Cutting-edge issues in objective Bayesian model comparison

Subjective views of a Bayesian barber

Luca La Rocca
luca.larocca@unimore.it

Dipartimento di Scienze Fisiche, Informatiche e Matematiche
Università di Modena e Reggio Emilia (UNIMORE)

Big data in biomedicine. Big models?
Centre for Research in Statistical Methodology (CRiSM)
University of Warwick, 27 February 2014

Outline

Bayes factors

Two sampling models for a sequence of discrete observations z^n,

$$\begin{aligned}\mathcal{M}_0^n &= \{f_0^n(\cdot|\gamma_0),\ \gamma_0 \in \Gamma_0\},\\ \mathcal{M}_1^n &= \{f_1^n(\cdot|\gamma_1),\ \gamma_1 \in \Gamma_1\},\end{aligned}$$

compared by means of the Bayes factor for $\mathcal{M}_1^n$ against $\mathcal{M}_0^n$,

$$\mathrm{BF}_{10}(z^n) = \frac{\int_{\Gamma_1} f_1^n(z^n|\gamma_1)p_1(\gamma_1)d\gamma_1}{\int_{\Gamma_0} f_0^n(z^n|\gamma_0)p_0(\gamma_0)d\gamma_0},$$

where $p_i(\gamma_i)$ is a parameter prior under $\mathcal{M}_i^n$ $(i = 0, 1)$; then

$$\Pr(\mathcal{M}_1^n|z^n) = \frac{\mathrm{BF}_{10}(z^n)}{1 + \mathrm{BF}_{10}(z^n)},$$

assuming $\Pr(\mathcal{M}_0^n) = \Pr(\mathcal{M}_1^n) = 1/2$, where $z^n = (z_1, \ldots, z_n)$.

Asymptotic learning rate

Let $\mathcal{M}_0^n$ be nested in $\mathcal{M}_1^n$ ($\Gamma_0 \equiv \tilde{\Gamma}_0 \subset \Gamma_1$) with dimensions $d_0 < d_1$.

Assume $p_0(\cdot)$ is a local prior (continuous and strictly positive on Γ_0).

Typically $p_1(\cdot)$ is also a local prior, so that (under regularity conditions)

$$\mathrm{BF}_{10}(z^n) = n^{-\frac{(d_1-d_0)}{2}} e^{O_P(1)},$$

as $n \to \infty$, if the sampling distribution of z^n belongs to $\mathcal{M}_0^n$,

$$\mathrm{BF}_{10}(z^n) = e^{Kn+O_P(n^{1/2})},$$

for some $K > 0$, if the sampling distribution of z^n belongs to $\mathcal{M}_1^n \setminus \mathcal{M}_0^n$.

This imbalance in the asymptotic learing rate motivated the introduction of non-local priors[1]...

[1] Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. J. R. Stat. Soc. Ser. B Stat. Methodol. 72, 143–170.

Generalized moment priors

...such as generalized moment priors[2] of order h:

$$p_1^M(\gamma_1|h) \propto g_h(\gamma_1)p_1(\gamma_1), \qquad \gamma_1 \in \Gamma_1,$$

where $g_h(\cdot)$ is a smooth function from Γ_1 to $\Re_+$,
vanishing on $\tilde{\Gamma}_0$ together with its first $2h-1$ derivatives,
while $g_h^{(2h)}(\gamma_1) > 0$ for all $\gamma_1 \in \tilde{\Gamma}_0$; let $g_0(\gamma_1) \equiv 1$.

Asymptotic learning rate changed to

$$\mathrm{BF}_{10}(z^n) = n^{-h-\frac{(d_1-d_0)}{2}} e^{O_P(1)},$$

as $n \to \infty$, if the sampling distribution of z^n belongs to $\mathcal{M}_0^n$;
unchanged if the sampling distribution of z^n belongs to $\mathcal{M}_1^n \setminus \mathcal{M}_0^n$.

[2] Consonni, G. , Forster, J. J. and La Rocca, L. (2013). The whetstone and the alum block: Balanced objective Bayesian comparison of nested models for discrete data. Statist. Sci. 38, 398–423.

Comparing two proportions

Let the larger model be the product of two binomial models,

$$f_1^{n_1+n_2}(y_1, y_2|\theta_1, \theta_2) = \text{Bin}(y_1|n_1, \theta_1)\text{Bin}(y_2|n_2, \theta_2), \qquad (\theta_1, \theta_2) \in]0, 1[^2,$$

and the null model assume $\theta_1 = \theta_2 = \theta$,

$$f_0^{n_1+n_2}(y_1, y_2|\theta) = \text{Bin}(y_1|n_1, \theta)\text{Bin}(y_2|n_2, \theta), \qquad \theta \in]0, 1[.$$

Starting from the conjugate local prior

$$p_1(\theta_1, \theta_2|a) = \text{Beta}(\theta_1|a_{11}, a_{12})\text{Beta}(\theta_2|a_{21}, a_{22}),$$

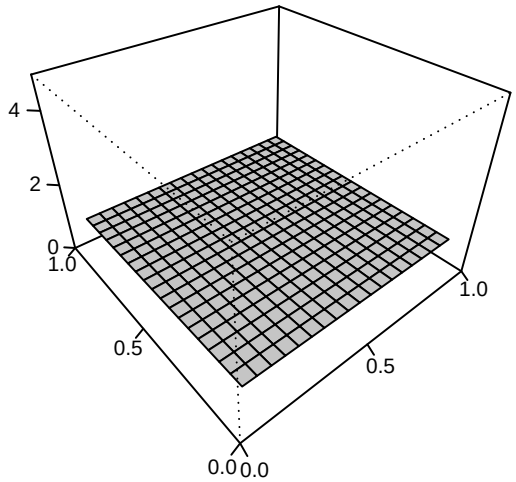
under $\mathcal{M}_1$, where a is 2×2 matrix of strictly positive real numbers, define the conjugate moment prior of order h as

$$p_1^M(\theta_1, \theta_2|a, h) \propto (\theta_1 - \theta_2)^{2h}\text{Beta}(\theta_1|a_{11}, a_{12})\text{Beta}(\theta_2|a_{21}, a_{22});$$

assume $a_{11} = a_{12} = b_1$ and $a_{21} = a_{22} = b_2$.

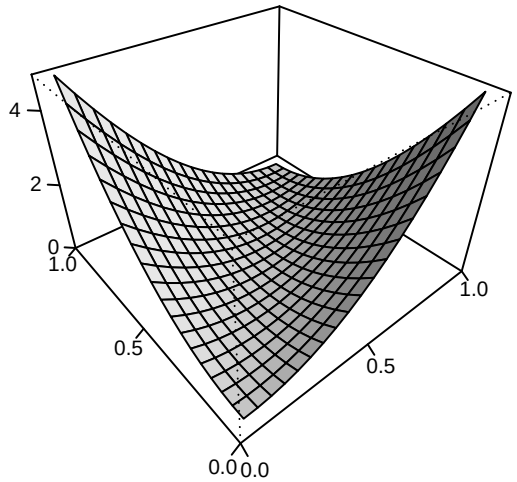
Going non-local from a default prior

h = 0



$b_1 = b_2 = 1$

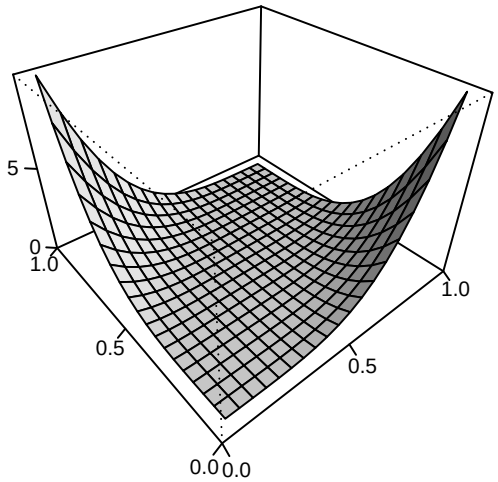
h = 1



$b_1 = b_2 = 1$

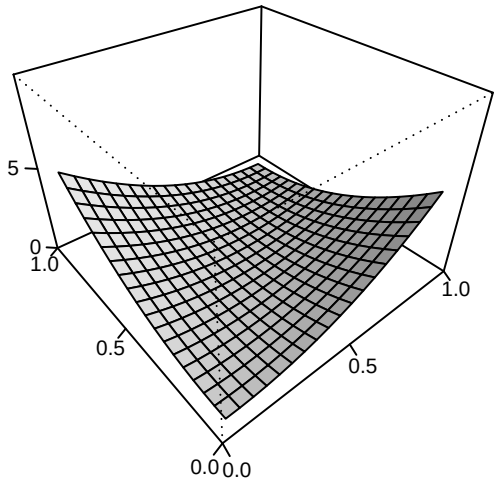
Increasing the order of a default moment prior

h = 2



$b_1 = b_2 = 1$

h = 1



$b_1 = b_2 = 1$

Jeffreys-Lindley-Bartlett paradox

Related to the limiting argument of the JLB paradox, the idea that probability mass should not be "wasted" in parameter areas too remote from the null is both old and new:

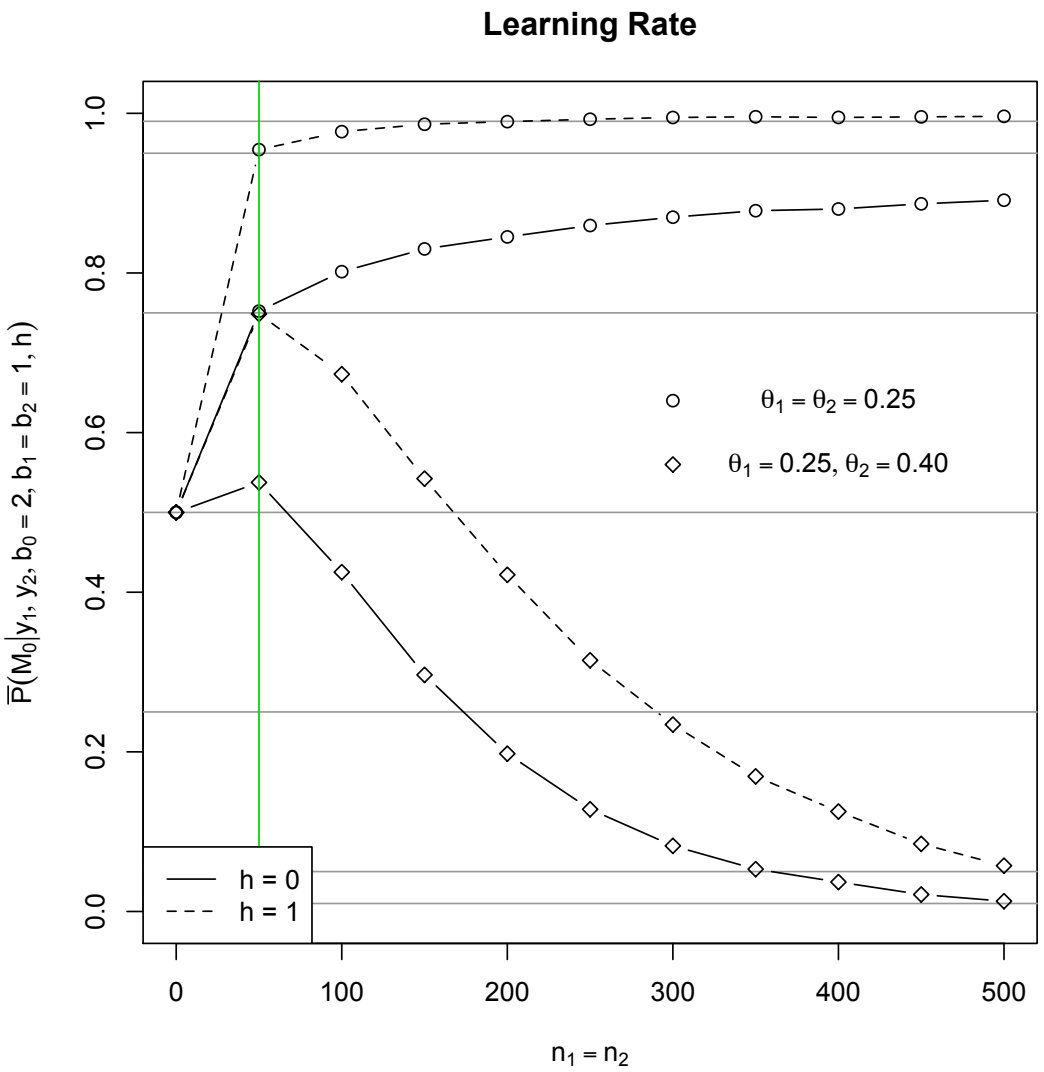
> *If a rare event for H_0 occurs that also is rare for typical H_1 values, it provides little evidence for rejecting H_0 in favor of H_1*[3]

> *A vague prior distribution assigns much of its probability on values that are never going to be plausible, and this disturbs the posterior probabilities more than we tend to expect—something that we probably do not think about enough in our routine applications of standard statistical methods*[4]

[3] Morris, C. N. (1987). Comments on "Testing a point null hypothesis: The irreconcilability of P values and evidence." J. Amer. Statist. Assoc. 82, 131–133.

[4] Gelman, A. (2013). P values and statistical practice. Epidemiology 24, 69–72.

Were you wearing a red tie, Sir?



Using a default local prior for θ under $\mathcal{M}_0$:

$$p_0(\theta|b_0) = \text{Beta}(\theta|b_0, b_0).$$

Vertical line at $n = 50$.

Intrinsic moment priors

Mixing[5] over all possible training samples $x^t = (x_1, \ldots, x_t)$ of size t, the intrinsic moment prior on γ_1 is given by

$$p_1^{IM}(\gamma_1|h,t) = \sum_{x^t} p_1^M(\gamma_1|x^t,h) m_0(x^t), \qquad \gamma_1 \in \Gamma_1,$$

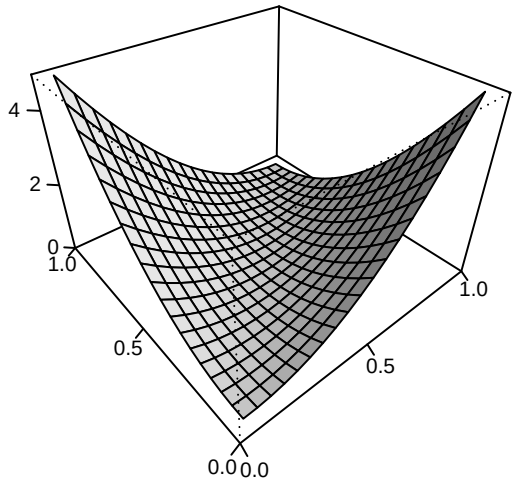
where $p_1^M(\cdot|x^t,h)$ is the posterior of γ_1 under $\mathcal{M}_1$, given x^t, and $m_0(x^t) = \int_{\Gamma_0} f_0^t(x^t|\gamma_0) p_0(\gamma_0) \mathrm{d}\gamma_0$ is the marginal of x^t under $\mathcal{M}_0$; let $p_1^{IM}(\cdot|h,0) = p_1^M(\cdot|h)$.

As the training sample size t grows, the intrinsic moment prior increases its concentration on regions around the subspace $\tilde{\Gamma}_0$, while the non-local nature of $p_1^M(\cdot|h)$ is preserved.

[5] Pérez, J. M. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. Biometrika 89, 491–511.

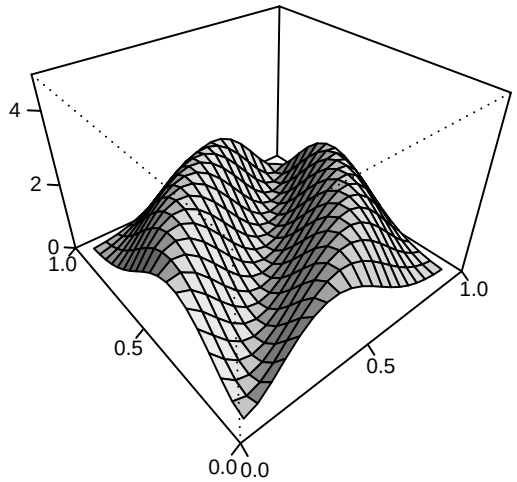
Pulling the mass back toward the null

h = 1, t = (0,0)



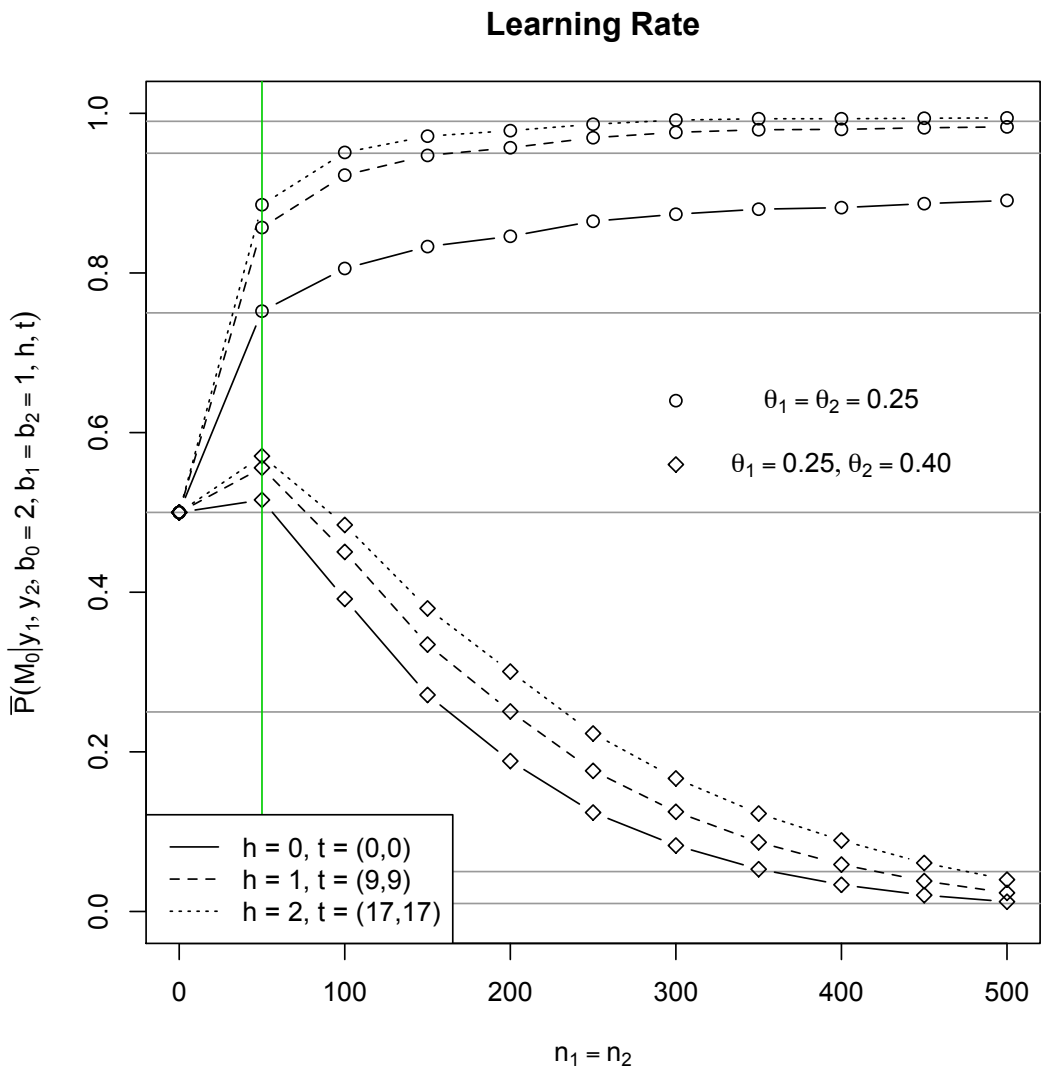
$b_0 = 2,\ b_1 = b_2 = 1$

h = 1, t = (9,9)



$b_0 = 2,\ b_1 = b_2 = 1$

Bleeding stopped

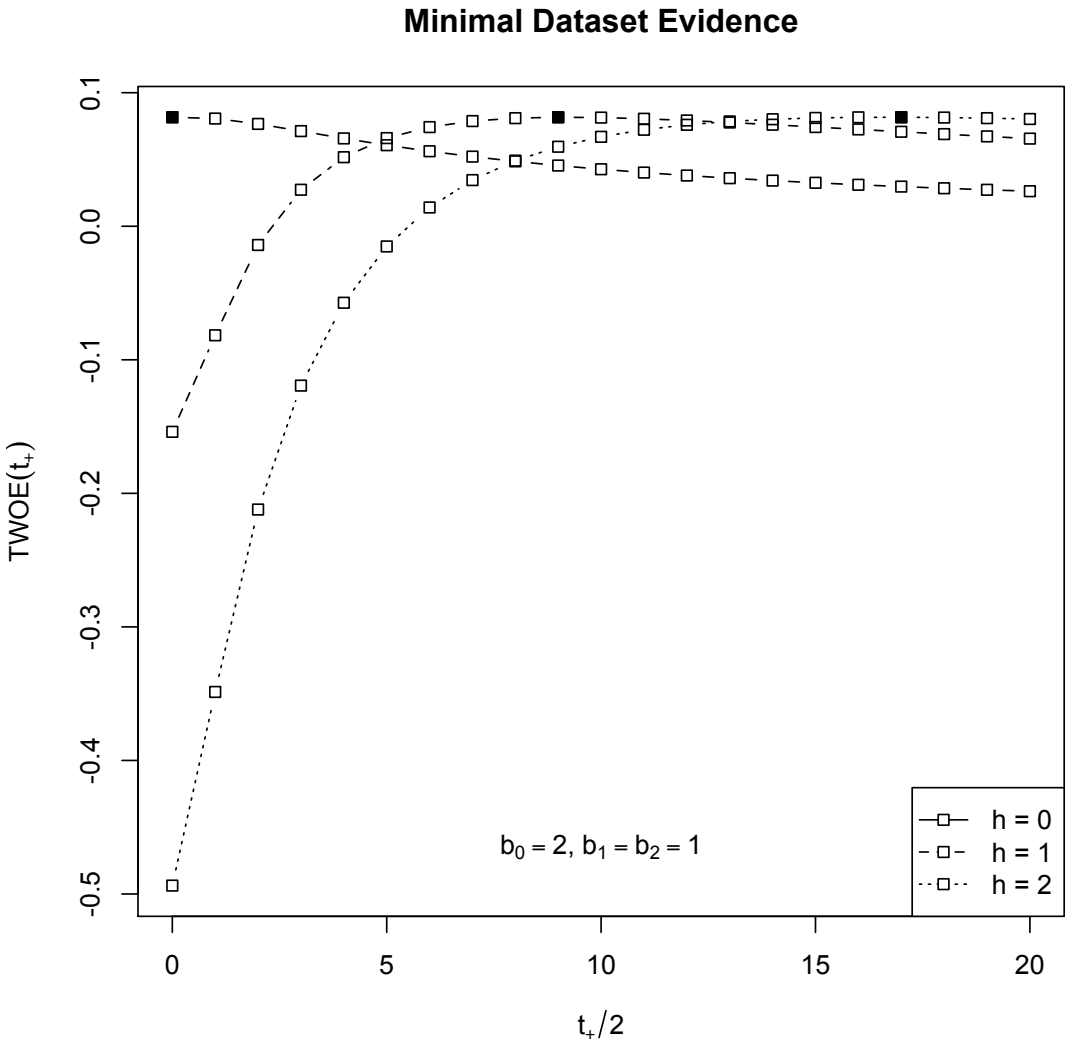


Total training sample size for the comparison of two proportions:

$$t_+ = t_1 + t_2.$$

Vertical line at $n = 50$.

How was t chosen?



Total Weight Of Evidence:

$$\sum_{z^m} \log BF_{10}^{IM}(z^m | b, h, t)$$

summing over all possible observations z^m with minimal sample size m such that data can discriminate between $\mathcal{M}_0$ and $\mathcal{M}_1$.

Here $m = (1, 1)$.

How about the choice of h?

Choice $h = 1$ recommended, based on the following considerations:

- switching from $h = 0$ to $h = 1$ changes the asymptotic learning rate from sublinear to superlinear (making a big difference);
- switching from $h = 1$ to $h = 2$ results in a less remarkable difference (while aggravating the problem with small samples).

Inverse moment priors (Johnson and Rossell, 2010, JRSS-B) achieve an exponential learning rate also when the sampling distribution belongs to the smaller model; do you really want to drop that fast a model with the sampling distribution on its boundary?

Predictive performance

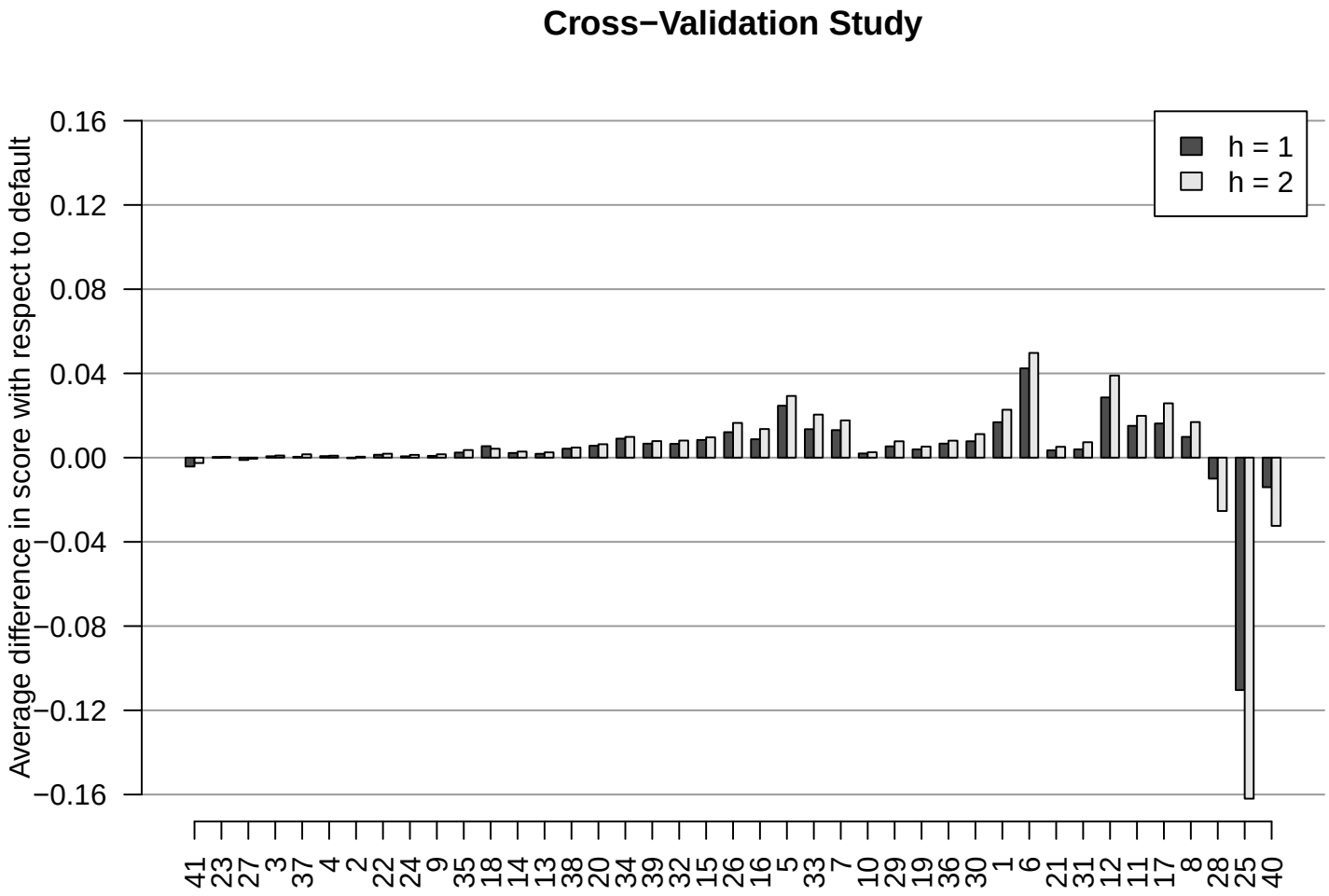


Table (Efron, 1996)

Computational burden

The Bayes factor against $\mathcal{M}_0^n$ using a generalized moment prior under $\mathcal{M}_1^n$ can be written as

$$BF_{10}^M(z^n|h) = \frac{\int_{\Gamma_1} g_h(\gamma_1)p_1(\gamma_1|z^n)d\gamma_1}{\int_{\Gamma_1} g_h(\gamma_1)p_1(\gamma_1)d\gamma_1} BF_{10}(z^n),$$

so that the extra effort required amounts to computing some generalized moments of the local prior and posterior.

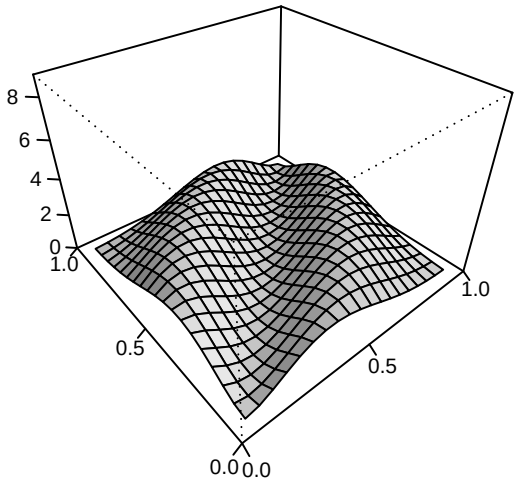
The Bayes factor against $\mathcal{M}_0^n$ using an intrinsic moment prior under $\mathcal{M}_1^n$ can be written as a mixture of conditional Bayes factors:

$$BF_{10}^{IM}(z^n|h,t) = \sum_{x^t} BF_{10}^M(z^n|x^t,h)m_0(x^t),$$

where $BF_{10}^M(z^t|x^t,h)$ is the Bayes factor using $p_1^M(\cdot|x^t,h)$ as prior under $\mathcal{M}_1^n$; recall that $m_0(x^t) = \int_{\Gamma_0} f_0^t(x^t|\gamma_0)p_0(\gamma_0)\mathrm{d}\gamma_0$.

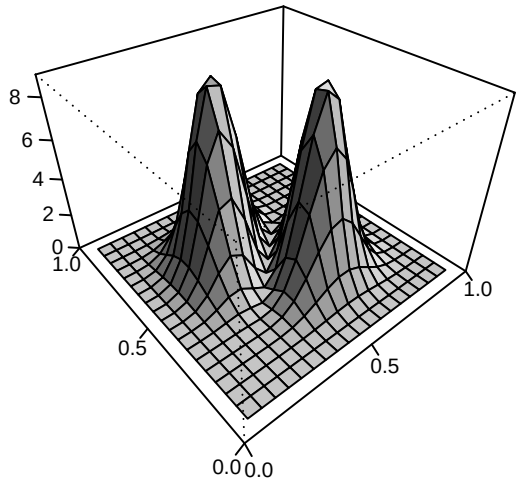
Why not just increase prior sample size?

h = 1, t = (9,9)



$b_0 = 2,\ b_1 = b_2 = 1$

h = 1, t = (0,0)



$b_0 = 2,\ b_1 = b_2 = 9$

Logistic regression models

Suppose we observe $y = (y_1, \ldots, y_N)$ with $f(y_i|\theta_i) = \text{Bin}(y_i|n_i, \theta_i)$, $i = 1, \ldots, N$, and we let

$$\log \frac{\theta_i}{1-\theta_i} = \beta_0 + \sum_{j=1}^{k} w_{ij}\beta_j, \qquad i = 1, \ldots, N,$$

where w_{ij}, $j = 1, \ldots, k$, are the values of k explanatory variables observed with y_i; the likelihood is $f_k^{n+}(y|\beta) = \left\{\prod_{i=1}^{N} \binom{n_i}{y_i}\right\} L_k(\beta|y, n)$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$, $n = (n_1, \ldots, n_N)$, and

$$L_k(\beta|y, n) = \prod_{i=1}^{N} e^{y_i\left(\beta_0 + \sum_{j=1}^{k} w_{ij}\beta_j\right) - n_i \log\left(1 + \exp\left\{\beta_0 + \sum_{j=1}^{k} w_{ij}\beta_j\right\}\right)}.$$

Special cases: $N = 2$, $k = 1$, $w_{ij} = (i-1)$ & $N = 2$, $k = 0$

Logistic regression priors

Conjugate local prior[6] given by $p_k^C(\beta|u,v) \propto L_k(\beta|u,v)$,
where $u = (u_1, \ldots, u_N)$ and $v = (v_1, \ldots, v_N)$;
default specification of these hyperparameters:

$$v_i = v_+ \frac{n_i}{n_+}, \quad u_i = \frac{v_i}{2}, \qquad i = 1, \ldots, N,$$

for some $v_+ > 0$ representing a prior sample size;
the condition $u_i = v_i/2$ ensures that the prior mode is at $\beta = 0$.

In the special cases corresponding to comparing two proportions, the induced prior on the common proportion is $\theta \sim \text{Beta}(v_+/2, v_+/2)$ while $(\theta_1, \theta_2) \sim \text{Beta}(v_1/2, v_1/2) \otimes \text{Beta}(v_2/2, v_2/2) \ldots$

[6] Bedrick, E. J., Christensen, R. and Johnson, W. (1996). A new perspective on priors for generalized linear models. J. Amer. Statist. Assoc. 91, 1450–1460.

Going non-local in a different parameterization

...whereas the product moment prior[7] on β

$$p_k^M(\beta|u,v,h) \propto \prod_{j=1}^{k} \beta_j^{2h} p_k^C(\beta|u,v),$$

in the special case $N = 2$, $k = 1$, $w_{ij} = (i-1)$,
induces on (θ_1, θ_2) a prior not in the family of conjugate moment priors considered before; how much can results differ?

Intrinsic procedure successfully applied to $p_k^M(\cdot|u,v,h)$,
but maybe increasing v_+ is an interesting alternative?

[7] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. J. Amer. Statist. Assoc. 107, 649–660.

Thank you!

http://xianblog.wordpress.com/2013/08/01/whetstone-and-alum-and-occams-razor/