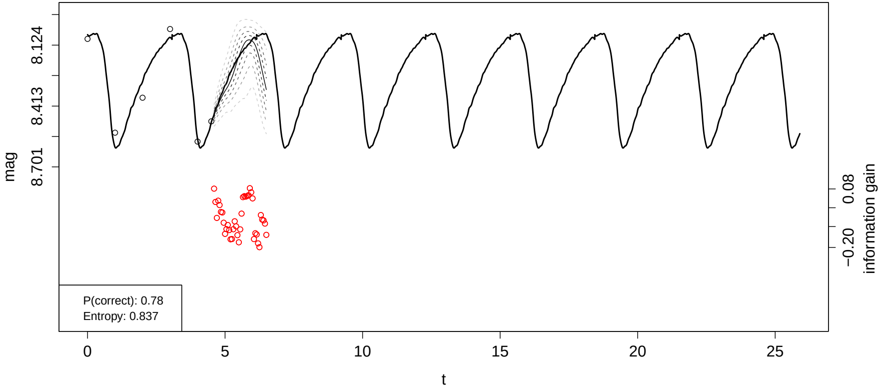
Light curve classification



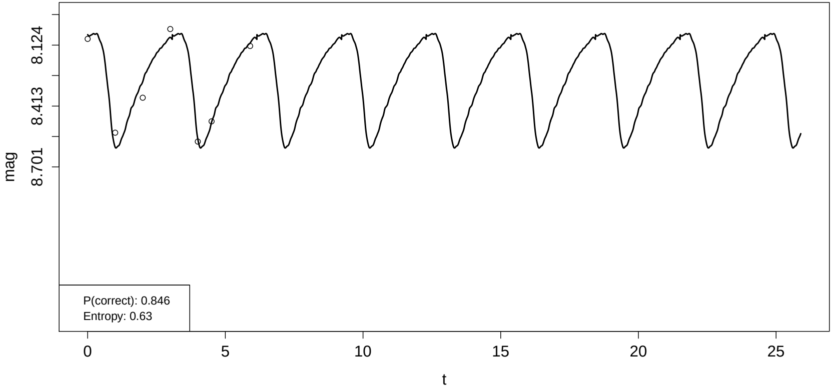
Light curve classification



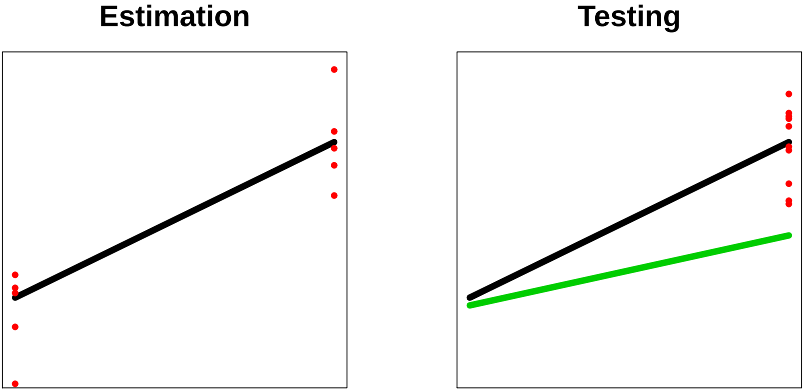
Light curve classification



Light curve classification



One slide version



Previous work?

- Nicolae et al. (2008): proposed some very natural measures e.g. $KL(f(\cdot|\theta_1)||f(\cdot|\theta_0))$
- Toman (1996): careful choice of loss function gives agreement of Bayes risk with estimation information

Estimation information review

- Shannon (1948) defined entropy: $H(\pi) = E_\theta[-\log \pi(\theta)]$
- Lindley (1956) defined *estimation* information provided by an experiment ξ with outcome X:

$$\begin{aligned} \mathcal{I}(\xi; \pi) &= \text{Prior entropy} - \text{Expected posterior entropy} \\ &= H(\pi) - E_X[H(p(\cdot|X))] \end{aligned}$$

- Linear regression: $\mathcal{I}(\xi; \pi)$ is essentially the D-optimality criterion

Generalization ... and our parallel version

DeGroot (1962) generalization

$$\mathcal{I}(\xi;\pi) = U(\pi) - E_X[U(p(\cdot|X))]$$

U = *uncertainty function*

Concave: $U(\lambda\pi_1 + (1-\lambda)\pi_2) \geq \lambda U(\pi_1) + (1-\lambda)U(\pi_2)$

Expected test information

Want to test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$. Define expected test info

$$\mathcal{I}^T_{\mathcal{V}}(\xi;\Theta_0,\Theta_1,\pi) = \mathcal{V}(1) - E_X[\mathcal{V}(\text{BF}(X|H_0,H_1))|H_1]$$

where $\text{BF}(X|H_0,H_1) = \frac{f(X|H_0)}{f(X|H_1)}$.

- Evidence function $\mathcal{V}$ (concave) e.g. $\mathcal{V}(z) = \log(z)$ gives $KL(f(\cdot|H_1)||f(\cdot|H_0))$
- Second term is f-divergence of Csiszár (1963), Ali and Silvey (1966)

Basic properties - non-negativity

(1) Non-negativity - use Jensen's inequality $\phi(E[Y]) > E[\phi(Y)]$

- DeGroot (1962):

$$E_X[p(\cdot|X)] = \int_{\mathcal{X}} p(\cdot|x) f(x) dx = \pi(\cdot)$$

- Testing:

$$E_X[\text{BF}(X|H_0, H_1)|H_1] = \int_{\mathcal{X}} \frac{f(x|H_0)}{f(x|H_1)} f(x|H_1) dx = 1$$

Jensen's inequality: $\mathcal{V}(1) \geq E_X[\mathcal{V}(\text{BF}(X|H_0, H_1))|H_1]$

Basic properties - additivity

(2) Additivity: for two-part experiment $\xi = (\xi_1, \xi_2)$ with outcome (X_1, X_2)

$$\underbrace{\mathcal{I}^T_{\mathcal{V}}(\xi;\pi)}_{\text{complete info.}} = \underbrace{\mathcal{I}^T_{\mathcal{V}}(\xi_1;\pi)}_{\text{experiment 1 info.}} + \underbrace{\mathcal{I}^T_{\mathcal{V}}(\xi_2|\xi_1;\pi)}_{\text{conditional info. of experiment 2}}$$

- Conditional test information

$$\mathcal{I}^T_{\mathcal{V}}(\xi_2|\xi_1;\pi) = E_{X_1}[\mathcal{V}(\mathsf{BF}(X_1))|H_1] - E_{X_1,X_2}[\mathcal{V}(\mathsf{BF}(X_1,X_2))|H_1]$$

- Additivity follows because $\mathcal{I}^T_{\mathcal{V}}(\xi;\pi) =$

$$\underbrace{\mathcal{V}(1) - \cancel{E_{X_1}[\mathcal{V}(\mathsf{BF}(X_1))|H_1]}}_{\mathcal{I}^T_{\mathcal{V}}(\xi_1;\pi)} + \underbrace{\cancel{E_{X_1}[\mathcal{V}(\mathsf{BF}(X_1))|H_1]} - E_{X_1,X_2}[\mathcal{V}(\mathsf{BF}(X_1,X_2))|H_1]}_{\mathcal{I}^T_{\mathcal{V}}(\xi_2|\xi_1;\pi)}$$

Model:

$$\begin{aligned} X|\theta, M &\sim N(M\theta, \sigma^2 I) \\ \theta &\sim N(\eta, \sigma^2 R) \end{aligned}$$

Estimation based D-optimality criterion:

Lindley (1956): $\mathcal{I}(M;\pi) = H(\pi) - E_X[H(p(\cdot|X))]$

M dependent part: $\phi_D(M) = \text{det}(M^T M + R^{-1})$
$= \text{det. of posterior precision matrix}$

Hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \sim N(\eta, \sigma^2 R)$

Expected test information: for $\mathcal{V}(z) = \log(z)$ we can calculate

$$\mathcal{I}^T_{\mathcal{V}}(\xi; \theta_0, \pi) \quad = \quad \mathbf{KL}(f(\cdot|H_1, M)||f(\cdot|\theta_0, M))$$

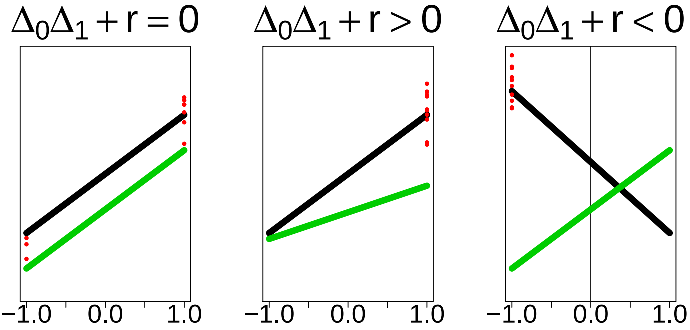
TK-optimality criterion

$$\phi_{TK}(M) = \frac{\text{Variance} \quad + \quad \text{“Bias”}}{\text{Standardize}} - \text{Penalty for relative vagueness of } H_1$$

Simple linear regression: $X_i = \theta_{\text{int}} + \theta_{\text{slope}} t_i + \epsilon_i$

Let $r = \text{Cov}(\theta_{\text{int}}, \theta_{\text{slope}} | H_1)$

$(\Delta_0, \Delta_1) = (\text{intercept diff., slope diff.}) = (\eta_{\text{int}} - \theta_{0,\text{int}}, \eta_{\text{slope}} - \theta_{0,\text{slope}})$



Probability based measures

Problems with power

1. Nuisance parameters and composite hypotheses
2. Observed power? Sequential design stopping rules
3. No maximal information interpretation
4. What if testing *and* estimation is of interest?

Probability based measures

Bayesian inspired measure:

- Posterior-prior ratio evidence function

$$\mathcal{V}(z) = \frac{z}{\pi_1 + \pi_0 z} = \frac{1}{\pi_0}\text{post. prob. of } H_0$$

- $\mathcal{I}_{\mathcal{V}}^{T}(\xi)$ = Relative expected reduction in “probability” of the null

$$1 - E_X\left[\frac{\text{BF}(X)}{\pi_1 + \pi_0\text{BF}(X)}\middle| H_1\right] = \frac{\pi_0 - E_X[\text{post. prob. of } H_0\ |H_1]}{\pi_0},$$

where $\text{BF}(X) = f(X|H_0)/f(X|H_1)$

Probability based measures

Coherence – "basic property (3)":

- "Dual" evidence function $\mathcal{V}_D(z) = \frac{1}{\pi_1+\pi_0 z}$, concave in $1/z$
- Dual measures

$$\begin{aligned} \mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1) &= 1 - E_X\left[\frac{\mathsf{BF}(X)}{\pi_1 + \pi_0 \mathsf{BF}(X)} \middle| H_1\right] \\ \mathcal{I}_{\mathcal{V}_D}^T(\xi; H_1, H_0) &= 1 - E_X\left[\frac{1}{\pi_1 + \pi_0 \mathsf{BF}(X)} \middle| H_0\right] \end{aligned}$$

Coherence identity

$$\frac{\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1)}{\mathcal{I}_{\mathcal{V}_D}^T(\xi; H_1, H_0)} = 1 \quad \text{or} \quad \mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1) = \mathcal{I}_{\mathcal{V}_D}^T(\xi; H_1, H_0) = 0$$

- **Consequence**: when finding optimal designs for testing it will not matter which hypothesis is true

Observed test information

Observed test information

$$\mathcal{I}_{\mathcal{V}}^{T}(\xi; \Theta_0, \Theta_1, \pi, x) = \mathcal{V}(1) - \mathcal{V}(\mathsf{BF}(x|H_0, H_1))$$

Observed coherence identity

$$\frac{\mathcal{V}(\mathsf{BF}(x))}{\mathcal{V}_D(\mathsf{BF}(x))} = \mathsf{BF}(x)$$

- More fundamental – Bayes factor is preserved
- Implies expected coherence identity
- Examples: posterior-prior ratio and evidence function for symmetrized KL-divergence $\frac{1}{2}KL(f(\cdot|H_1)||f(\cdot|H_0)) + \frac{1}{2}KL(f(\cdot|H_0)||f(\cdot|H_1))$ i.e.

$$\mathcal{V}(z) = \frac{1}{2}\log(z) - \frac{1}{2}z\log(z)$$

Coherence identity in sequential design

Observed conditional information

$$\mathcal{I}_{\mathcal{V}}^{T}(\xi_2|\xi_1; x_1) = \mathcal{V}(\mathsf{BF}(x_1|H_0, H_1)) - E_{X_2}[\mathcal{V}(\mathsf{BF}(x_1, X_2|H_0, H_1))|H_1, x_1]$$

Observed conditional coherence identity

$$\frac{\mathcal{I}_{\mathcal{V}}^{T}(\xi_2|\xi_1; x_1)}{\mathcal{I}_{\mathcal{V}_D}^{T}(\xi_2|\xi_1; x_1)} = \mathsf{BF}(x_1)$$

- Implied by observed coherence identity
- Optimal sequential designs do not depend on which hypothesis is true

Simulations

1. Binary regression non-nested models (link function)
2. Sequential design for cubic regression models

Sequential design example

- **Model:**

$$X|\theta, M \sim N(M\theta, I_4),$$

where $\theta = (\theta_{\text{int}}, \theta_{\text{slope}}, \theta_{\text{quad}}, \theta_{\text{cubic}})$

- **Hypotheses:**

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \sim N(\eta, R)$$

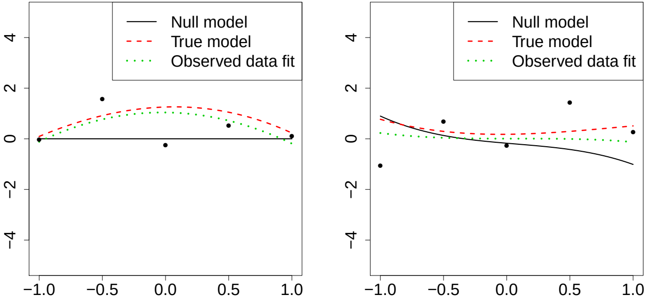
- **Observed data:** design matrix M_1 for x_1

$$M_1^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_{1,1} & t_{1,2} & \cdots & t_{1,n_1} \\ t_{1,1}^2 & t_{1,2}^2 & \cdots & t_{1,n_1}^2 \\ t_{1,1}^3 & t_{1,2}^3 & \cdots & t_{1,n_1}^3 \end{pmatrix} \tag{1}$$

Set $n_1 = 5$ and $\mathbf{t}_1 = (-1, -0.5, 0, 0.5, 1)$

- **Task:** for $n_2 = 5$ choose design M_2 for missing data

Sequential design example



Three settings ($R = 0.2I_4$):

1. Parabola: $\theta_0 = (0, 0, 0, 0)$ and $\eta = (1.1, 0, -1.3, 0)$
2. High curvature:

$$\theta_{0,\text{int}}, \theta_{0,\text{slope}} \sim \text{Uniform}(-1, 1)$$

$$\theta_{0,\text{quad}}, \theta_{0,\text{cubic}} \sim \text{Uniform}(-10, 10)$$

$$\eta = \theta_0$$

3. Standard curvature: same except $\theta_{0,\text{quad}}, \theta_{0,\text{cubic}} \sim \text{Uniform}(-1, 1)$

Sequential design example

Method: optimize three criteria

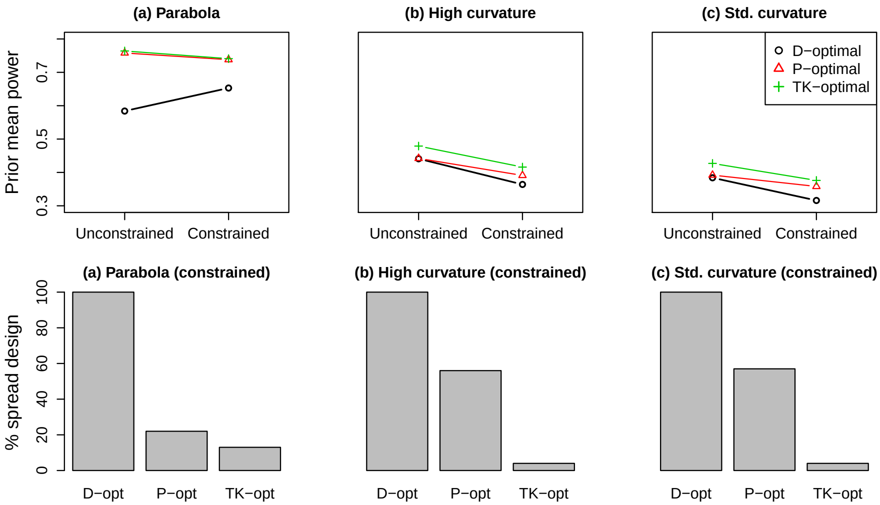
1. $\mathcal{I}_{\mathcal{V}}^{T}(\xi_2|\xi_1; x_1)$ for posterior-prior ratio evidence function
2. $\mathcal{I}_{\mathcal{V}}^{T}(\xi_2|\xi_1; x_1)$ for $\mathcal{V}(z) = \log$
3. D-optimality criterion

Evaluation: average power for fixed θ over H_1 dist. for θ

$$\int_{\Theta_1} \text{Power}(\theta, \text{procedure k})\pi(\theta|H_1)d\theta,$$

for $k = 1, 2, 3$

Sequential design example



Constrained optimization: either $\mathbf{t}_2 = \mathbf{t}_1$ or put all points near where null and posterior (for x_1) mean model differ most

Future goal: design for testing and estimation

Fraction of observed information

$$\mathcal{F}\mathcal{I}_{\mathcal{V}}^{T}(\xi_2|\xi_1;x_1) = \frac{\mathcal{I}_{\mathcal{V}}^{T}(\xi_1;x_1)}{\mathcal{I}_{\mathcal{V}}^{T}(\xi_1;x_1) + \mathcal{I}_{\mathcal{V}}^{T}(\xi_2|\xi_1;x_1)}$$

Single numerical summary of

- How much more test information may be obtainable
- How difficult it is to collect that test information

Fisher information analogue (estimation):

$$\frac{I_{\text{ob}}}{I_{\text{ob}} + I_{\text{mis}}},$$

where

$$I_{\text{ob}} = -\frac{\partial^2 \log f(x_1|\theta)}{\partial \theta^2}\bigg|_{\theta=\theta_{\text{MLE}}}, I_{\text{mis}} = E_{X_2}\left[-\frac{\partial^2 \log f(x_1, X_2|x_1, \theta)}{\partial \theta^2} \bigg| x_1, \theta \right]\bigg|_{\theta=\theta_{\text{MLE}}}$$

Future goal: design for testing and estimation

No evidence approximation

Conditions:

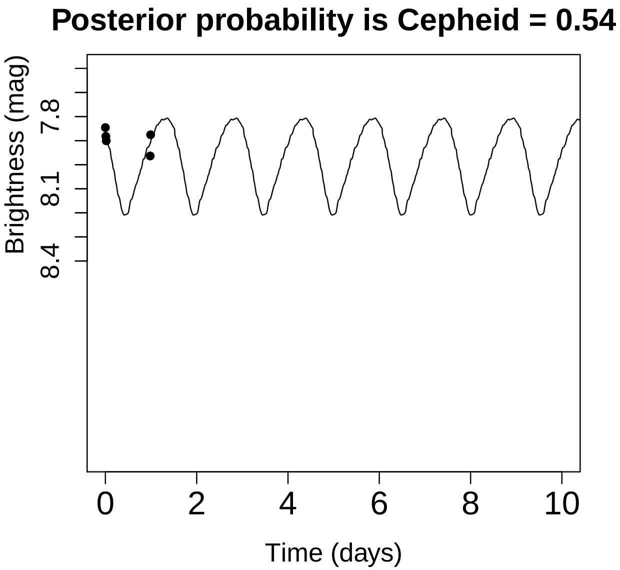
1. Precise prior: $H_1 : \theta \sim \text{Uniform}(\theta_1 - \delta, \theta_1 + \delta)$ for small δ
2. Null is approximately correct: $|\theta_0 - \theta_{\text{MLE}}|$ small
3. Prior mean better still: $|\theta_1 - \theta_{\text{MLE}}|$ smaller

Then:

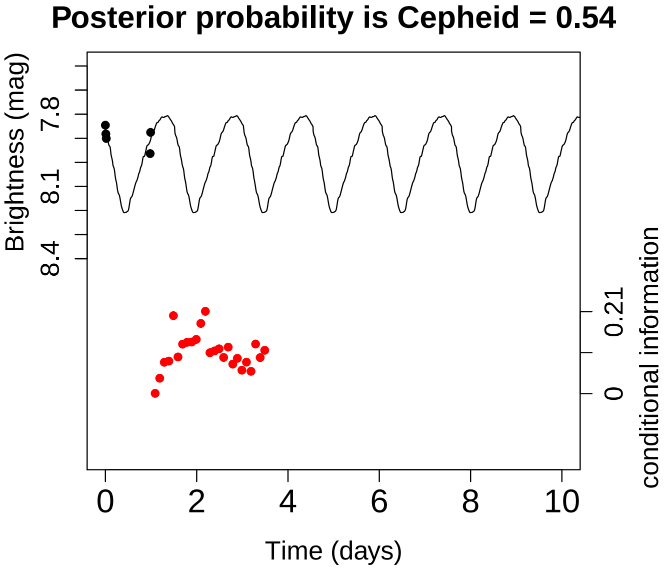
$$\mathcal{FI}_{\mathcal{V}}^{T}(\xi_2|\xi_1; x_{\text{ob}}) \approx \frac{I_{\text{ob}}}{I_{\text{ob}} + \frac{-\mathcal{V}''(1)}{\mathcal{V}'(1)} I_{\text{mis}}},$$

- Conversion number: $C_{\mathcal{V}} = \frac{-\mathcal{V}''(1)}{\mathcal{V}'(1)}$
- Characterization: LRT $C_{\mathcal{V}} = 1$, Bayesian hypothesis testing $C_{\mathcal{V}} = \infty$

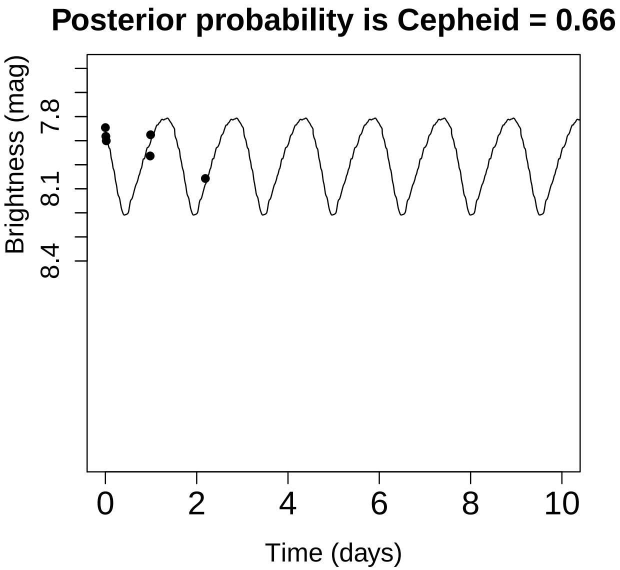
Questions regarding work on lightcurve classification



Questions regarding work on lightcurve classification



Questions regarding work on lightcurve classification



Questions regarding work on lightcurve classification

1. Taking a step back, what should the model be?
2. How should we assess the success of our optimal designs?

Lightcurve model?

Current model: Gaussian process with class specific priors

$$y_i \sim f_i + \epsilon_i$$
$$\epsilon_i \sim N(0, V_i),\ V_i \text{ known}$$
$$\mathbf{f} \sim N(\mu\mathbf{1}, K_c(\mathbf{t}, \mathbf{t}; \phi))$$

e.g. Periodic kernel: $K_c(s, t; \phi) = \sigma^2 \exp\left(-\beta \sin\left(\frac{\pi(t-s)}{\tau}\right)^2\right)$

Class C specific prior based on previously classified lightcurves:

$$\begin{pmatrix}\mu \\ \log\phi\end{pmatrix}\Bigg| C \sim N\left(\begin{pmatrix}\mu_{0,C} \\ \tilde{\phi}_{0,C}\end{pmatrix}, \Sigma_{0,C}\right)$$

Best way to assess design performance?

- Should we measure how close we get to the optimal gain in posterior probability of the correct class? (Through simulation from a precisely fit lightcurve).
- For general $\mathcal{V}$, should we still consider posterior probability?
- Which measures are more robust when there are few observations?
- We could also base the assessment on success of “the test” but it is not clear what the test should be

References I

S. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitiit von markoffschen ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, 8: 85–108, 1963.

M. H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, pages 404–419, 1962.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

D. L. Nicolae, X.-L. Meng, and A. Kong. Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies. *Statistical Science*, 23(3):pp. 287–312, 2008. ISSN 08834237. URL `http://www.jstor.org/stable/20697638`.

References II

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

B. Toman. Bayesian experimental design for multiple hypothesis testing. *Journal of the American Statistical Association*, 91(433):185–190, 1996.