Designing Test Information and Test Information in Design

David Jones
Joint work with Xiao-Li Meng
Harvard University

October 13, 2015
The priors for the Gaussian Process parameters are class specific.

- 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

P(correct): 0.688
Entropy: 1.076
Light curve classification

P(correct): 0.688
Entropy: 1.076

Information gain: 0.31
Light curve classification

P(correct): 0.78
Entropy: 0.837
Light curve classification

\[ P(\text{correct}) : 0.78 \]
\[ \text{Entropy} : 0.837 \]
Light curve classification

P(correct): 0.78
Entropy: 0.837

Information gain

\[ \text{Information gain} = - \sum_{i} P(i) \log P(i) \]

\( P(i) = \frac{\text{count}_{i}}{\text{total count}} \)
Light curve classification

![Light curve graph]

- P(correct): 0.846
- Entropy: 0.63
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
Shannon (1948) defined entropy: \( H(\pi) = E_\theta[-\log \pi(\theta)] \)

Lindley (1956) defined estimation information provided by an experiment \( \xi \) with outcome \( X \):

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

Linear regression: \( \mathcal{I}(\xi; \pi) \) is essentially the D-optimality criterion
DeGroot (1962) generalization

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

\( U = \text{uncertainty function} \)

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

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

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

- Evidence function \( V \) (concave) e.g. \( V(z) = \log(z) \) 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):

$$E_X[p(\cdot|X)] = \int_X p(\cdot|x) f(x) dx = \pi(\cdot)$$

- Testing:

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

Jensen’s inequality: $\mathcal{V}(1) \geq E_X[\mathcal{V}(BF(X|H_0, H_1))|H_1]$
(2) Additivity: for two-part experiment $\xi = (\xi_1, \xi_2)$ with outcome $(X_1, X_2)$

$$\mathcal{I}_T^T(\xi; \pi) = \mathcal{I}_T^T(\xi_1; \pi) + \mathcal{I}_T^T(\xi_2|\xi_1; \pi)$$

- Complete info.  
- Experiment 1 info.  
- Conditional info. of experiment 2

**Conditional test information**

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

**Additivity follows because** $\mathcal{I}_T^T(\xi; \pi) =$

$$\mathcal{V}(1) - E_{X_1}[\mathcal{V}(\operatorname{BF}(X_1))|H_1] + E_{X_1}[\mathcal{V}(\operatorname{BF}(X_1))|H_1] - E_{X_1, X_2}[\mathcal{V}(\operatorname{BF}(X_1, X_2))|H_1]$$

$$\mathcal{I}_T^T(\xi_1; \pi) + \mathcal{I}_T^T(\xi_2|\xi_1; \pi)$$
Canonical example: Bayesian linear regression

Estimation

Model:

\[ X \mid \theta, M \sim N(M\theta, \sigma^2 I) \]
\[ \theta \sim N(\eta, \sigma^2 R) \]

Estimation based D-optimality criterion:

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

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

$$\mathcal{I}_V^T(\xi; \theta_0, \pi) = KL(f(\cdot | H_1, M) || f(\cdot | \theta_0, M))$$

**TK-optimality criterion**

$$\phi_{TK}(M) = \frac{\text{Variance}}{\text{Standardize}} + \text{“Bias”} - \text{Penalty for relative vagueness of } H_1$$
Canonical example: Bayesian linear regression

Sense check

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

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

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

Delta plots:

- $\Delta_0\Delta_1 + r = 0$
- $\Delta_0\Delta_1 + r > 0$
- $\Delta_0\Delta_1 + r < 0$
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?
Bayesian inspired measure:

- Posterior-prior ratio evidence function

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

- \( I_V^T(\xi) = \) Relative expected reduction in “probability” of the null

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

where \( BF(X) = f(X|H_0)/f(X|H_1) \)
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

$$
\mathcal{I}_V^T(\xi; H_0, H_1) = 1 - EX \left[ \frac{BF(X)}{\pi_1 + \pi_0 BF(X)} \middle| H_1 \right]
$$
$$
\mathcal{I}_V^T(\xi; H_1, H_0) = 1 - EX \left[ \frac{1}{\pi_1 + \pi_0 BF(X)} \middle| H_0 \right]
$$

Coherence identity

$$
\frac{\mathcal{I}_V^T(\xi; H_0, H_1)}{\mathcal{I}_V^T_{D}(\xi; H_1, H_0)} = 1 \quad \text{or} \quad \mathcal{I}_V^T(\xi; H_0, H_1) = \mathcal{I}_V^T_{D}(\xi; H_1, H_0) = 0
$$

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

\[ \mathcal{I}_T^T(\xi; \Theta_0, \Theta_1, \pi, x) = \mathcal{V}(1) - \mathcal{V}(\text{BF}(x|H_0, H_1)) \]

Observed coherence identity

\[ \frac{\mathcal{V}(\text{BF}(x))}{\mathcal{V}_D(\text{BF}(x))} = \text{BF}(x) \]

- 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

\[ \mathcal{I}_T^T(\xi_2|\xi_1; x_1) = \mathcal{V}(\text{BF}(x_1|H_0, H_1)) - \mathbb{E}_{X_2}[\mathcal{V}(\text{BF}(x_1, X_2|H_0, H_1))|H_1, x_1] \]

Observed conditional coherence identity

\[ \frac{\mathcal{I}_T^T(\xi_2|\xi_1; x_1)}{\mathcal{I}_V^T(\xi_2|\xi_1; x_1)} = \text{BF}(x_1) \]

- 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 \( x_1 \)

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

  \[ (1) \]

  Set \( n_1 = 5 \) and \( t_1 = (-1, -0.5, 0, 0.5, 1) \)

- **Task:** for \( n_2 = 5 \) choose design \( M_2 \) for missing data
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. \( \mathcal{I}_V^T(\xi_2|\xi_1; x_1) \) for posterior-prior ratio evidence function
2. \( \mathcal{I}_V^T(\xi_2|\xi_1; x_1) \) for \( V(z) = \log \)
3. D-optimality criterion

Evaluation: average power for fixed \( \theta \) over \( H_1 \) dist. for \( \theta \)

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

for \( k = 1, 2, 3 \)
Sequential design example

Constrained optimization: either $t_2 = t_1$ 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

\[
F \mathcal{I}_V^T(\xi_2|\xi_1; x_1) = \frac{\mathcal{I}_V^T(\xi_1; x_1)}{\mathcal{I}_V^T(\xi_1; x_1) + \mathcal{I}_V^T(\xi_2|\xi_1; x_1)}
\]

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):

\[
\frac{I_{ob}}{I_{ob} + I_{mis}},
\]

where

\[
I_{ob} = - \left. \frac{\partial^2 \log f(x_1|\theta)}{\partial \theta^2} \right|_{\theta=\theta_{MLE}},
I_{mis} = \mathbb{E}_{X_2} \left[ - \left. \frac{\partial^2 \log f(x_1, X_2|x_1, \theta)}{\partial \theta^2} \right|_{x_1, \theta} \right]_{\theta=\theta_{MLE}}
\]
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 $\delta$
2. Null is approximately correct: $|\theta_0 - \theta_{\text{MLE}}|$ small
3. Prior mean better still: $|\theta_1 - \theta_{\text{MLE}}|$ smaller

Then:

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

- Conversion number: $C_V = \frac{-V''(1)}{V'(1)}$
- Characterization: LRT $C_V = 1$, Bayesian hypothesis testing $C_V = \infty$
Questions regarding work on lightcurve classification

Posterior probability is Cepheid = 0.54
Questions regarding work on lightcurve classification

Posterior probability is Cepheid = 0.54
Questions regarding work on lightcurve classification

Posterior probability is Cepheid = 0.66
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?
Current model: Gaussian process with class specific priors

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

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:

\[
\left( \frac{\mu}{\log \phi} \right) \mid C \sim N \left( \left( \frac{\mu_0, C}{\tilde{\phi}_0, C} \right), \Sigma_{0,C} \right)
\]
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.


