# General Bayesian updating

Chris Holmes[1]

Department of Statistics and,
Wellcome Trust Centre for Human Genetics,
University of Oxford

UCL Big Data 2015

[1] joint work with Stephen Walker

# Overview

- Bayesian statistics in a "big-data" world
- The problem of $\mathcal{M}$-open
- Decision theoretic solutions
- Illustrations

# Background Motivation

- Bayesian analysis provide a coherent approach to updating of beliefs typically through the use of "Bayes Theorem"

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$

  where

  - $f(x|\theta)$ is a sampling distribution (likelihood) for the data
  - $\pi(\theta)$ represents prior beliefs on the unknown true value of $\theta$
  - $\pi(\theta|x)$ represents updated beliefs about the unknown $\theta$ in light of the data $x$

- Bayesian analysis is rooted in decision theory (Savage 1954), it is axiomatic, intuitive, and coherent; where all aspects of uncertainty are accommodated through the specification of a joint probability model used as a vehicle to quantify uncertainty on all unknowns, $\pi(x,\theta) = f(x|\theta)\pi(\theta)$
  - All of Bayesian statistics is model based

# Challenges from a big-data world

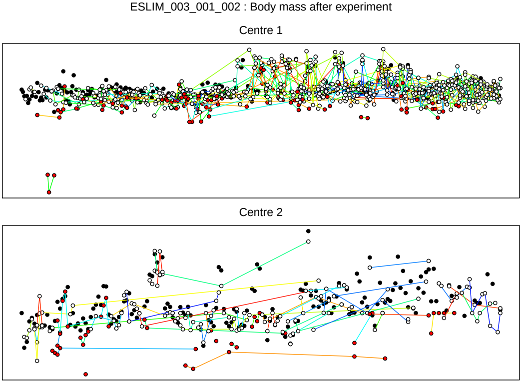
- However, Bayesian updating is also highly restrictive in the need to assume a joint probability for everything observed, and moreover assume that the model is true,
  - $f(x|\theta)$, true likelihood for all measurements
  - $f(x) = \int_\theta f(x|\theta)\pi(\theta)d\theta$, true joint density ("the model") for $x$
- In modern applications such a requirement can be highly restrictive and cumbersome ($\mathcal{M}$-open problem)
- Information maybe highly heterogeneous, high-dimensional and non-stochastic
  - news snippets, twitter feeds,
  - $x = \{$your genome, medical image, electronic health record$\}$
  - partial information under privacy constraints, $p$-values

  it's difficult to think of joint models for $x$, yet $x$ is highly relevant to learning about $\theta$
- Taken together, Bayesian inference can be challenging, even for supposedly simple problems

# Motivation: International Mouse Phenotyping Consortium

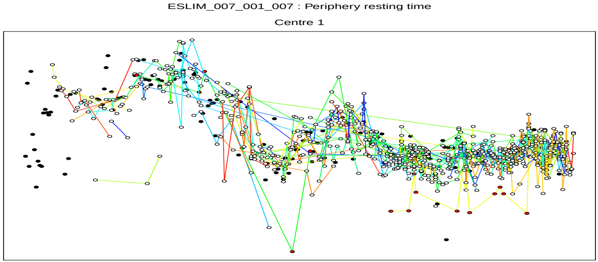
- The International Mouse Phenotyping Consortium (http://www.mousephenotype.org/) is a 10 year study to systematically characterise the functional consequences of each of around 20,000 genes in the mouse genome
- Recording over 1500 measurements per mouse (leading to around 700 phenotypes), around 7 mice per knockout ($\times$ 2 sexes) and matched controls
- IMPC will deliver complex multivariate measurements on around 560,000 mice $\times$ 690 dependent phenotypes across 8 Centres
  - costing $100M's

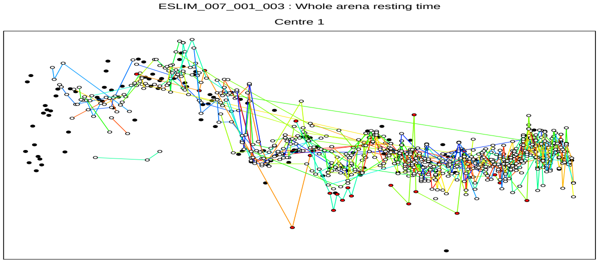
# Example Data



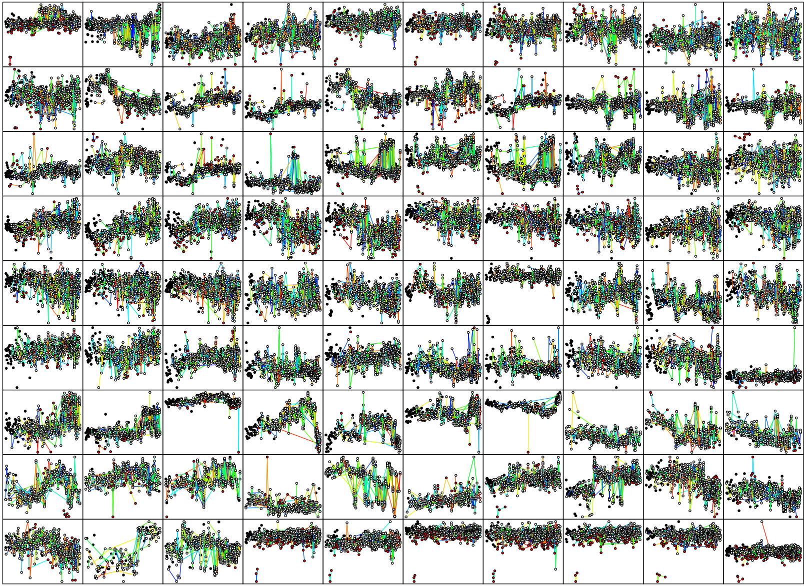

- Time points represents a (robust) mean of (transformed) measurement on a litter, on a particular day, at a specific Centre
- Lines connect repeated measurements; Black dots are controls; Circles are mutant lines
- Red dots are putative mutants that show systematic differences
  - controlling for meta-data collected on technician, reagents, ...

# Many of the phenotypes show high dependence



# ....and there are many phenotypes (90 of 690) of 35,000,000 data points



# Bayesian analysis

- Formal Bayesian analysis approaches of such data structures are hard to formulate
- Of course we could use approximate models, for example Variational Bayes, but then what are we targeting?
- That is, what does $\pi(\theta|x)$ actually represent if I know that $x$ does not arise from $f(x|\theta)$?

# Our research

- We have been considering a more general framework for updating of beliefs

$$\pi(\theta) \to \pi(\theta|x)$$

  on a well defined $\theta$ of interest given information $x$, without having to assume a known component $x \sim f(x)$

- The update needs to be coherent (to be defined later), principled (decision theoretic) and open to inspection

- The central idea is the replacement of $f(x|\theta)$ with a general loss function $l(x,\theta)$ that is used to connect information in the data to the value of $\theta$ minimising the population expected loss
  - and $l(x,\theta)$ can accommodate partial information, non-stochastic information, ...

- Importantly the procedure should coincide with Bayesian inference if $f(x) = \int f(x|\theta)\pi(\theta)\mathrm{d}\theta$ is assumed known

# Toy Example – not only big problems cause problems

- Consider that you want to infer the *median* patient survival time, $\theta$, for a particular population,

$$\theta = F_0^{-1}(0.5)$$

  where $F_0$ is the unknown distribution of survival times

Suppose:

- You hold subjective prior beliefs on $\theta$, expressible via $\pi(\theta)$
- You don't know $F_0$
- You obtain independent observations of survival times $\boldsymbol{x} = \{x_1, \ldots, x_n\}$

It feels that an update of beliefs, $\pi(\theta) \to \pi(\theta|x)$, should be possible

Yet the Bayesian solution to this problem is highly non-trivial

# Functionals of interest

- Instead we consider learning about the minimiser of some functional,

$$\begin{aligned} L(\theta) &= \int l(\theta, x)\, \mathrm{d}F_0(x), \\ \theta_0 &= \arg \inf_{\theta \in \Theta} L(\theta) \end{aligned}$$

  for some loss function $l(\theta, x)$ introduced to target $\theta_0$, where $F_0(x)$ is the unknown distribution function from which i.i.d. observations arise

- It may be easier to think of this as

$$\begin{aligned} \theta_0 &= \arg\min_{\theta} \left[ \sum_{i=1}^{n \to \infty} l(\theta, x_i) \right] \\ x_i &\sim f_0(x) \end{aligned}$$

  for $f_0$ unknown, and $\theta_0$ represents the optimal value of $\theta$ under an infinite sample size

# Update

- If $\pi(\theta)$ represents prior beliefs about this $\theta_0$, and $x$ is observed from $F_0$, then we will argue that a valid and coherent update of $\pi(\cdot)$ is to the posterior $\pi(\cdot|x)$, where

$$\pi(\theta|x) \propto \exp\{-l(\theta, x)\}\, \pi(\theta).$$

- It is important to note that:
  - $\pi(\theta|x)$ does not involve the unknown $f_0(x)$ and
  - this update is not an approximation, but a valid representation of beliefs about the value of $\theta_0$ in more general circumstances when $f(x)$ is unknown ($\mathcal{M}$-open problems)

- We have replaced the more ambitious task of learning about a "true" parameter for $f(x|\eta)$, with that of learning about a $\theta_0$

# Model Sufficiency

- Underlying the justification is the notion of model sufficiency, namely that $\theta_0$ is sufficient for the analyst to make a decision and that if $\theta_0$ was ever known then the data $x$ contains no further information to the decision process
- That is, given $\theta_0$ then the inference task is solved and the optimal action will be revealed $U(a, \theta_0)$, where $U$ denotes a utility function on action space $a$
- In this sense $\pi(\theta|x)$ is sufficient for the decision task, and the remaining information in $x$ can be discarded
- For example, the use of a logistic regression classification model, is a statement that knowledge of he MAP estimates under an infinite sample reveals the optimal action

# Constructing the update

- We have two independent pieces of information in $\{\pi(\theta), x\}$
- We consider a coherent scoring rule on the space of probability measures, given $\{\pi(\theta), x\}$, and then show that the optimal distribution with highest score, $\pi(\theta|x)$, can be identified
- As the data and the prior represent independent pieces of information we will naturally assume additivity of loss
- So we can score any distribution (model), $\pi'(\theta)$, on $\theta$ using

$$\begin{aligned} S(\pi'; \{x, \pi\}) &= L_x(\pi', x) + L_\pi(\pi', \pi) \\ &= \text{loss to data} + \text{loss to prior} \end{aligned}$$

and we will then select the optimal model (distribution) $\pi'$ which minimizes expected loss, over the space of all valid probability measures

$$\tilde{\pi} = \arg\min_{\pi'} S(\pi'; \{x, \pi\})$$

This is optimisation of probability measures, rather than parameters. This is the formal way to proceed (Key, Pericchi, Smith; B&S)

# Scoring belief distributions

- The empirical loss to each datum, $L_x(\pi', x_i)$, is given by

$$L_x(\pi', x_i) = \int_\theta l(\theta, x_i)\pi'(\theta)\mathrm{d}\theta$$

  where $l(\theta, x_i)$ is the loss-function targeting $\theta_0$

- The loss to the prior, $L_\pi(\pi', \pi)$, will be some divergence score between probability measures,

$$D(\pi', \pi) = \int g(\mathrm{d}\pi'/\mathrm{d}\pi)\mathrm{d}\pi'$$

  where g is a convex function measuring divergence from $(0, \infty)$ to the real line and $g(1) = 0$. See Ali and Silvey (1966).

- From the convexity of the $g$-divergence we can equivalently write the optimisation as

$$\tilde{\pi} = \arg\min_{\pi'} \left[L_x(\pi', x)\right] \quad \text{s.t. } D(\pi', \pi) \leq C$$

# Equivalent constraint based optimisation

- From the convexity of the $g$-divergence we can equivalently write the optimisation as

$$\tilde{\pi} = \arg\min_{\pi'} \left[ L_x(\pi', x) \right] \quad \text{s.t. } D(\pi', \pi) \leq C$$

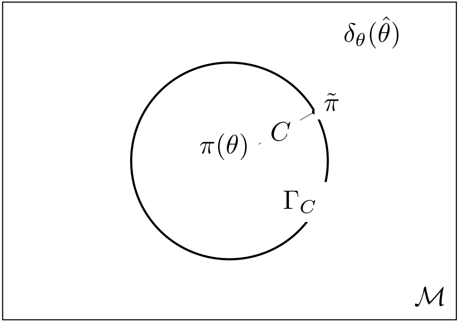



Figure: Graphical representation of solution $\tilde{\pi}$ as the minimiser of $L_x(\pi', x)$ subject to a constraint that $D(\pi', \pi) < C$

# Canonical forms for $l(\theta, x_i)$

- If we have a good proxy model for $F_0$, or if we know we're in $\mathcal{M}$-closed, then the natural choice for $l(\theta, x_i)$ is the self-information loss (negative log-likelihood),

$$l(\theta, x_i) = -\log f(x_i; \theta)$$

  and for $\mathcal{M}$-closed this is the "honest" loss function (proper local scoring rule)

- though equally, say for survival analysis, a partial-likelihood provides a valid update

$$l(\theta, x_i) = -\log g(x_i; \theta)$$

- or for inference on the median of a population

$$l(\theta, x_i) = |\theta - x_i|$$

The key point is that $l(\cdot)$ is targeting $\theta_0$, sufficient for Your decision

# Loss to prior

- The $g$-divergence $D(\pi', \pi) = \int g(\mathrm{d}\pi'/\mathrm{d}\pi)\mathrm{d}\pi'$ provides a large class of loss function; and some special cases include
  - $g(s) = 1 - \sqrt{s}$, the Hellinger divergence, which is equivalent to the L1 metric;
  - $g(s) = s^{-1} - 1$ yields the chi-squared divergence

- For coherency it turns out $D(\pi', \pi)$ must be the Kullback Leibler loss between updated $\pi'$ and the prior ; i.e. $g(s) = -\log s$,

$$L_\pi(\pi', \pi) = KL(\pi', \pi) = \int_\Theta \pi'(\theta) \log \frac{\pi'(\theta)}{\pi(\theta)} d\theta$$

  for proof see Bissiri, Holmes & Walker

- By coherency we mean

$$\pi(\theta) \to \tilde{\pi}(\theta | x_{1:n}) \equiv \pi(\theta) \to \tilde{\pi}(\theta | x_{1:j}) \to \tilde{\pi}(\theta | x_{1:j}, x_{j+1:n})$$

# Updating

- We require $\tilde{\pi}$ to minimize $[L_x(\pi', x) + L_\pi(\pi', \pi)]$,

$$\begin{aligned}
\tilde{\pi} &= \arg\min_{\pi'} [L_x(\pi', x) + L_\pi(\pi', \pi)] \\
&= \text{a.m}\left[\int_\Theta \pi'(\theta) l(\theta, x) d\theta + \int_\Theta \pi' \log \frac{\pi'(\theta)}{\pi(\theta)} d\theta\right] \\
&= \text{a.m.}\left[\int_\Theta \pi'(\theta) \log\left(\frac{\pi'(\theta)}{\pi(\theta)\exp[-l(\theta, x)]}\right) d\theta\right]
\end{aligned}$$

- From which we see that the optimal measure $\tilde{\pi}$ follows

$$\tilde{\pi}(\theta) \propto \arg\min_{\pi'} [KL(\pi', \pi(\theta) \;\times\; \exp[-l(\theta, x)])]$$

# Best beliefs

- Hence under this decision theoretic construction we are led to use

$$\tilde{\pi}(\theta) = \frac{\exp[-l(\theta, x)]\pi(\theta)}{\int_\Theta \exp[-l(\theta, x)]\pi(\theta)d\theta} \quad (1)$$

  as our best updated measure of beliefs for $\theta$
- where $\int_\Theta \exp[-l(\theta, x)]\pi(\theta)d\theta$ is the prior predictive utility of the model $\pi(\theta)$
- We have not had to assume knowledge of $f(x)$, i.e. $\mathcal{M}$-closed, to get here
- The solution coincides with other recent ideas on risk minimisation
  - Gibbs posteriors – (Zhang, 2006)
  - PAC-Bayes – (Langford, 2005)

  although we arrive at (1) through an axiomatic principle of coherency

# Points to Note

- If you really believe your model to be true then you're in $\mathcal{M}$-closed then we are led to use $l(\theta, x_i) = -\log f(x_i; \theta)$ and we recover Bayes Theorem

- So one way to view Bayesian updating is by maximising the posterior predictive log-likelihood

$$\int_\theta \left[ \sum_i \log f(x_i; \theta) \right] \pi'(\theta) \mathrm{d}\theta$$

  Subject to a KL constraint,

$$KL(\pi'(\theta|x), \pi(\theta)) \leq C$$

- However, the update here has been obtained under much weaker conditions – just loss functions and a KL loss on the prior

- In particular, we have treated the prior $\pi$ as just another piece of information; so $\pi$ could be elicited after the data has arrived, or during, or updated based on additional knowledge obtained

# Illustration

- We illustrate the General Bayesian updating for understanding the contribution of genetic variation to risk of colon cancer involving right-censored time-to-event data
- Collaborators at the Wellcome Trust Centre for Human Genetics, University of Oxford, obtained survival times on 918 cancer patients with germline genotype data at 100,000's of markers genome-wide
- For demonstration purposes we only consider one chromosomal previously identified as holding a potential association signal containing 15,608 genotype measurements

# Illustration

- The data table $X$ then has $n = 918$ rows and $p = 15,608$ columns, where $(X)_{ij} \in \{0, 1, 2\}$ denotes the genotype of the $i$'th individual at the $j$'th marker.

- Alongside this we have the corresponding $(n \times 2)$ response table of survival times $Y$ with a column of event-times, $y_{i1} \in \Re^+$ and a column of indicator variables $y_{i2} \in \{0, 1\}$, denoting whether the event is observed or right-censored at $y_{i1}$.

# Full Bayesian Model

- For the full Bayesian model we require a joint model for the data and parameters
- For example, a log-linear proportional hazards model

$$p(y \mid x, \beta) = h_0(y) \prod_i \frac{\exp(x_i \beta)}{\sum_{j \in R_i} \exp(x_j \beta)} \pi(\beta) \pi[h_0(\cdot)]$$

  where $h_0(y)$ is the baseline hazard, assumed a nuisance parameter (process), and $\pi[h_0(\cdot)]$ would usually be a NP measure
- If interest is in $\pi(\beta|x, y)$ then this is obtained from the marginal

$$\pi(\beta|x, y) = \int_{h_0} \pi(\beta, h_0|x, y) \mathrm{d}h_0$$

- But this is challenging as $h_0(y)$ is an infinite dimensional nuisance parameter for the decision

# Use of Bayesian partial loss

- Using our construction we can consider only the conditional order of events as partial-information relevant to the decision, $\beta$, via the cumulative loss function,

$$l(\beta, \mathbf{x}) = \sum_{i=1}^{n} \log \left( \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{l \in R_i} \exp\left(\sum_{j=1}^{p} x_{lj}\beta_j\right)} \right), \tag{2}$$

  where $R_i$ denotes the risk set, those individuals not censored or at time $t_i$, and in this way obtain a conditional distribution $\pi(\beta|\boldsymbol{x})$

- We assume, $\beta_j \sim N(0, v_j)$ and set $v_j = 0.5$ for our study, reflecting beliefs that associated coefficients will be modest; although we note that one advantage of our approach is that subjective prior information can be integrated into the analysis.
  - Note: this is substantive prior knowledge as we know that $||\beta_j||$'s will be small

# General Bayes factors

- To initially explore for evidence of effects; i.e. $\beta_j \neq 0$, we can calculate the general Bayes Factor of association at the $j$ th marker as,

$$BF_j = \frac{\int_{\beta_j} \exp\left[-l(\beta_j|\boldsymbol{x}_j)\right] \pi(\beta_j) \mathrm{d}\beta_j}{\exp\left[-l(\beta_j = 0|\boldsymbol{x}_j)\right]}$$

- This involves a one-dimensional integral via importance sampling for the prior expected loss in using $\beta_j$ on the numerator

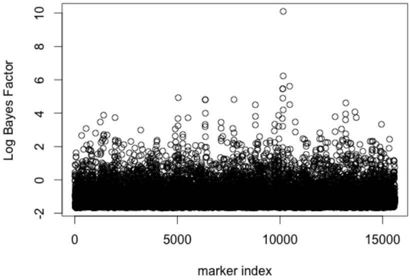



Figure: Log Bayes Factor vrs marker index along chromosome

# Comparing BFs with p-values

- It is interesting to compare the evidence of association provided by the Bayes Factor to that obtained using a conventional Cox PH analysis

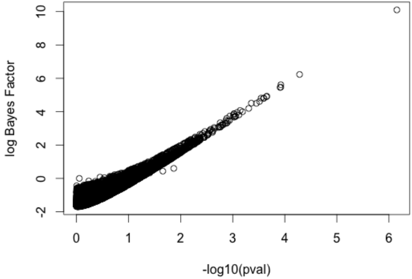



Figure: Log Bayes Factor vrs -log10 p-value of association

- We see general agreement, although interestingly there appears to be greater dispersion at markers of weaker association

# Comparing BFs with p-values

- We colour the points by the standard error of the MLE

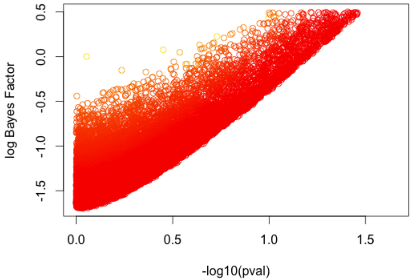



Figure: Log BF vrs -log10 p-value coloured by standard error in MLE

- We can see a tendency for markers with less information, greater standard error, to get attenuated towards a logBF of 0

# Comparing BFs with p-values

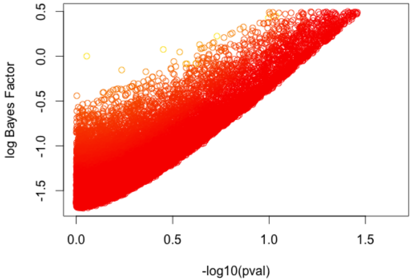



Figure: Log BF vrs -log10 p-value coloured by standard error in MLE

- High standard errors relate to genotypes of rarer alleles and the attenuation reflects a greater degree of uncertainty for association at these markers that contain less information; whereas the p-value is uniform under the null no matter what the power is in the alternative

# Multivariate variable selection

- We can explore the uncertainty in the multiple regression model via the cumulative loss function

$$l(\beta, \mathbf{x}) = \sum_{i=1}^{n} \log \left( \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{l \in R_i} \exp\left(\sum_{j=1}^{p} x_{lj}\beta_j\right)} \right),$$

- We assume proper priors, $\pi(\beta)$ on the regression coefficient,

$$\pi(\beta_j) = \begin{cases} 0 & \text{if } \delta_j = 0 \\ \mathsf{N}(0, v_j) & \text{otherwise,} \end{cases}$$

  where $\delta_j \in \{0, 1\}$ is an indicator variable selection on covariate relevance with, $\pi(\delta_j) = \mathsf{Bin}(a_j)$

- In this way the joint marginal posterior $\pi(\delta|\boldsymbol{x})$ quantifies beliefs about which variables are important to the regression

# Prior-predictive Utility

- As we are using the partial-loss (likelihood) model we have

$$\pi(\delta|x) = \left[\int_\beta \exp[-l(\beta,\delta,\theta)]\pi(\beta|\delta)d\beta\right]\pi(\delta)$$

  where the first term is the marginal partial-loss or prior-predictive utility

- We can implement a MCMC algorithm for this General Bayesian model (with efficient independence proposal densities) without specifying a full probability model

# Posterior probability of marker inclusion

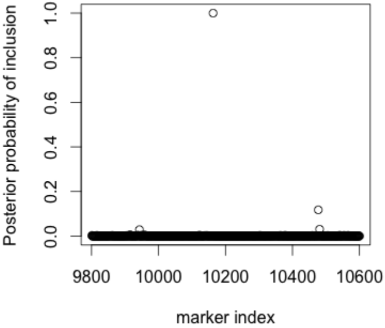



Figure: Posterior marginal inclusion probability from multiple marker model

- The model suggest overwhelming evidence for a single marker in the region of index 10200 but also weaker evidence of independent signal in a couple of other regions

# Current work / Open Questions

- We have a constructive, decision theoretic, approach to coherent Bayesian updating in the absence of a true model
  - this is not an approximation but a valid representation of beliefs
- This allows the modeller to concentrate on those aspects important to the decision
- The method has clear connections with penalised log-likelihood (c.f. Lasso, splines etc) but here for penalised probability measures
  - We are selecting $\widehat{\pi(\theta)}$ rather than $\hat{\theta}$
- Interpretation of the normalising constant $\int_\Theta \exp[-l(\theta, x)]\pi(\theta)d\theta$ which arises in model-choice $\pi(M_i)$ for models $M \in \{M_1, \ldots, M_k\}$ as,

$$L(M_i, x) = \int_\Theta \exp[-l(\theta, x)]\pi_{M_i}(\theta)d\theta$$

- But in general $\int_x \exp[-l(\theta, x)]dx \neq 1$
- Do we obtain the same parsimony as for $\mathcal{M}$-closed Bayes Factors? Does it make sense to consider normalised relative loss and impose this constraint?