# Bayesian Experimental Designs

Quan Long*, **Marco Scavino**† *, Raúl Tempone*, Suojin Wang††

* *SRI Center for Uncertainty Quantification*
*Computer, Electrical and Mathematical Sciences & Engineering Division,*
*King Abdullah University of Science and Technology, KSA*
† *IESTA, Universidad de la República, Montevideo, Uruguay*
†† *Department of Statistics, Texas A&M University, USA*


KAUST, AMCS Graduate Seminar
October 23, 2014 – Thuwal, Kingdom of Saudi Arabia

# Outline

# Introduction - a motivating example I

> *In designing an experiment, decisions must be made* **before** *data collection, and data collection is restricted by limited resources (K. Chaloner & I. Verdinelli [1], p.273).*

NASA News (July 23$^{rd}$, 2014)
`http://www.nasa.gov/jpl/spitzer/pia18463/`
Using data from NASA's Kepler and Spitzer Space Telescopes, scientists have made the most precise measurement ever of the size of a world outside our solar system, as illustrated in this artist's conception.

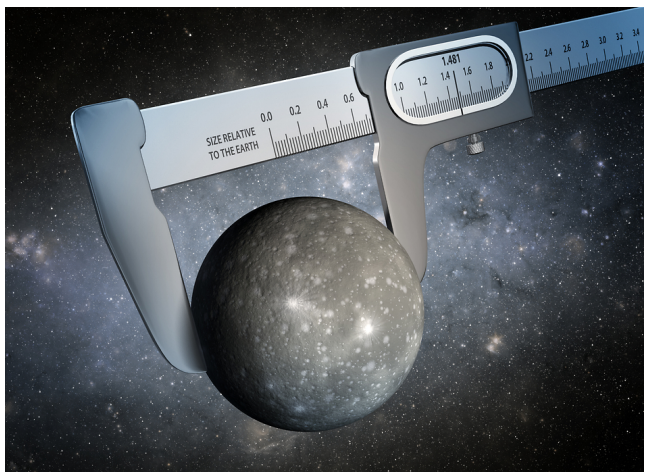



Figure : NASA/JPL-Caltech - Gauging an Alien World's Size.

# Introduction - a motivating example II

The diameter of the exoplanet, dubbed Kepler-93b, is now known with an uncertainty of just one percent.

According to this new study, the diameter of Kepler-93b is about 18,800 km, $\pm$ 240 km – the approximate distance between Washington, D.C., and Philadelphia, Penn.

Kepler-93b is 1.481 times the width of Earth, the diameter of which is 12,742 km.

> *Although light from the planet is too faint to be detected, the gravitational tug of the planet on the star is sufficient to produce a measurable Doppler shift in the velocity of absorption lines in the stars spectrum. By fitting a Keplerian orbit to the measured radial velocity data, say $y_k$, it is possible to obtain information about the orbit and a lower limit on the mass of the unseen planet (P. Gregory [2], p.331).*

# Introduction - a motivating example III

A basic tool in this context is the
**Keplerian radial velocity model for a single planet.**

**The extrasolar planet Kepler problem** ([3], p.60))

> *Making radial velocity measurements of a star in order to best determine the parameters of the orbit of an unseen Jupiter-mass companion.*

**Goal** ([3], pp.60-61))

> *To choose* **future times of observations** *to best improve the estimates of the planet's orbital parameters.*

# Introduction - a motivating example IV

The *predicted* time-dependent Keplerian star's radial velocity for a single planet takes the form

$$v(t) = v_0 + K\,[e\,cos(\omega_P) + cos(\omega_P + \nu(t))]$$

where

- $v_0$ is the systemic velocity,
- $K$ is the radial velocity semi-amplitude (m/s),
- $e$ is the eccentricity of the elliptical orbit and
- $\omega_P$ is the argument of periastron (the point nearest to a star in the path of a planet orbiting that star).

The so-called *true anomaly* $\nu(t)$ (the angle between the periastron and the position of the planet), that depends on the three parameters $e$,

- $\tau$ the orbital period (days), and
- $t_P$ the time of periastron passage,

# Introduction - a motivating example V

can be computed by solving the two equations

$$E(t) - e\sin(E(t)) = \frac{2\pi}{\tau}(t - t_P)\,,$$

$$\tan\left(\frac{\nu(t)}{2}\right) = \sqrt{\frac{1+e}{1-e}}\tan\left(\frac{E(t)}{2}\right)\,.$$

Statistical assumptions:

- the single measurement $y_k$ is Gaussian distributed with mean $v(t_k;\tau,e,K)$,
- the measurement errors are uncorrelated, i.e. the error covariance matrix is given by $\mathbf{\Sigma}_\epsilon = \boldsymbol{I}\sigma^2$, with standard deviation $\sigma = 8\,\mathrm{m\ s^{-1}}$.

# Introduction - a motivating example VI

The *measured* velocities are given by

$$y_k = v(t_k; \tau, e, K) + e_k\,, \qquad e_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)\,. \tag{1}$$

We now describe the main stages that characterize a simulation-based optimal Bayesian experimental design.

To exemplify the virtuous cycle that may lead to the uncertainty reduction about the unknown parameters assume, as an initial *observation stage*, that 10 observations are available from model (1) with $\tau = 800$ days, $e = 0.5$, and $K = 50\,\text{m s}^{-1}$.

# Introduction - a motivating example VII

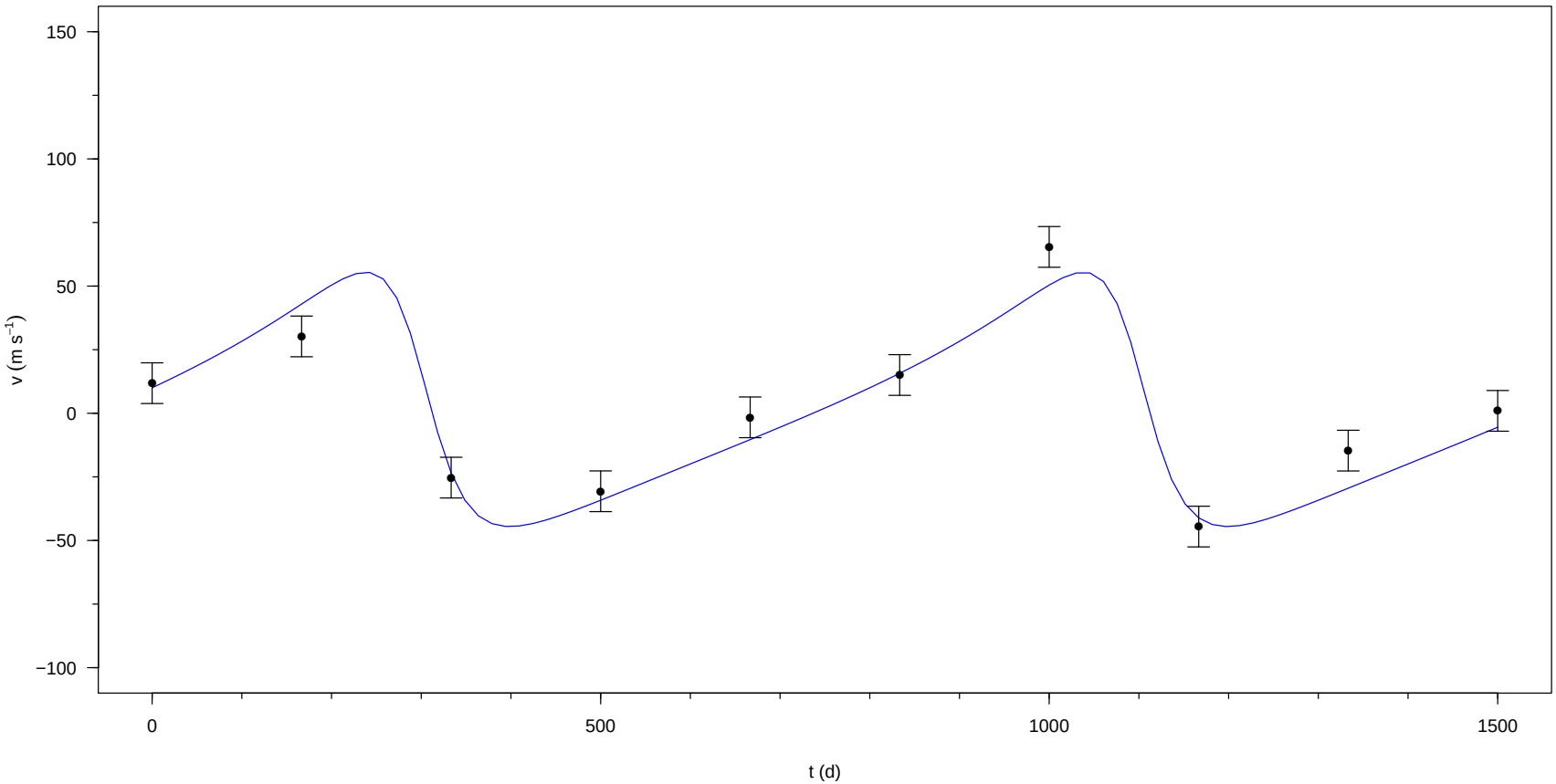



Figure : The true velocity curve and the 10 simulated observations.

# Introduction - a motivating example VIII

We now introduce - the *inference stage* - the *posterior probability density* for the unknown *parameter vector* $\boldsymbol{\theta} := (\tau, e, K)$ by means of the Bayes' theorem:

$$p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \mathcal{M}) = \frac{p(\boldsymbol{\theta}|\mathcal{M})\, p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}, \mathcal{M})}{p(\bar{\boldsymbol{y}}|\mathcal{M})}$$

where $\bar{\boldsymbol{y}}$ and $\mathcal{M}$ denote the *data vector* and the *modeling assumptions*, respectively, and

- $p(\boldsymbol{\theta}|\mathcal{M})$ is the *prior probability density* for the parameter vector $\boldsymbol{\theta} \in \Theta$,
- $p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}, \mathcal{M})$ is the *likelihood function*, and
- $p(\bar{\boldsymbol{y}}|\mathcal{M}) = \int p(\boldsymbol{\theta}|\mathcal{M})\, p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}, \mathcal{M})\, d\boldsymbol{\theta}$ is the *evidence* of the data.

Assumption: flat prior.
Then

$$p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \mathcal{M}) \propto \exp\left[-\frac{1}{2}\sum_{k=1}^{10}\left(\frac{y_k - v(t_k; \tau, e, K)}{\sigma}\right)^2\right].$$

We may now draw samples from the posterior distribution $p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \mathcal{M})$.

# Introduction - a motivating example IX

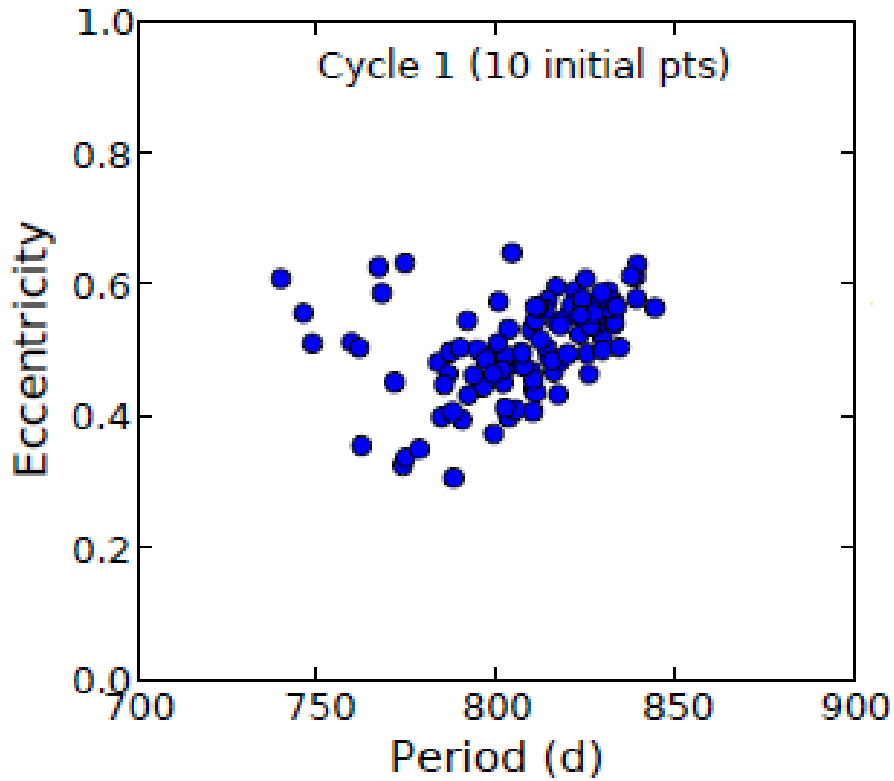



Figure : Samples from the posterior distribution $p(\tau, e|\bar{\boldsymbol{y}}, \mathcal{M})$. T. J. Loredo ([4], p.341)

# Introduction - a motivating example X

Finally, in the *design stage*, our goal is to choose the time $t$ at which to take a new observation in order to reduce the uncertainty into the unknown parameters.

Remark

*In this example the time $t$ is the design parameter.*

First we consider the *predictive distribution* of a future data $y$ at time $t$:

$$p(y|t,\bar{\boldsymbol{y}},\mathcal{M}) = \int p(y,\boldsymbol{\theta}|t,\bar{\boldsymbol{y}},\mathcal{M})d\boldsymbol{\theta} = \int p(y|t,\boldsymbol{\theta},\mathcal{M})\, p(\boldsymbol{\theta}|\bar{\boldsymbol{y}},\mathcal{M})d\boldsymbol{\theta}\,,$$

where $p(y|t,\boldsymbol{\theta},\mathcal{M})$ is the sampling distribution for $y$.

$$\begin{aligned} p(y|t,\bar{\boldsymbol{y}},\mathcal{M}) &= \int p(\boldsymbol{\theta}|\bar{\boldsymbol{y}},\mathcal{M})\,\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-v(t;\boldsymbol{\theta}))^2\right]d\boldsymbol{\theta} \\ &\approx \frac{1}{N}\sum_{\boldsymbol{\theta}_j}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-v(t;\boldsymbol{\theta}_j))^2\right]\,. \end{aligned}$$

# Introduction - a motivating example XI

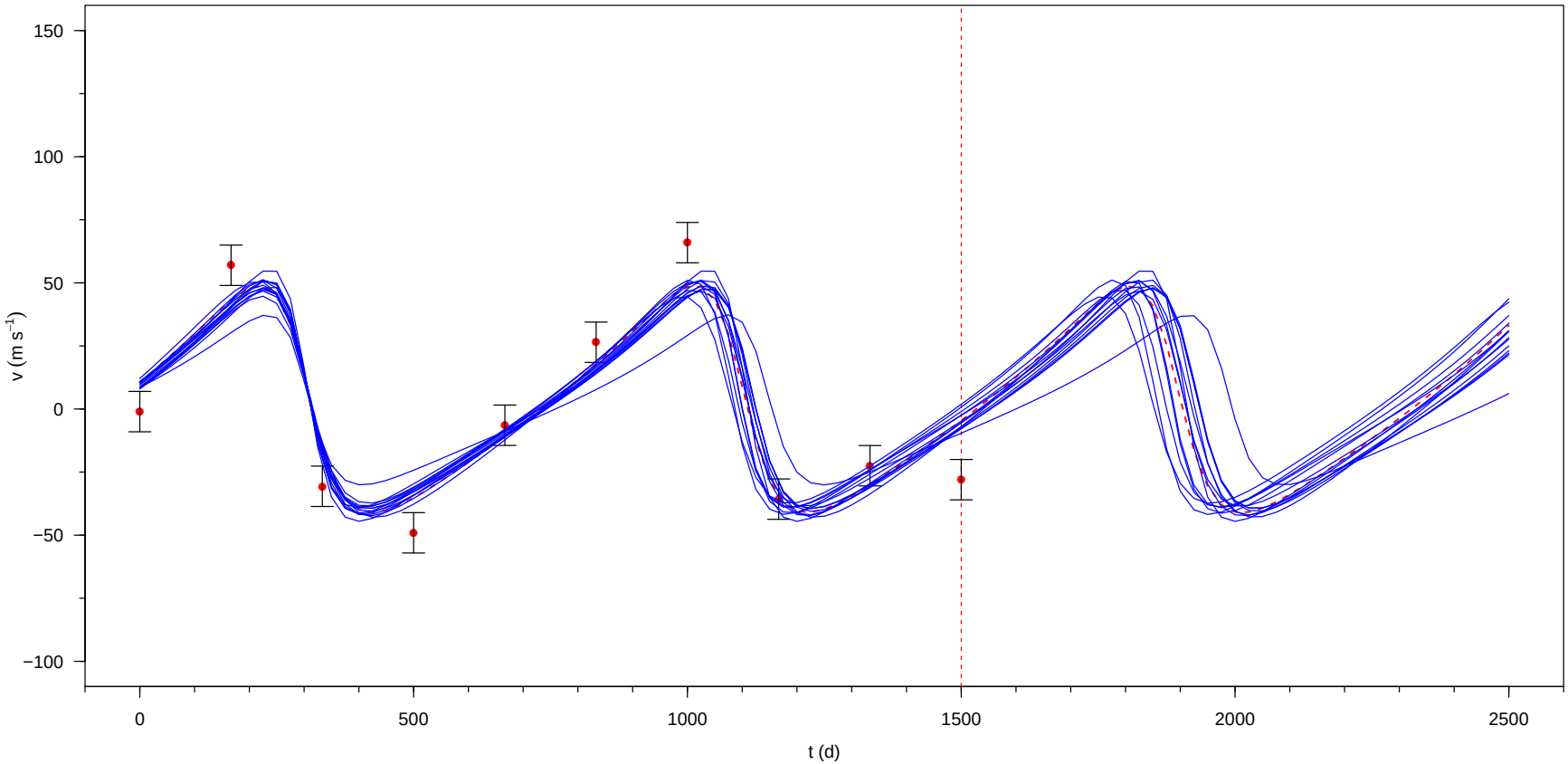



Figure : The true velocity curve and the 15 velocity curves for samples from the posterior distribution.

# Introduction - a motivating example XII

**What is the best sampling time?** This is a *decision problem*.
For each time $t$ we may choose - the *action* - that is for each experiment we may perform - there will be a *consequence*, that is the corresponding value $y$ of the future data.
**Consequences** are evaluated by means of an *utility function* $U(y, t)$.
The *optimal* experiment is the one that maximizes the *expected utility*:

$$\widehat{t} := \arg\max_t \int p(y|t, \bar{\boldsymbol{y}}, \mathcal{M})\, U(y, t)\, dy\,.$$

D. V. Lindley ([5], 1956) proposed, on the basis of Shannon's ideas developed into the theory of information in communication engineering, the following utility function:

$$U(y, t) = \int p(\boldsymbol{\theta}|y, \bar{\boldsymbol{y}}, t, \mathcal{M}) \log p(\boldsymbol{\theta}|y, \bar{\boldsymbol{y}}, t, \mathcal{M})\, d\boldsymbol{\theta}\,.$$

which is the amount of information about $\boldsymbol{\theta}$ with respect to the posterior distribution (also known as negative Shannon entropy of the posterior distribution).

# Introduction - a motivating example XIII

The expected utility function (*expected information gain*)

$$\begin{aligned} I(t) &:= \int p(y|t,\bar{\boldsymbol{y}},\mathcal{M})\, U(y,t)\, dy \\ &= \int p(y|t,\bar{\boldsymbol{y}},\mathcal{M}) \int p(\boldsymbol{\theta}|y,\bar{\boldsymbol{y}},t,\mathcal{M}) \log p(\boldsymbol{\theta}|y,\bar{\boldsymbol{y}},t,\mathcal{M})\, d\boldsymbol{\theta}\, dy\,, \end{aligned}$$

should be maximized with respect to $t$ to provide indications about the best experiment to be performed in order to achieve an optimal reduction of the uncertainty of the model's parameters.
Sebastiani and Wynn ([6]) showed that, if the width of the noise distribution is independent of the underlying signal, then

$$I(t) = -\int p(y|t,\bar{\boldsymbol{y}},\mathcal{M}) \log p(y|t,\bar{\boldsymbol{y}},\mathcal{M})\, dy\,, \tag{2}$$

which is the entropy of the predictive distribution.
The choice of $t$ that maximizes (2) will provide the maximum amount of information about $\boldsymbol{\theta}$ (*maximum entropy sampling principle*).
$I(t)$ can be estimated by means of a Monte Carlo integration technique.

# Introduction - a motivating example XIV

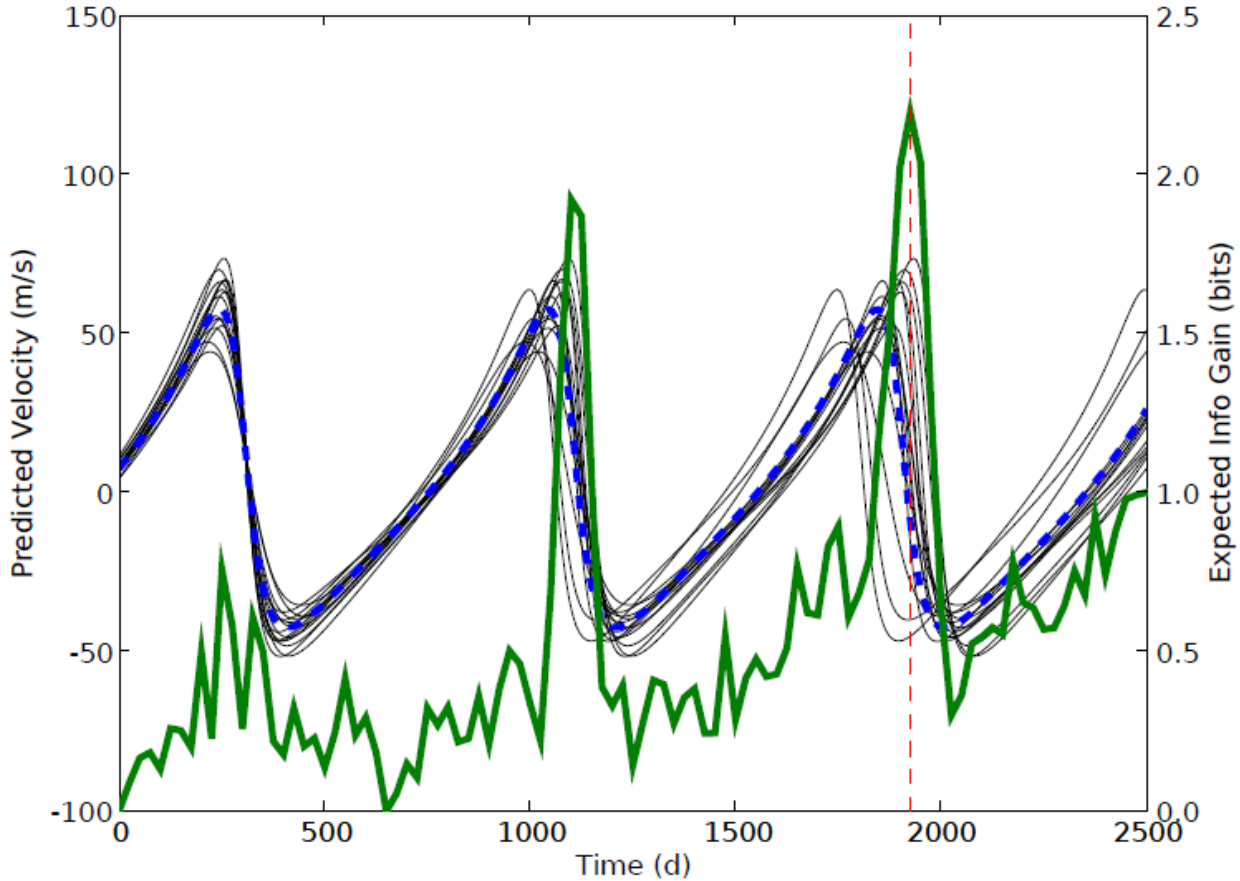



Figure : The true velocity curve, predicted velocity curves for samples from the posterior distribution, Monte Carlo evaluation of the expected information gain. T. J. Loredo ([4], p.341)

# Introduction - a motivating example XV

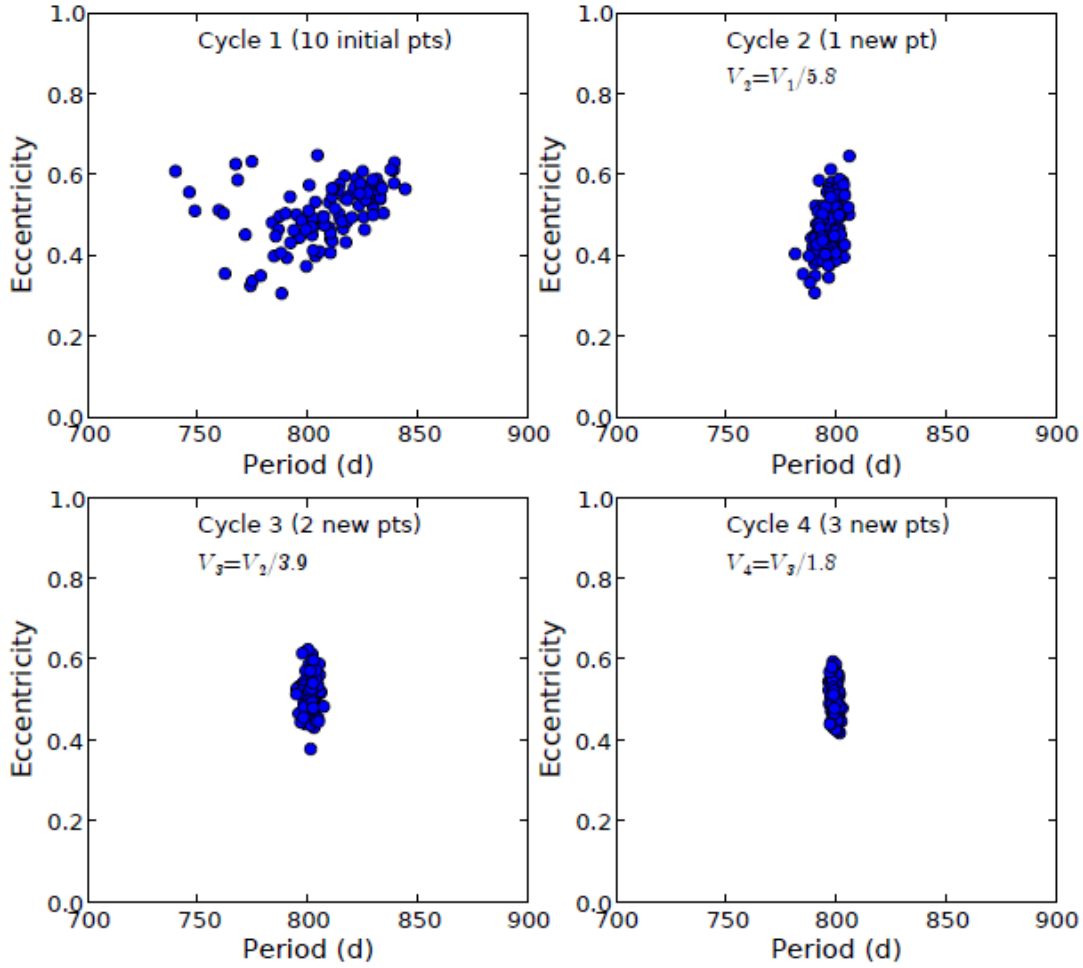



Figure : Sequential uncertainty reduction in the model's parameters $(\tau, e)$. T. J. Loredo ([4], p.342)

# The expected information gain I

Assume that the experiment is performed to make inference (uncertainty reduction) on the vector $\boldsymbol{\theta}$ of unknown parameters in the model

$$\boldsymbol{y}_i = \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\xi}) + \boldsymbol{\epsilon}_i \,, i = 1, \ldots, M \,,$$

where

- $\boldsymbol{\xi}$ denotes the vector of design parameters,
- $\boldsymbol{y}_i$ is the $i$th observation vector, and
- we suppose that the additive noise is such that $\boldsymbol{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ .

Let $p(\boldsymbol{\theta})$ be the prior density of $\boldsymbol{\theta}$ .

An alternative definition of the expected information gain can be based on the so-called *Kullback-Leibler (K-L) divergence* between the prior distribution $Q$ and the posterior distribution $P$ for the parameter vector $\boldsymbol{\theta}$.

# The expected information gain II

The **K-L divergence** (information gain) between prior and posterior of $\theta$ is

$$D_{KL}(\bar{y}, \xi) = \int_{\Theta} \log\left(\frac{p(\theta|\bar{y}, red\xi)}{p(\theta)}\right) p(\theta|\bar{y}, \xi)\, d\theta\,.$$

(if $p(\theta|\bar{y}, \xi) = p(\theta)$, then $D_{KL} = 0.$ )
The **expected information gain** in $\theta$ is given by

$$\begin{aligned} I(\xi) &:= ED_{KL}(\bar{y}, \xi) = \int_{\mathcal{Y}} D_{KL}(\bar{y}, \xi)\, p(\bar{y}|\xi)\, d\bar{y} \\ &= \int_{\mathcal{Y}} \int_{\Theta} \log\left(\frac{p(\theta|\bar{y}, \xi)}{p(\theta)}\right) p(\theta|\bar{y}, \xi)\, d\theta\, p(\bar{y}|\xi)\, d\bar{y} \geq 0\,. \end{aligned}$$

After using Bayes' theorem

$$I(\xi) = \int_{\mathcal{Y}} \int_{\Theta} \log\left(\frac{p(\bar{y}|\theta, \xi)}{p(\bar{y}|\xi)}\right) p(\theta)\, p(\bar{y}|\theta, \xi) d\theta\, d\bar{y}\,. \qquad (3)$$

The *optimal* design is:

$$\widehat{\xi} := \arg\max_{\xi} I(\xi)\,.$$

# The expected information gain III

Observe that the optimality criterion is an average of the amount of information given by the decrease in the Shannon entropy of the distribution of $\boldsymbol{\theta}$:

$$-\int_{\Theta} p(\boldsymbol{\theta}) \log p(\boldsymbol{\theta})\, d\boldsymbol{\theta} + \int_{\Theta} p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi}) \log p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi})\, d\boldsymbol{\theta}\,,$$

and, after taking expectation with respect to all possible experimental scenarios, we get

$$\int_{\mathcal{Y}} \left[ \int_{\Theta} \left( p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi}) \log p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi}) - p(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}) \right) d\boldsymbol{\theta} \right] p(\bar{\boldsymbol{y}}|\boldsymbol{\xi})\, d\bar{\boldsymbol{y}}\,,$$

that equals the preposterior expectation of the Kullback-Leibler divergence between the prior and the posterior distribution of the parameter vector $\boldsymbol{\theta}$:

$$\int_{\mathcal{Y}} \int_{\Theta} \log \left( \frac{p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi})}{p(\boldsymbol{\theta})} \right) p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}, \boldsymbol{\xi}) d\boldsymbol{\theta}\, p(\bar{\boldsymbol{y}}|\boldsymbol{\xi}) d\bar{\boldsymbol{y}} = I(\boldsymbol{\xi})\,.$$

Observe that the K-L divergence will be hardly available in a closed form.

# The expected information gain IV

However, as an illustration, we may compute easily the K-L divergence between, for example, a Gaussian prior $\mathcal{N}(\mu_0, \sigma_0^2))$ and a Gaussian posterior $\mathcal{N}(\mu_1, \sigma_1^2)$:

$$D_{KL}(\text{post}, \text{prior}) = \log\left(\frac{\sigma_0}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_0^2}\right) + \frac{(\mu_0 - \mu_1)^2}{2\sigma_0^2} - \frac{1}{2}\,.$$

The K-L divergence between a multivariate Gaussian prior $\mathcal{N}_d(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and a multivariate Gaussian posterior $\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is given by

$$D_{KL}(\text{post}, \text{prior}) = \frac{1}{2}\left\{\log\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} - d + \text{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right\}.$$

Useful formulas have been obtained for broad class of distributions, for example, multivariate skew-elliptical distributions. See ([7], Genton et al., 2013) where these tools have been applied for the optimal design of an ozone monitoring station network.

To determine the *optimal* design, a *fast numerical integration strategy* is needed in order to compute, for each design parameter $\boldsymbol{\xi} \in \mathcal{D}$, the value of the criterion function $I(\boldsymbol{\xi})$.

# Double–loop Monte Carlo I

The integral over the data and the parameter space that defines the expected information gain

$$I(\boldsymbol{\xi}) = \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} \log\left(\frac{p(\bar{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{\xi})}{p(\bar{\boldsymbol{y}}|\boldsymbol{\xi})}\right) p(\bar{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{\xi})\, d\bar{\boldsymbol{y}}\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}\,,$$

can be evaluated, for each $\boldsymbol{\xi}$, by means of a Monte Carlo sampling strategy ([8], [9]):

$$I(\boldsymbol{\xi}) \approx \frac{1}{M_o} \sum_{k=1}^{M_o} \log\left(\frac{p(\bar{\boldsymbol{y}}_k|\boldsymbol{\theta}_k,\boldsymbol{\xi})}{p(\bar{\boldsymbol{y}}_k|\boldsymbol{\xi})}\right)\,,$$

where $\boldsymbol{\theta}_k$ is drawn from $p(\boldsymbol{\theta})$, $\bar{\boldsymbol{y}}_k$ is drawn from $p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_k,\boldsymbol{\xi})$. The term "double–loop" originates from the nested Monte Carlo sampling that is needed to evaluate the marginal density

$$p(\bar{\boldsymbol{y}}_k|\boldsymbol{\xi}) = \int_{\boldsymbol{\Theta}} p(\bar{\boldsymbol{y}}_k|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \approx \frac{1}{M_i} \sum_{j=1}^{M_i} p(\bar{\boldsymbol{y}}_k|\boldsymbol{\theta}_j)\,,$$

where $\boldsymbol{\theta}_j$ is drawn from $p(\boldsymbol{\theta})$.

# Double–loop Monte Carlo II

$M_o \times M_i$ drawings are needed to estimate $I(\boldsymbol{\xi})$.
The outer loop ($M_o$) controls the variance of the estimate, the inner loop ($M_i$) its bias.
The application of this method can be extremely costly when the data becomes available after solving, for each $\boldsymbol{\theta}$, a numerical problem involving the approximation of the solution of partial differential equations.
Ryan ([8], p.589) proposed to use the estimator of $I(\boldsymbol{\xi})$ given by

$$\frac{1}{M}\sum_{l=1}^{M} \log \frac{p(\{\boldsymbol{Y}_i\}_l|\boldsymbol{\theta}_l, \boldsymbol{\xi})}{\frac{1}{L}\sum_{j=1}^{L} p(\{\boldsymbol{Y}_i\}|\boldsymbol{\theta}^*_{lj}, \boldsymbol{\xi})}$$

with an importance sampling based estimator for $p(\{\boldsymbol{y}_i\}_l|\boldsymbol{\xi})$, where $\boldsymbol{\theta}^*_{lj}, l = 1, \ldots, M, j = 1, \ldots, L$ are $M$ samples of size $L$ from $p(\boldsymbol{\theta})$ obtained independently of the $N$ pairs $(\{\boldsymbol{y}_i\}_l, \boldsymbol{\theta}_l)$.

**Our approach to estimate the expected information gain $I(\boldsymbol{\xi})$ will rely upon the Laplace approximation, which is closely connected to the asymptotic normality of the posterior distribution ([10], p.115).**

# Laplace's method - a reminder I

*Approximation is ubiquitous in both statistical theory and practice ([11], p.1358)*

For a classical treatment of Laplace's method see, for instance, the book by Wong ([12], pp.55–66) or the book by Evans and Swartz ([13], pp.62–70).
Laplace's method provides an approximation for integrals of the form

$$I_M = \int_a^b f(\theta) e^{M\,L(\theta)}\, d\theta$$

when $M$ is large.
The idea is that if $L$ has a unique maximum at $\hat{\theta} \in (a, b)$, where $L'(\hat{\theta}) = 0$ and $L''(\hat{\theta}) < 0$, then for $M$ large the value of $I_M$ depends on the behavior of $L$ near its maximum

$$I_M \approx f(\hat{\theta})\, e^{M\,L(\hat{\theta})} \left( \frac{-2\pi}{ML''(\hat{\theta})} \right)^{1/2} \quad \text{as } M \to \infty\,.$$

# Laplace's method - a reminder II

In the multiparameter case, under regularity conditions, the integral

$$I_M = \int_{\mathcal{R}^d} f(\boldsymbol{\theta})\, e^{ML(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\,,$$

can be expressed by

$$I_M = f(\hat{\boldsymbol{\theta}})\, e^{ML(\hat{\boldsymbol{\theta}})} (2\pi)^{d/2} M^{-d/2} |-\nabla\nabla L(\hat{\boldsymbol{\theta}})|^{-1/2} \left(1 + O(M^{-1})\right),$$

where $\nabla\nabla L(\boldsymbol{\theta})$ denotes the Hessian of $L$

$$\underset{d\times d}{\nabla\nabla L(\boldsymbol{\theta})} = \left(\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(\boldsymbol{\theta})\right)\right)_{ij}.$$

*Tierney and Kadane ([14], 1986) convincingly demonstrated that Laplace's approximation holds significant practical value for Bayesians. ...(Their) article clearly renewed general interest in Laplace's method among statisticians working in many different area of statistics ([11], p.1362).*

# Estimation of $I(\xi)$ for determined models I

- **Synthetic Data model**:

$$\boldsymbol{y}_i = \boldsymbol{g}(\boldsymbol{\theta}_0, \boldsymbol{\xi}) + \boldsymbol{\epsilon}_i\,, \quad i = 1, \ldots, M\,,$$

$\boldsymbol{\theta}_0$ is the "true" parameter vector that generates, through the "true" model $\boldsymbol{g}$, the synthetic data, $\boldsymbol{\xi}$ is the experimental set-up, $\boldsymbol{y}_i$ is the $i$th measurement, $\boldsymbol{\epsilon}$ is the measurement noise, $M$ is the number of repeated experiments.
Let $p(\boldsymbol{\theta})$ be the prior density for $\boldsymbol{\theta}$ and define $h(\boldsymbol{\theta}) := \log(p(\boldsymbol{\theta}))$.

- The posterior density for $\boldsymbol{\theta}$ is proportional to

$$p(\boldsymbol{\theta}|\bar{\boldsymbol{y}}) \propto \prod_{i=1}^{M} \exp\left(-\frac{1}{2}\boldsymbol{r}_i(\boldsymbol{\theta})^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{r}_i(\boldsymbol{\theta})\right) p(\boldsymbol{\theta})\,, \tag{4}$$

where $\boldsymbol{r}_i$ is the residual for the $i^{th}$ measurement,

$$\boldsymbol{r}_i(\boldsymbol{\theta}) = \boldsymbol{y}_i - \boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}_0) + \boldsymbol{\epsilon}_i - \boldsymbol{g}(\boldsymbol{\theta})\,,$$

and $\boldsymbol{\Sigma}_\epsilon$ is the covariance matrix of the experimental Gaussian additive noise.

# Estimation of $I(\xi)$ for determined models II

- The Laplace's method leads to the following normal approximation for the posterior density of $\boldsymbol{\theta}$

$$\tilde{p}(\boldsymbol{\theta}|\bar{\boldsymbol{y}}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{2}\right), \quad (5)$$

where $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{2}\sum_{i=1}^{M} \boldsymbol{r}_i(\boldsymbol{\theta})^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{r}_i(\boldsymbol{\theta}) - h(\boldsymbol{\theta})\right]$ is the maximum posterior estimate for $\boldsymbol{\theta}$, and $\boldsymbol{\Sigma} = \boldsymbol{H}_F(\hat{\boldsymbol{\theta}}) := \nabla\nabla F(\theta)$, with

$$F(\boldsymbol{\theta}) := -\log(p(\boldsymbol{\theta}|\bar{\boldsymbol{y}})) = \frac{1}{2}\sum_{i=1}^{M} \boldsymbol{r}_i(\boldsymbol{\theta})^T\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{r}_i(\boldsymbol{\theta}) - h(\boldsymbol{\theta}) + C_1\,.$$

# Estimation of $I(\xi)$ for determined models III

We derived ([15], pp.26–27) the following asymptotic approximation for $I(\boldsymbol{\xi})$:

$$
\begin{aligned}
I(\boldsymbol{\xi}) &= \int_{\mathcal{Y}} D_{KL}(\bar{\boldsymbol{y}}, \boldsymbol{\xi}) p(\bar{\boldsymbol{y}}|\boldsymbol{\xi})\, d\bar{\boldsymbol{y}} \\
&= \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} D_{KL}(\bar{\boldsymbol{y}}, \boldsymbol{\xi})\, p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0, \boldsymbol{\xi})\, d\bar{\boldsymbol{y}}\, p(\boldsymbol{\theta}_0)\, d\boldsymbol{\theta}_0 \\
&= \int_{\boldsymbol{\Theta}} \int_{\mathcal{Y}} \left[ -\frac{1}{2} \log((2\pi)^d |\boldsymbol{\Sigma}|) - \frac{d}{2} - h(\hat{\boldsymbol{\theta}}) - \frac{\mathrm{tr}(\boldsymbol{\Sigma}\, \nabla\nabla h(\hat{\boldsymbol{\theta}}))}{2} \right] \\
&\times \quad p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}_0) d\bar{\boldsymbol{y}} p(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0 + O\left(\frac{1}{M^2}\right), \qquad (6)
\end{aligned}
$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the posterior distribution, $d$ is the dimension of $\boldsymbol{\theta}$, $h(\boldsymbol{\theta})$ is the log prior, and $\hat{\boldsymbol{\theta}}$ is the maximum posterior estimate for $\boldsymbol{\theta}$.

# Estimation of $I(\xi)$ for determined models IV

By using the first order approximation $\hat{\theta} = \theta_0 + O_p(\frac{1}{\sqrt{M}})$ we can show that

$$\begin{aligned} I(\xi) &= \int_{\Theta} \left[ -\frac{1}{2} \log((2\pi)^d |\mathbf{\Sigma}|) - \frac{d}{2} - h(\theta_0) - \frac{\mathrm{tr}(\mathbf{\Sigma}\, \nabla\nabla h(\theta_0))}{2} \right] \\ &\times p(\theta_0) d\theta_0 + O\left(\frac{1}{M}\right) . \end{aligned} \tag{7}$$

An estimator of $I(\xi)$ based on the Laplace approximation and a simple Monte Carlo strategy is given by

$$I(\xi) \approx \frac{1}{N} \sum_{i=1}^{N} \left( -\frac{1}{2} \log((2\pi)^d |\mathbf{\Sigma}|) - \frac{d}{2} - h(\theta_i) - \frac{\mathrm{tr}(\mathbf{\Sigma}\, \nabla\nabla h(\theta_i))}{2} \right) . \tag{8}$$

Remark

*The estimator (8) is consistent only when the conditions for the validity of the Laplace approximation are fulfilled.*

# Estimation of $I(\xi)$ for determined models V

As an illustration, consider the following model (proposed in ([9], 2013):

$$y(\theta, \xi) = \theta^3 \xi^2 + \theta\, e^{-|0.2-\xi|} + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma_m^2)$$

and assume a uniform prior distribution $\mathcal{U}(0.5, 1.5)$ for $\theta$.

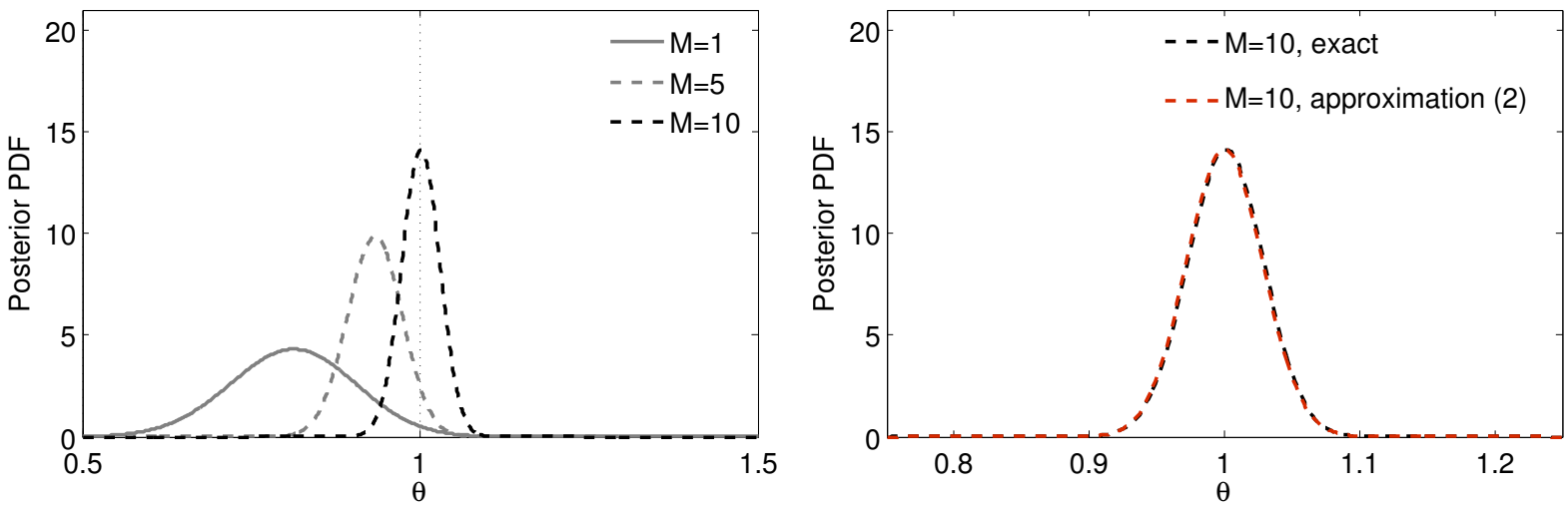
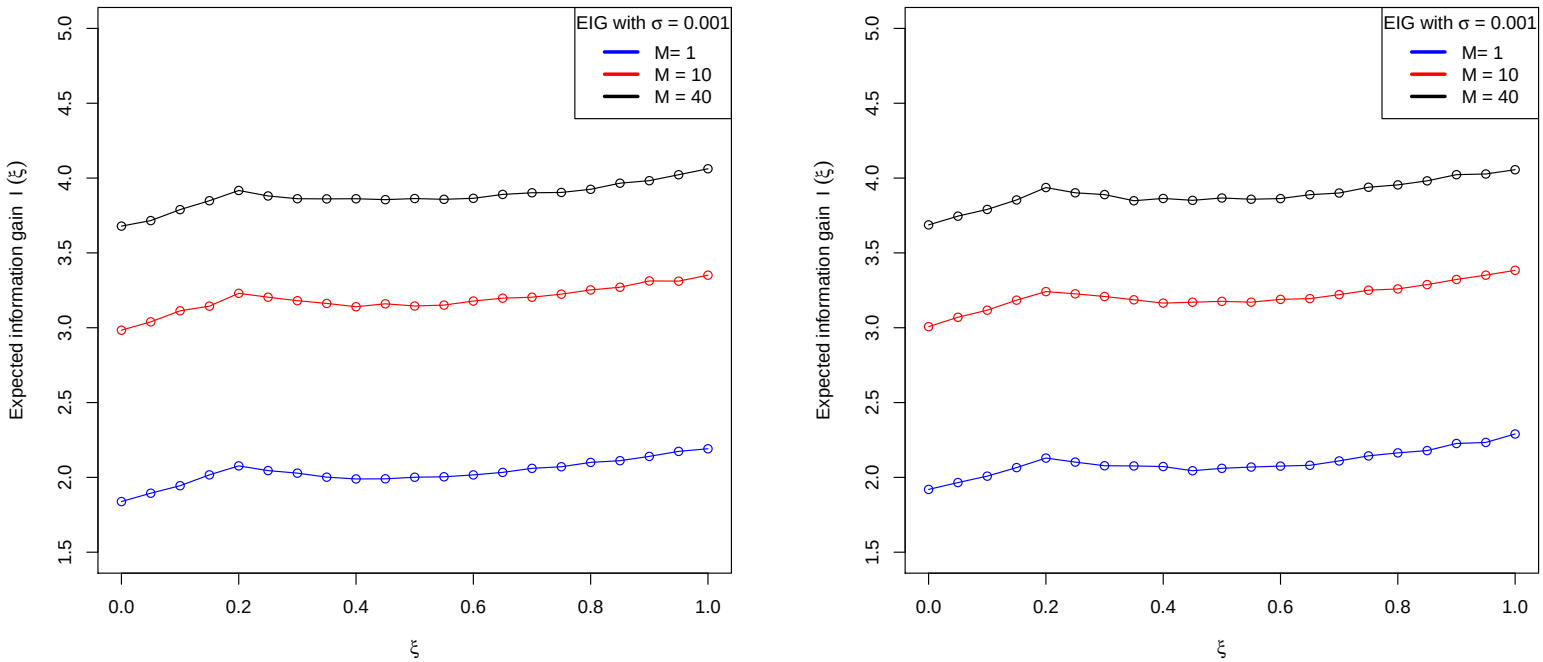



Figure : Posterior densities for $\theta$ (generated using the following parameters: $\sigma_m^2 = 0.01$, $\xi = 0.2$, $\theta_0 = 1$) and the Laplace-based approximation density for $\theta$.

# Estimation of $I(\xi)$ for determined models VI




(a) (b)

Figure : Left: Expected information gains estimated by Laplace approximation. Right: Expected information gains estimated by DLMC with $10^4 \times 10^4$ samples.

# Estimation of $I(\xi)$ for determined models VII

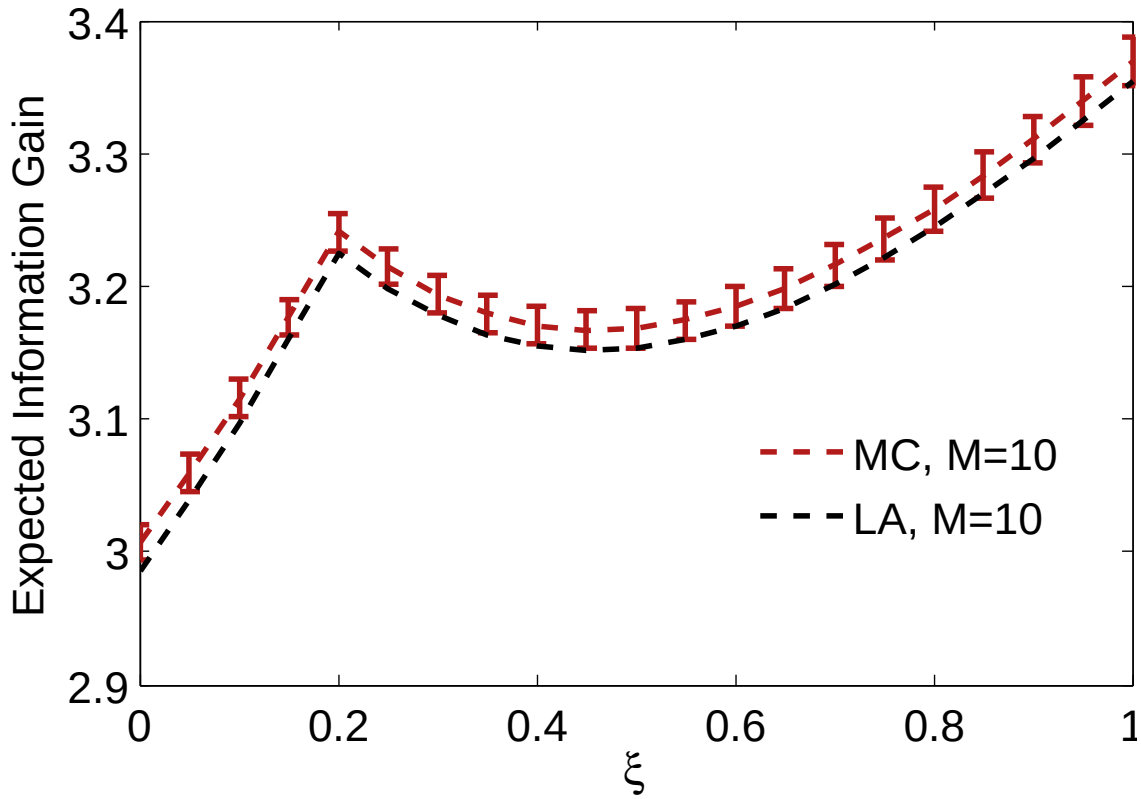



Figure : A zoomed–in version for the case $M = 10$. The error bar represents the 97.5% confidence interval, whose width is approximately 0.5% of the magnitude of the expected information gain.

# Estimation of $I(\xi)$ for under-determined models I

How to deal with the cases where there are unidentifiable parameters?

**Example:** $g(\boldsymbol{\theta}, \xi) = (\theta_1^2 + \theta_2^2)^3 \xi^2 + (\theta_1^2 + \theta_2^2)\, e^{-|0.2-\xi|}$

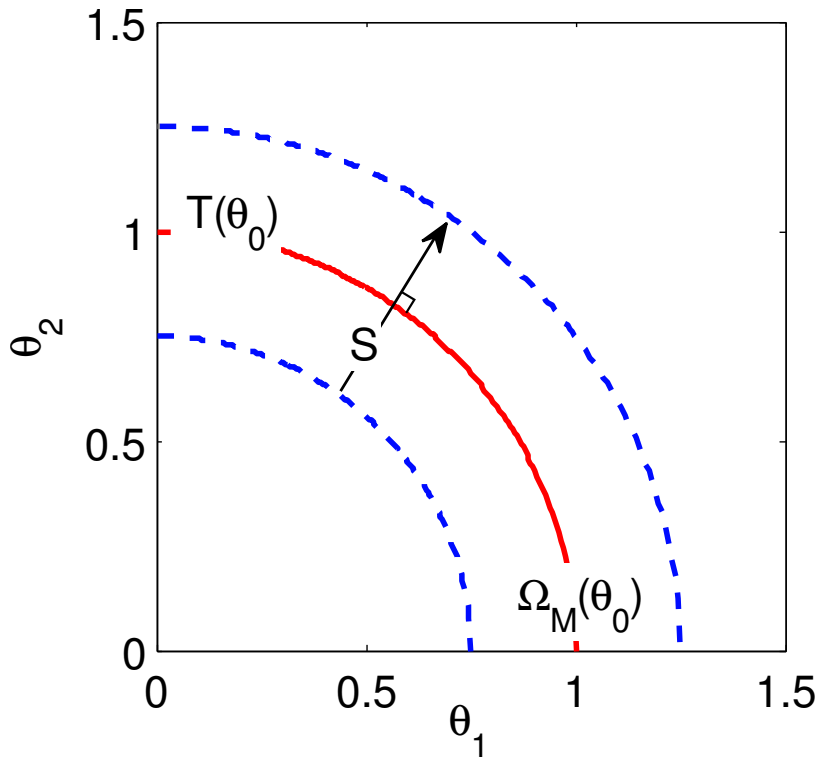



Figure : Illustrative unidentifiable submanifold.

# Estimation of $I(\xi)$ for under-determined models II

- **Data model:**

$$y = (\theta_1 + \theta_2)^3 \xi^2 + (\theta_1 + \theta_2)\, e^{-|0.2-\xi|} + \epsilon\,,$$

with $\epsilon \sim \mathcal{N}(0, \sigma_m^2)$.

- Gaussian mixture model prior for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}) = \frac{1}{2}\, p_1(\boldsymbol{\theta}) + \frac{1}{2}\, p_2(\boldsymbol{\theta})\,,$$

where $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$ are the densities of two multivariate Gaussian distributions with mean vectors $[2, 0]'$ and $[0, 2]'$, respectively, and covariance matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

# Illustrative example

$y = (\theta_1 + \theta_2)^3 \xi^2 + (\theta_1 + \theta_2)\, e^{-|0.2-\xi|} + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 10^{-3})$.
Gaussian mixture model prior for $\theta$

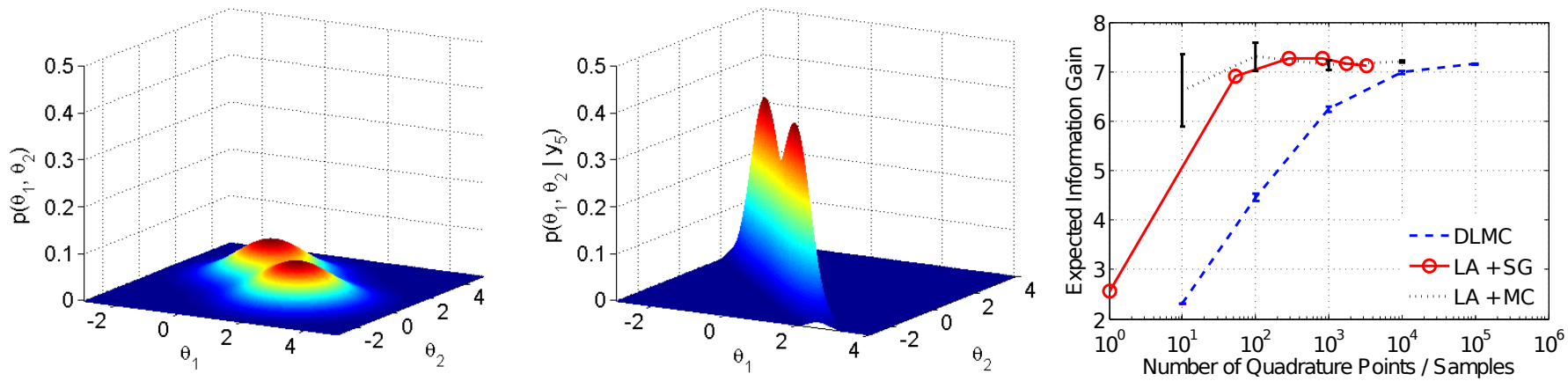



Figure : Left: the prior pdf; middle: the posterior pdf (M = 5); right: the convergence of the expected information gain computed by projective Laplace approximation using Monte Carlo Sampling and sparse grid numerical integration, and by DLMC. $\xi = 1$.

# Illustrative example

Log Gaussian mixture model prior for $\gamma = \log \theta$

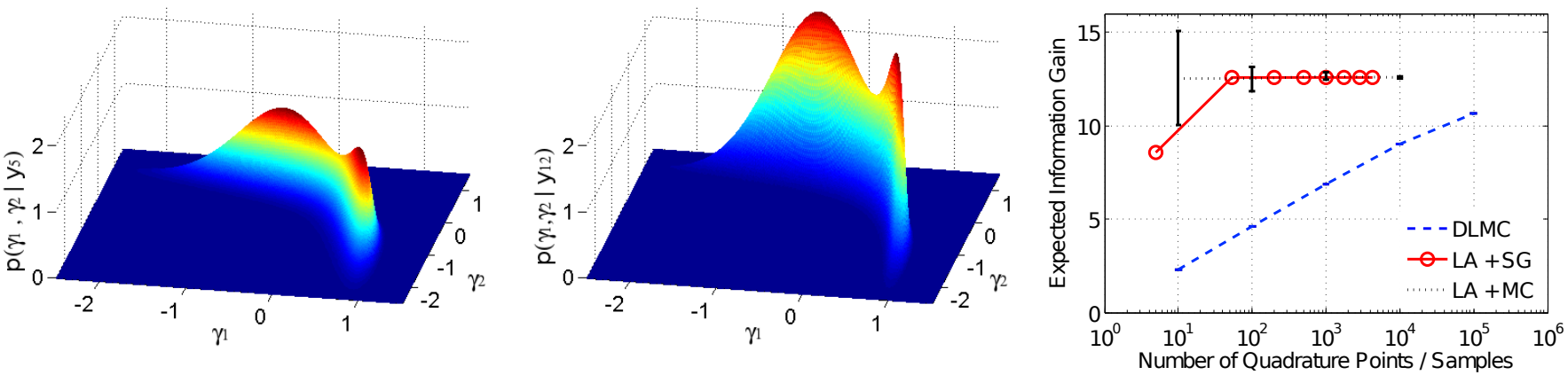



Figure : Left: the posterior pdf (M = 5); middle: the posterior pdf (M = 12); right: the convergence of the expected information gain computed by projective Laplace approximation using Monte Carlo Sampling and sparse grid numerical integration, and by DLMC. $\xi = 1$.

# Conclusions

- We presented some fundamental issues that characterize the Bayesian learning process based on experimental data, when the main interest is to make inference about the unknown parameters of the proposed model.
- We stressed the role of information theoretic measures to obtain optimal designs.
- We proposed a strategy for fast numerical computation of the expected information gain by means of the Laplace method. Some illustrative examples show the remarkable computational advantage of such method when compared with the double–loop Monte Carlo sampling integration technique.
- In our recent work we explored how to adapt the Laplace method for unidentifiable problems.

# References I


K. Chaloner, I. Verdinelli

*Bayesian experimental design: a review*, Statistical Science **10** 273–304, 1995.

P. Gregory Bayesian Logical Data Analysis for the Physical Sciences, Cambridge University Press, 2005.

T.J. Loredo, D.F. Chernoff

*Bayesian adaptive exploration*, in Statistical Challenges in Astronomy, 57–70, Springer, 2003.

T.J. Loredo,

*Bayesian adaptive exploration*, in Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 330-346, 2004.

D.V. Lindley

*On a measure of the information provided by an experiment*, in Annals of Statistics, **27** 986–1005, 1956.

P. Sebastiani, H.P. Wynn

*Maximum entropy sampling and optimal Bayesian experimental design*, Journal of the Royal Statistical Society, series B, **62** 145-157, 2000.

R.B. Arellano-Valle, J.E. Contreras-Reyes, M. Genton

*Shannon entropy and mutual information for multivariate skew-elliptical distributions*, Scandinavian Journal of Statistics, **40** 42-62, 2013.

# References II


K.J. Ryan

*Estimating expected information gains for experimental designs with application to the random fatigue-limit model,* Journal of Computational and Graphical Statistics, **12** 585–603, 2003.

X. Huan, Y.M. Marzouk

*Simulation–based optimal Bayesian experimental design for nonlinear systems,* Journal of Computational Physics, **232** 288–317, 2013.

J.K. Ghosh, M. Delampady, T. Samanta An Introduction to Bayesian Analysis, Springer, 2006.

R.L. Strawderman

*Higher–order asymptotic approximation: Laplace, saddlepoint, and related methods,* Journal of the American Statistical Association, **95** 1358–1364, 2000.

R. Wong Asymptotic Approximation of Integrals, SIAM, 2001.

M. Evans, T. Swartz Approximating Integrals via Monte Carlo and Deterministic Methods, Oxford University Press, 2000.

L. Tierney, J.B. Kadane

*Accurate approximations for posterior moments and marginal densities,* Journal of the American Statistical Association, **81** 82–86, 1986.

Q. Long, M. Scavino, R. Tempone and S. Wang

*Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations,* Computer Methods in Applied Mechanics and Engineering **259 24–39, 2013.**

# References III

Q. Long, M. Scavino, R. Tempone and S. Wang

*A Laplace method for under-determined Bayesian optimal experimental designs*, Resubmitted after first round of refereeing, 2014.