UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Bayesian Statistics

Model comparisons

Leonardo Egidi

A.A. 2018/19

Motivations

Bayesian models can be evaluated and compared in several ways. Most simply, any model or set of models can be taken as an exhaustive set, in which case all inference is summarized by the posterior distribution.

The fit of model to data can be assessed using posterior predictive checks, prior predictive checks or, more generally, mixed checks for hierarchical models.

However, we may need a pure comparison set of tools: when several candidate models are available, they can be compared and averaged using:

- Bayes factors (which is equivalent to embedding them in a larger discrete model)
- Predictive information criteria (AIC, DIC, BIC, WAIC,...)

Indice

Hypothesis testing

In the classical framework we compare two alternatives:

$$H_0 : \theta \in \Theta_0; H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 form a partition of the parameter space.

- In the classical approach, data are used to verify whether they are compatible with the null hypothesis through the calculation of the *p*-value, that is the probability that, under the null hypothesis, we may observe a sample which would give a result which is even *less convincing* under the null hypothesis (compared with the one we observe).

Hypothesis testing

- In a Bayesian framework, the most natural way to proceed is to quantify the posterior weights of the two competing hypootheses, that is, $\Pr(H_0|y)$, defined as:

$$\Pr(H_0|y) = \Pr(\Theta_0|y) = \int_{\Theta_0} \pi(\theta|y)d\theta,$$

where we assume that the random variable Θ is continuous.

The simplest case: Two simple hypotheses

This is the most elementary case (enough to present the irreconcilability)

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1$$

We may elicit the following prior distributions:

$$\pi_0 = \Pr(H_0), \quad \pi_1 = \Pr(H_1) = 1 - \pi_0.$$

Likelihood:

$$p(y|\theta_0), \quad p(y|\theta_1)$$

The relative weights of the two hypotheses is then given by the ratio:

$$\frac{\Pr(H_0|y)}{\Pr(H_1|y)} = \frac{\pi(\theta_0|y)}{\pi(\theta_1|y)} = \frac{\pi_0}{\pi_1}\frac{p(y|\theta_0)}{p(y|\theta_1)} \tag{1}$$

The simplest case: Two simple hypotheses

The posterior odds are the product of two terms:

- π_0/π_1 is the relative weight of the two hypotheses before observing the data.
- the second factor is usually denoted as:

$$BF_{01} = \frac{p(y|\theta_0)}{p(y|\theta_1)} \tag{2}$$

and is called Bayes factor: it represents the multiplicative factor which transforms the prior odds into the posterior odds. BF_{01} represents a measure of evidence in favour of H_0, where $BF_{01} > 1 (< 1)$ indicates that data favour H_0 (H_1).

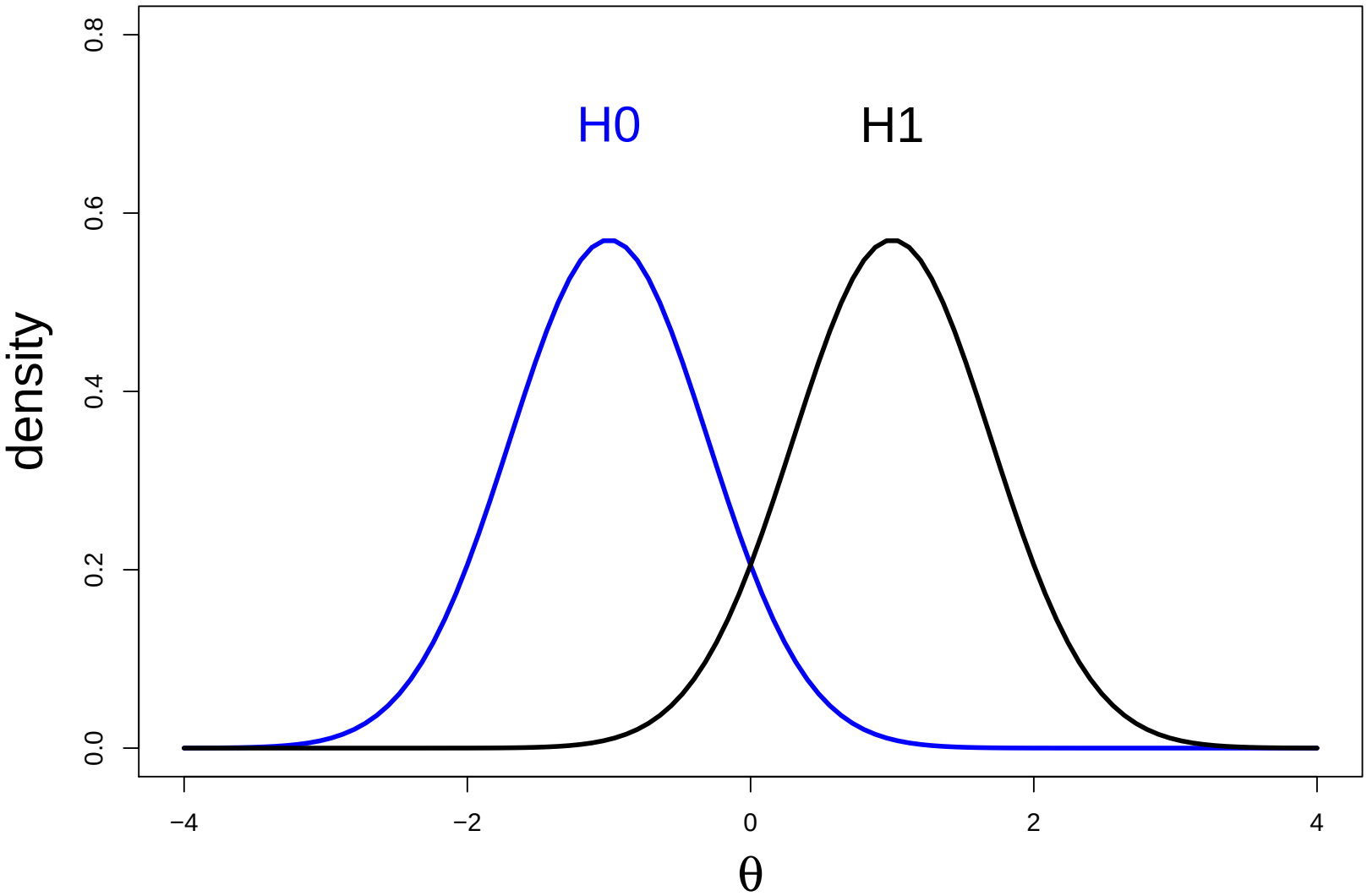
Toy example

Observe $\bar{y} \sim \mathcal{N}(\theta, \sigma^2)$, with $\sigma = 0.7$, we want to test:

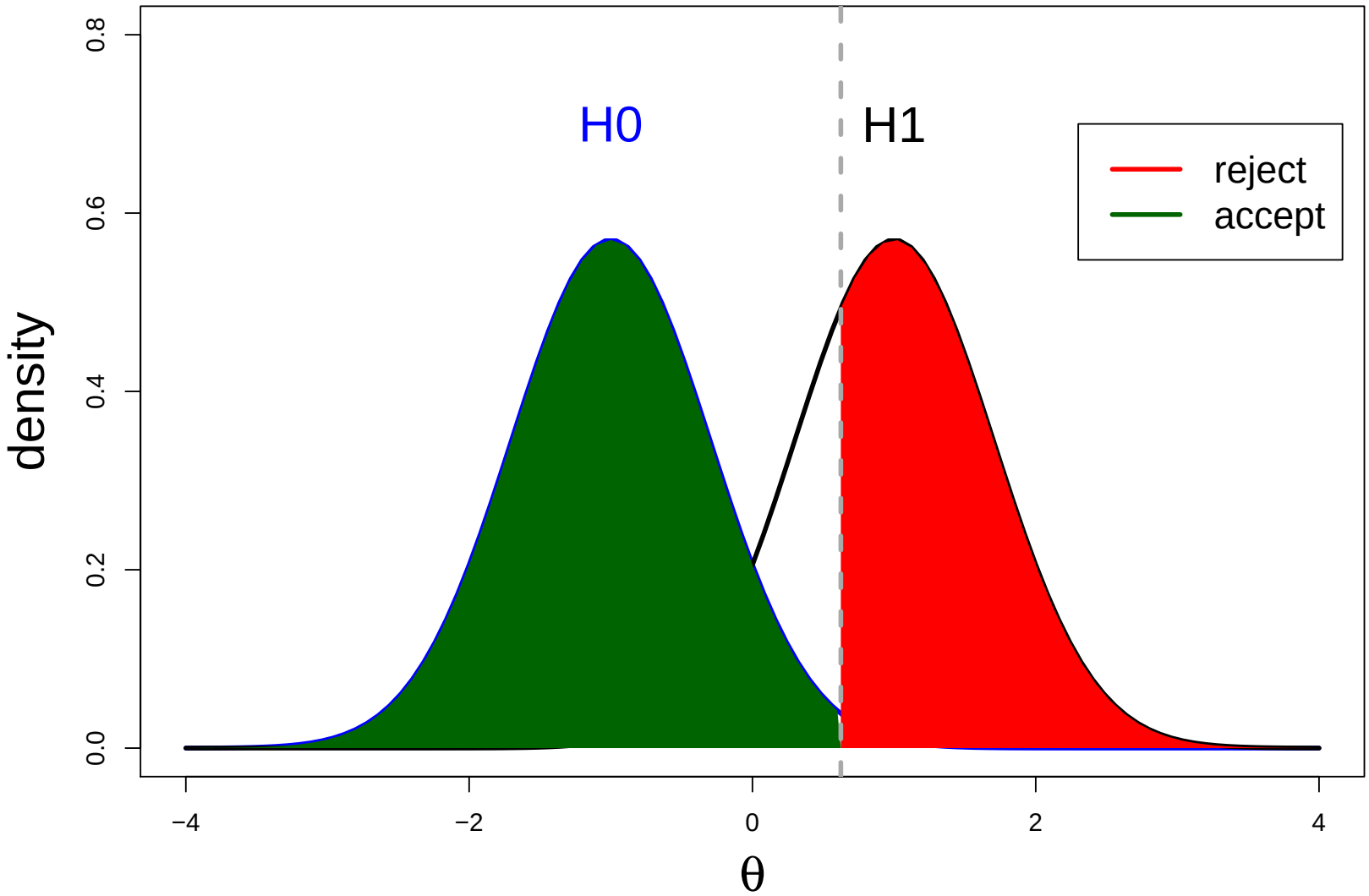
$$H_0 : \theta = \theta_0 = -1, \quad \text{vs.} \quad H_1 : \theta = \theta_1 = 1.$$

- Frequentist approach: fix $\alpha = 0.01$ and calculate the rejection region as $\frac{\bar{y}-\theta_0}{\sigma} > 2.32$, that is $\bar{y} > 0.624$.
- Bayesian approach: the data information is summarized by the Bayes factor $BF_{01} = p(y|\theta_0)/p(y|\theta_1)$.

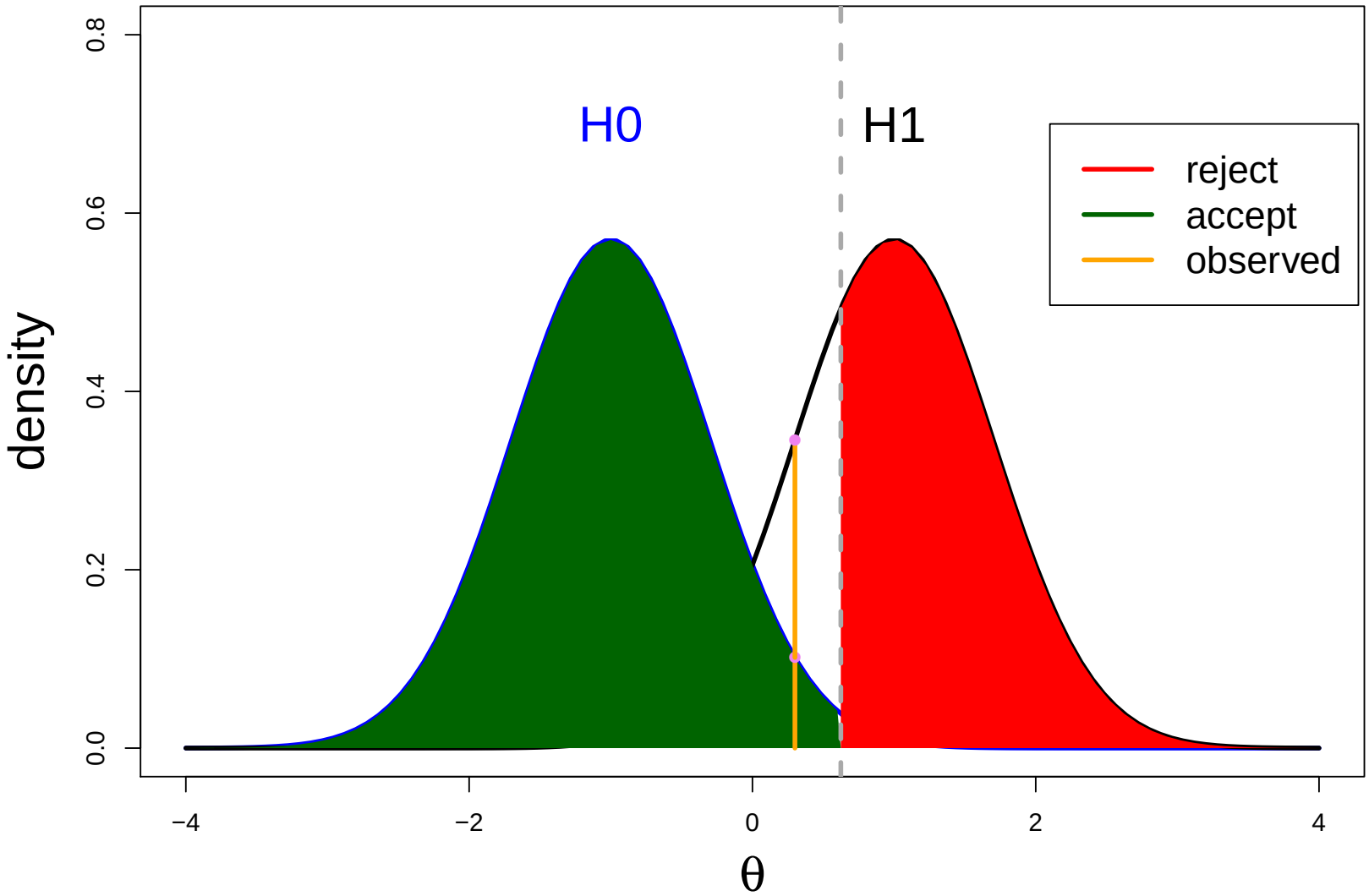
Toy example



Toy example: classical testing



Toy example: classical testing. If $\bar{y} = 0.3$...



Toy example

Comments:

- Frequentist approach tends to accept H_0 even for values which are more likely under H_1 (see what happens between 0 and 0.624...).
- Bayesian approach tends to accept H_0 only if the posterior probability of H_0 is higher than the posterior probability of H_1. If $\pi_0 = \pi_1 = 1/2$, H_0 is 'accepted' if the marginal likelihood under H_0 is higher than the marginal likelihood under H_1.

The general case

$$H_0 : \theta \in \Theta_0; \quad \text{vs.} \quad H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 form a partition of the parameter space.
Prior: two steps

1. Prior probabilities:

$$\pi_0 = \Pr(\Theta_0), \quad \pi_1 = \Pr(\Theta_1).$$

2. For H_i, $i = 0, 1$, $g_i(\theta)$ is the prior density of θ. Then the global prior is:

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \theta \in \Theta_0 \\ (1 - \pi_0) g_1(\theta) & \theta \in \Theta_1 \end{cases}$$

The general case

The beliefs about the two hypotheses are summarized by the posterior odds ratio:

$$\frac{\Pr(\theta \in \Theta_0|y)}{\Pr(\theta \in \Theta_1|y)} = \frac{\int_{\Theta_0} \pi(\theta|y)d\theta}{\int_{\Theta_1} \pi(\theta|y)d\theta} = \frac{\pi_0}{1-\pi_0} \frac{\int_{\Theta_0} p(y|\theta)g_0(\theta)d\theta}{\int_{\Theta_1} p(y|\theta)g_1(\theta)d\theta}.$$

- The Bayes factor $BF_{01} = \frac{\int_{\Theta_0} p(y|\theta)g_0(\theta)d\theta}{\int_{\Theta_1} p(y|\theta)g_1(\theta)d\theta}$ is a ratio between marginal likelihoods. Their evaluation is central in testing hypotheses and model selection!
- Prior information plays a minor role through the densities g_0 and g_1.

Bayes factors and improper priors

Bayes factor cannot be used with improper priors. This happens because the marginal distribution of the data is not well defined.

Example Suppose an *iid* sample $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma_0^2)$, with σ_0^2 known and $\pi(\mu) = c$. Then:

$$p(y) = \int_\mu p(y|\mu)\pi(\mu)d\mu = c \times constant.$$

If $H_0 : \mu = \mu_0 = 0$ and $H_1 : \mu = \mu_1 \neq 0$, the Bayes factor will depend on c! In fact:

$$BF_{01} = \frac{p(y|\theta_0)}{c \int_{\theta \neq \theta_0} p(y|\theta)d\theta}$$

This has caused a great effort in producing new methods for proper priors for testing.

Jeffreys-Lindley's paradox

Jeffreys-Lindley's paradox describes a counterintuitive situation in which the Bayesian and frequentist approaches to a hypothesis testing problem give opposite results for certain choices of the prior distribution, which favor H_0 weakly.

The paradox occurs when: the frequentist test indicates sufficient evidence to reject H_0, say, at the 5% level, and $P(H_0|y)$ is high, say, 95%, indicating strong evidence that H_0 is in fact true.

Jeffreys-Lindley's paradox

Let us assume again $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known, and let $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$. Let assume as a prior for $\pi(\theta_1)$ under H_1 a normal $\mathcal{N}(\theta_0, \tau^2)$ with large τ (or any other flat enough or vague or diffuse prior).

It can be shown that:

$$BF_{01} = \frac{\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp(-n(\bar{y} - \theta_0)^2/2\sigma^2)}{\frac{\sqrt{n}}{\sqrt{\sigma^2+n\tau^2}\sqrt{2\pi}} \exp(-n(\bar{y} - \theta_0)^2/2(\sigma^2 + n\tau^2))}.$$

Let $u = \sqrt{n}(\bar{y} - \theta_0)/\sigma$ and let $\rho = \tau^2/\sigma^2$. Then:

$$BF_{01} = \sqrt{1 + n\rho^2} \exp(-\frac{u^2}{2} \frac{n\rho^2}{1 + n\rho^2})$$

In classic approach, the more data I observe, the most probable is to reject H_0

For $n \to \infty$, the BF indicates strong evidence for H_0.

Testing and model selection with Bayes factors

Rather than verifying a particular value for a parameter, we are often more interested in assessing some model comparisons. In such a viewpoint, distinct models represent distinct hypotheses, and data y are assumed to have arisen from one of several models:

$$
\begin{aligned}
M_1 &: y \sim p_1(y|\theta_1) \\
M_2 &: y \sim p_2(y|\theta_2) \\
&\dots \\
M_j &: y \sim p_j(y|\theta_j) \\
&\dots \\
M_q &: y \sim p_q(y|\theta_q)
\end{aligned}
$$

Assign prior probabilities, $\Pr(M_j)$ to each model

Indice

Testing and model selection with Bayes factors

Under model M_j:

- Prior density of θ_j: $\pi_j(\theta_j)$.
- Marginal density of y:

$$p_j(y) = \int p_j(y|\theta_j)\pi_j(\theta_j)d\theta_j,$$

  which measures *how likely* is y under model M_j.
- Posterior density:

$$\pi_j(\theta_j|y) = \frac{\pi_j(\theta_j)p_j(y|\theta_j)}{p_j(y)}$$

  .
- Bayes factor of M_j to M_i is defined as the ratio between posterior odds and prior odds for the two competing models:

$$BF_{ji} = p_j(y)/p_i(y). \quad (3)$$

Testing and model selection with Bayes factors

Posterior probability of a model:

$$\Pr(M_j|y) = \frac{\Pr(M_j)p_j(y)}{\sum_{k=1}^{q}\Pr(M_k)p_k(y)} = \left[\sum_{k=1}^{q}\frac{\Pr(M_k)}{\Pr(M_j)}BF_{kj}\right]^{-1} \quad (4)$$

If $\Pr(M_j) = 1/q$,

$$\Pr(M_j|y) = \bar{p}_j = \frac{p_j(y)}{\sum_{k=1}^{q}p_k(y)} = \left[\sum_{k=1}^{q}BF_{kj}\right]^{-1}$$

Reporting: it is useful to separately report the $\bar{p}_j(y)$'s and the $\Pr(M_j)$'s. Knowing the $\bar{p}_j(y)$'s allows computation of the posterior probabilities for any prior probabilities.

Bayes factors

Posterior odds of model M_j relative to model M_k:

$$\underbrace{\frac{\Pr(M_j|y)}{\Pr(M_k|y)}}_{Posterior\ \ odds} = \underbrace{\frac{\Pr(M_j)}{\Pr(M_k)}}_{Prior\ \ odds} \times \underbrace{\frac{p(y|M_j)}{p(y|M_k)}}_{Bayes\ \ factor} \tag{5}$$

The Bayes Factor is the weighted likelihood ratio of M_j relative to M_k or a ratio of marginal (wrt the prior) likelihoods.

Jeffreys (1961) recommends the use of the following rule of thumb to decide between models M_j and M_k:

$BF_{jk} > 100$ decisive evidence against M_k; $10 < BF_{jk} \leq 100$ strong evidence against M_k; $3 < BF_{jk} \leq 10$ substantial evidence against M_k.

though very subjective

Example: soccer goals models.

Major League Soccer model

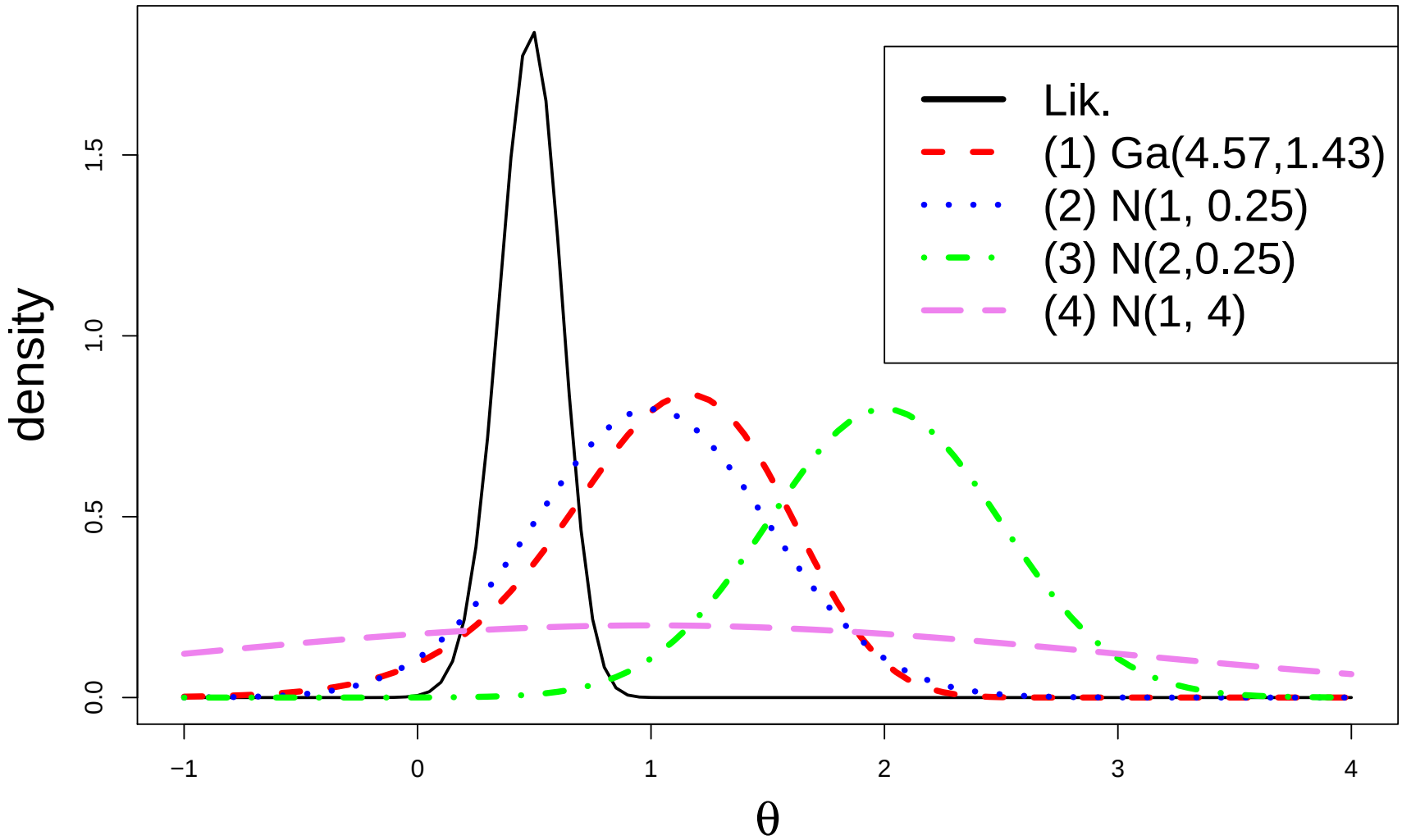
Suppose we are interested in assessing the average number of goals scored by a given team in Major League Soccer, and denote the goals for n games as $y_1, \ldots, y_n$. Since goals are relatively rare events, we assume that:

$$y_i \sim \text{Poisson}(\lambda), \ \ i = 1, \ldots, n.$$

Four different priors, then, four possible models:

1. $\lambda \sim \text{Gamma}(4.57, 1.43)$
2. $\log(\lambda) \sim \mathcal{N}(1, .5^2)$
3. $\log(\lambda) \sim \mathcal{N}(2, .5^2)$
4. $\log(\lambda) \sim \mathcal{N}(1, 2^2)$

Example: soccer goals models. Priors



Example: soccer goals

Once we know how to compute marginal likelihoods, we could compute the Bayes factors for each pair of models. The Bayes factor in support of Prior 1 over Prior 2 is:

$$BF_{12} = \frac{p_1(y)}{p_2(y)} = \frac{\int_{\Theta_1} p(y|\theta)\pi_1(\theta)d\theta}{\int_{\Theta_2} p(y|\theta)\pi_2(\theta)d\theta}.$$

	M_1	M_2	M_3	M_4
M_1	1.00	0.78	35.63	1.89
M_2	1.28	1.00	45.66	2.42
M_3	0.03	0.02	1.00	0.05
M_4	0.53	0.41	18.90	1.00

Prior 2 is always favored over the other priors. Generally, the marginal probability for a prior decreases as the prior density becomes more diffuse.

Bayesian model selection

The basic ingredient for model selection is then the marginal density:

$$p_j(y) = \int p(y|\theta_j)\pi_j(\theta_j)d\theta_j,$$

Gen from prior
Gen from Lik.

that is the normalizing constant of the posterior distribution under model M_j, also seen as the *likelihood* of the model M_j.

For any given model it can be written as:

$$\mathrm{E}_j[p(y|\theta_j)],$$

where the expectation is taken wrt the prior $\pi_j(\theta_j)$. Several aproximations are available: normal, Laplace, Monte Carlo, Importance sampling, composition method...

Computation of the marginal density

Evaluation of the integral in $p(y)$ can be performed using Monte Carlo computing methods or asymptotic expansions.

The Laplace expansion plays a central role in Bayesian inference since not only $p(y)$ but also many posterior summaries are expressible in terms of integrals (or as ratio of integrals) of the form:

$$I = \int h(\theta)\pi(\theta)p(y|\theta)d\theta,$$

for suitable functions $h(\theta)$. For I, the Laplace expansion gives:

$$\hat{I} = \frac{h(\hat{\theta})\pi(\hat{\theta})p(y|\hat{\theta})(2\pi)^{p/2}}{|j(\hat{\theta})|^{1/2}}\{1 + O(n^{-1})\},$$

-> NO IN EXAM

with $\hat{\theta}$ MLE of θ, p the parameter vector dimension and $j(\theta)$ the observed information matrix.

Indice

Evaluate predictive accuracy

BF → NO IMPROPER PRIORS, NO MULTI MODEL COMPARISON (ONLY PAIRWISE)

One way to evaluate a model is through the accuracy of its predictions. Sometimes we care about this accuracy for its own sake, as when evaluating a forecast. In other settings, predictive accuracy is valued not for its own sake but rather for comparing different models. We are interested in prediction accuracy for two reasons:

- to measure the performance of a model that we are using;
- second, to compare models. (More than pairwise)

PPC is biased bc supposes model good

If we consider data $y_1, \ldots, y_n$ modelled as independent given parameters θ, thus $p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$. A general summary of predictive fit is the log predictive density, $\log(p(y|\theta))$. When comparing models of differing size, it is important to make some adjustment for the natural ability of a larger model to fit data better, even if only by chance.

→ log likelihood

Evaluate predictive accuracy

The general form for the predictive information criteria that we will encounter is the following:

$$\text{crit} = -2\text{lpd} + \text{penalty}$$

- lpd is a measure of the log predictive density of the fitted model.
- penalty is a penalization accounting for the effective number of parameters of the fitted model.

The interpretation is the following: the lower is a particular value for an information criteria, and the better is the model fit. Moreover, if two competing models share the same value for the log predictive density, the model with less parameters is favored.

This is the Occam's Razor occurring in statistics:

Frustra fit per plura quod potest fieri per pauciora

Evaluate predictive accuracy

- Difficulty All the proposed measures are attempting to perform what is, in general, an impossible task: to obtain an unbiased and accurate measure of out-of-sample prediction error that will be valid over a general class of models and that requires minimal computation.

Indice

Akaike Information Criteria (AIC)

Let

- p be the number of parameters estimated in the model.
- $\hat{\theta}$ be the maximum likelihood estimate for θ.

The simplest bias correction is based on the asymptotic normal posterior distribution. In this limit (or in the special case of a normal linear model with known variance and uniform prior distribution), subtracting p from the log predictive density given the MLE is a correction for how much the fitting of p parameters will increase predictive accuracy, by chance alone:

$$\log(p(y|\hat{\theta})) - p.$$

As defined by Akaike (1973), Akaike Information Criteria (AIC) is the above multiplied by -2, thus:

$$\text{AIC} = -2\log(p(y|\hat{\theta})) + 2p. \tag{6}$$

→ Easiest possible solution
Non Bayesian
Hierarchical model: in case of partial pooling, θ's are p but the deg. of free. $< p$ (bc. a COMPLETE POOLING is a compromise bw. NO & COMPLETE)
PARAMS CONTRIBUTE LESS THAN 1

AIC

Bayesian "violation" → discard everything just care for DEVIANCE EVALUATED IN JUST SINGLE POINT

- It makes sense to adjust the deviance for fitted parameters, but once we go beyond linear models with flat priors, we cannot simply subtract p.
- Informative prior distributions and hierarchical structures tend to reduce the amount of overfitting, compared to what would happen under simple least squares or maximum likelihood estimation.
- For models with informative priors or hierarchical structure, the effective number of parameters strongly depends on the variance of the group-level parameters.
- Under the hierarchical model in the eight schools example, we would expect the effective number of parameters to be somewhere between 8 (one for each school) and 1 (for the average of the school effects).

Informative pri ⇒ add info ⇒ add shrinkage to the parameters
⇒ reduce dof of params → reduce effective #params.

Indice

Deviance Information Criteria (DIC)

A very popular approach has been proposed recently by Spiegelhalter et al (2002). It replaces the MLE $\hat{\theta}$ in (6) with the posterior mean $\hat{\theta}_{\text{Bayes}} = \mathrm{E}(\theta|y)$ and p with a data-based bias correction. The new measure of predictive accuracy is:

$$\log(p(y|\hat{\theta}_{\text{Bayes}})) - p_{\text{DIC}},$$

where p_{DIC} is the effective number of parameters, defined as:

$$p_{\text{DIC}} = 2(\log(p(y|\hat{\theta}_{\text{Bayes}})) - \mathrm{E}_{\theta|y}[\log(p(y|\theta)],$$

where the expectation in the second term is an average of θ over its posterior distribution, and is usually computed through the S draws from the posterior distribution. Then:

$$\text{DIC} = -2\log(p(y|\hat{\theta}_{\text{Bayes}})) + 2p_{\text{DIC}} \tag{7}$$

Indice

Bayes Information Criteria (BIC)

A simple way to approximate a marginal density, and thus the BF, is by the Laplace method. For one of the hypotheses, we apply the Laplace expansion to $p(y)$. Then, for large sample size n and up to order $O(1)$, we have:

$$-2\log(p(y)) \equiv \text{BIC} = -2\log(p(y|\hat{\theta})) + p\log n,$$

that is the Bayes Information Criterion (BIC), which is used for rough comparison of competing models, with $\hat{\theta}$ MLE of θ, p the parameter vector dimension.

BIC is splitted in two components:

- $-2\log(p(y|\hat{\theta}))$: this is the deviance, or the log predictive density of the data given a point estimate of the fitted model, multiplied by -2;
- $p\log n$: a penalty term, which is bigger as p and n increase.

BIC

Consider two models:

$$M_1 : (y_1, \ldots, y_n) \sim p_1(y|\theta_1), \theta_1 \in \Theta_1 \subset \mathbb{R}^{p_1}$$
$$M_2 : (y_1, \ldots, y_n) \sim p_2(y|\theta_2), \theta_2 \in \Theta_2 \subset \mathbb{R}^{p_2},$$

and let $\pi(\theta_1)$ and $\pi(\theta_2)$ be the priors. Then:

BF can be proxied by diff of BICs

$$2 \log BF_{12} = \triangle \text{BIC} = W - (p_2 - p_1) \log n,$$

with $W = 2(\log(p(y|\hat{\theta}_2)) - \log(p(y|\hat{\theta}_1)))$ the usual log-likelihood ratio test statistic.

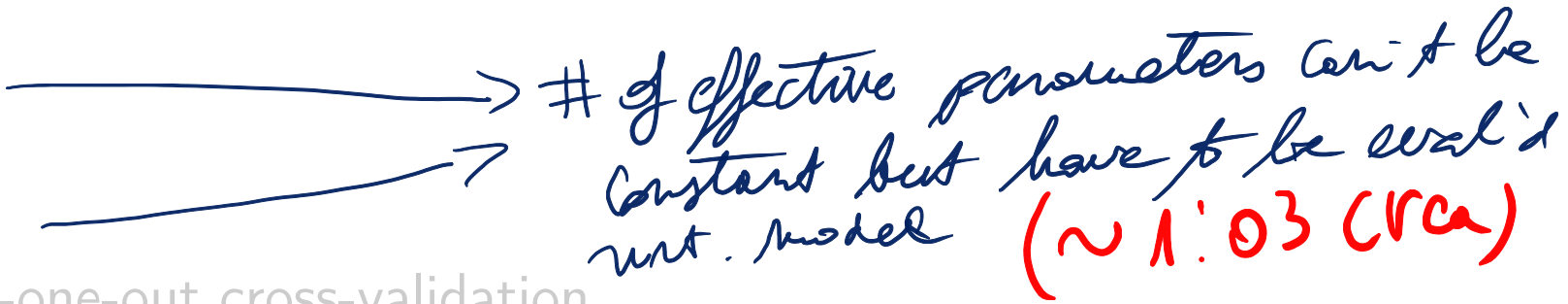
Comment The lower is the BIC for model 1 (2) when compared to model 2 (1) and the better is considered model 1 (2).

BIC

- BIC and its variants differ from the other information criteria considered here in being motivated not by an estimation of predictive fit but by the goal of approximating the marginal probability density of the data, $p(y)$, under the model, which can be used to estimate relative posterior probabilities in a setting of discrete model comparison.
- It is completely possible for a complicated model to predict well and have a low AIC, DIC, and WAIC, but, because of the penalty function, to have a relatively high (that is, poor) BIC. Given that BIC is not intended to predict out-of-sample model performance but rather is designed for other purposes, we do not consider it further here.

Penalizing for n can be dangerous!!

Indice



Watanabe-Akaike Information Criteria (WAIC)

We define the log pointwise predictive density for a single value y_i:

$$\text{lppd} = \sum_{i=1}^{n} \log(p(y_i|y)) = \sum_{i=1}^{n} \log \int p(y_i|\theta)\pi(\theta|y)d\theta. \quad (8)$$

To compute the lppd in practice, we can evaluate the expectation using draws from $\pi(\theta|y)$, the usual posterior simulations, which we label $\theta^{(s)}$, $s = 1, \ldots, S$, defining the computed log pointwise predictive density:

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{(s)}) \right) \quad (9)$$

WAIC

Then, we define the WAIC as follows:

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}, \tag{10}$$

where the quantity p_{WAIC} is defined as:

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\theta|y}(\log(p(y_i|\theta)),$$

→ Now use pointwise posterior variability

which computes the variance separately for each data point. We can practically compute this quantity by using:

$$\sum_{i=1}^{n} \text{Var}_{s=1}^{S}(\log(p(y_i|\theta^{(s)})),$$

→ CALCULATE & SUM SAMPLE VARIANCE ON EACH OF THE SIMULATION

where $\text{Var}_{s=1}^{S}$ represents the sample variance, $\text{Var}_{s=1}^{S} a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$.

WAIC

- Compared to AIC and DIC, WAIC has the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate.
- This is especially relevant in a predictive context, as WAIC is evaluating the predictions that are actually being used for new data in a Bayesian context. AIC and DIC estimate the performance of the plugin predictive density, but Bayesian users of these measures would still use the posterior predictive density for predictions.
- WAIC works also with singular models and thus is particularly helpful for models with hierarchical and mixture structures in which the number of parameters increases with sample size and where point estimates often do not make sense.

if prior variance is low, also n. eff. params. stays low, even if p high
NB: in Bayes, $p > n$ is no problem due to SHRINKAGE induced by the prior. The level of shrinkage is proportional to the informativity. Eg. 8 schools $N(\mu, \tau)$. $\tau \to 0$ → Complete pooling → 1 param; $\tau \to \infty$ → No pooling → J params; if I choose τ high → Peff high; τ small → Peff low, EVEN IF I STILL HAVE J PARAMS IN THE MODEL.

Indice

Leave-one-out cross-validation

In Bayesian cross-validation, the data are repeatedly partitioned into a training set y_{train} and a holdout set y_{holdout}, and then the model is fit to y_{train} (thus yielding a posterior distribution $\pi(\theta|y_{\text{train}})$), with this fit evaluated using an estimate of the log predictive density of the holdout data, $\log(p_{\text{train}}(y_{\text{holdout}})) = \log \int p_{\text{pred}}(y_{\text{holdout}}|\theta)\pi_{\text{train}}(\theta)d\theta$. The Bayesian leave-one-out cross-validation (LOO-CV) estimate of out-of-sample predictive fit is:

If regular data → don't care
Otherwise → see next slide

Log pointwise pred. dens. — All sample except i — Lik for unit i — Post. for all $-i$

$$\text{lppd}_{\text{loo-cv}} = \sum_{i=1}^{n} \log(p(y_i|y_{-i})) = \sum_{i=1}^{n} \log \int p(y_i|\theta)\pi(\theta|y_{-i})d\theta, \qquad (11)$$

where y_{-i} represents the data without the i-th data point. This quantity is usually calculated as:

$$\sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{i=1}^{n} p(y_i|\theta^{(s)}) \right)$$

Leave-one-out cross-validation

(no exam will be asked only for ideas)

- Each prediction is conditioned on $n-1$ data points, which causes underestimation of the predictive fit. For large n the difference is negligible, but for small n (or when using K-fold cross-validation) we can use a first order bias correction.
- Cross-validation is like WAIC in that it requires data to be divided into disjoint, ideally conditionally independent, pieces. This represents a limitation of the approach when applied to structured models.
- In addition, cross-validation can be computationally expensive except in settings where shortcuts are available to approximate the distributions $p(y_i|y_{-i})$.
- The purpose of using LOO or WAIC is to estimate the accuracy of the predictive distribution $p(\tilde{y}_i|y)$.

Importance sampling LOO (IS-LOO)

If the n points are conditionally independent in the data model we can then evaluate $p(y_i|y_{-i})$ with draws $\theta^{(s)}$ from the full posterior $\pi(\theta|y)$ using importance ratios:

$$r_i^{(s)} = \frac{1}{p(y_i|\theta^{(s)})} \propto \frac{\pi(\theta^{(s)}|y_{-i})}{\pi(\theta^{(s)}|y)}$$

Ratios may be very high if tails are ≠

to get the importance sampling leave-one-out (IS-LOO) predictive distribution,

$$p(\tilde{y}_i|y_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^{(s)} p(\tilde{y}_i|\theta^{(s)})}{\sum_{s=1}^{S} r_i^{(s)}} \tag{12}$$

Evaluating this LOO log predictive density at the held-out data point y_i, we get

$$p(y_i|y_{-i}) \approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\theta^{(s)})}}$$

PSIS-LOO

A direct use of r_i induces instability because the importance ratios can have high or infinite variance. We can improve the LOO estimate using Pareto smoothed importance sampling (PSIS), which applies a smoothing procedure to the importance weights. Here the main steps:

1. Since the distribution of the importance weights used in LOO may have a long right tail, we fit a generalized Pareto distribution to the tail (20% largest importance ratios $r^{(s)}$). The computation is done separately for each held-out data point i.
2. Stabilize the importance ratios by replacing the largest ratios by the expected values of the order statistics of the fitted generalized Pareto distribution. Label these values as $\tilde{\omega}_i^{(s)}$.
3. To guarantee finite variance of the estimate, truncate each vector of weights at $S^{3/4}\bar{w}_i$, where $\bar{w}_i$ is the average of the S smoothed weights corresponding to the distribution holding out data point i . Finally, label these truncated weights as $\omega_i^{(s)}$.

PSIS-LOO

The PSIS estimate of the LOO expected log pointwise predictive density (PSIS-LOO) is the same as in (12), but with the new weights ω_i in place of r_i. The LOOIC criteria is then defined as:

$$\text{LOOIC} = -2 \sum_{i=1}^{n} \log \left(\frac{\sum_{s=1}^{S} \omega_i^{(s)} p(\tilde{y}_i | \theta^{(s)})}{\sum_{s=1}^{S} \omega_i^{(s)}} \right) \tag{13}$$

Shrinked weights such that variance is not ∞

Strength of such criterion: use posterior predictive density (~1:26)

The estimated shape parameter $\hat{k}$ of the generalized Pareto distribution can be used to assess the reliability of the estimate:

- $k < 1/2$: the variance of the raw importance ratios is finite, the central limit theorem holds, and the estimate converges quickly.
- $k > 1/2$: the variance of the PSIS estimate is finite but may be large.

Simple applied example: election forecasting (Hibbs, 2008)

Forecast elections based on economic growth

We propose now a simple model to forecast elections based solely on economic growth. Better forecasts are possible using additional information such as incumbency and opinion polls, but what is impressive here is that this simple model does pretty well all by itself. Next table shows the year-by-year data, whereas next figure shows a quick summary of economic conditions and presidential elections over the past several decades. There is a clear linear relationship between economic growth and incumbent party's share of the popular vote. For simplicity, we predict y (vote share) solely from x (economic performance), using a linear regression,

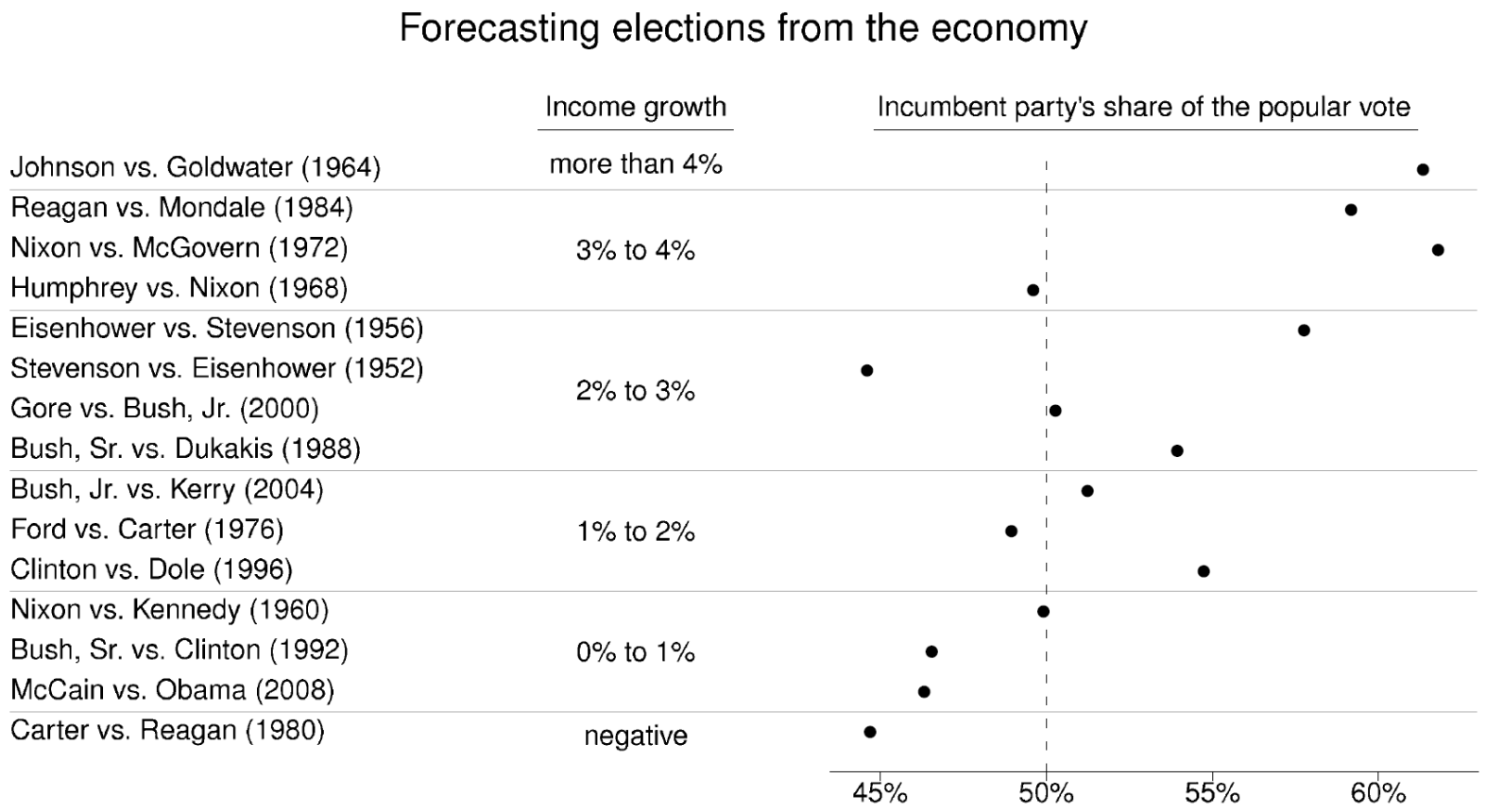
$$y \sim \mathcal{N}(a + bx, \sigma^2),$$

with a noninformative prior distribution, $p(a, b, \log \sigma) \propto 1$, so that the posterior distribution is normal-inverseχ^2.

Simple applied example: election forecasting (Hibbs, 2008)

Year	Growth rate	Vote share for incumbent party	Candidate of incumbent party	Candidate of other party
1952	2.40	44.60	Stevenson	Eisenhower
1956	2.89	57.76	Eisenhower	Stevenson
1960	0.85	49.91	Nixon	Kennedy
1964	4.21	61.34	Johnson	Goldwater
1968	3.02	49.60	Humphrey	Nixon
1972	3.62	61.79	Nixon	McGovern
1976	1.08	48.95	Ford	Carter
1980	-0.39	44.70	Carter	Reagan
1984	3.86	59.17	Reagan	Mondale
1988	2.27	53.94	Bush, Sr.	Dukakis
1992	0.38	46.55	Bush, Sr.	Clinton
1996	1.04	54.74	Clinton	Dole
2000	2.36	50.27	Gore	Bush, Jr.
2004	1.72	51.24	Bush, Jr.	Kerry
2008	0.10	46.32	McCain	Obama

Simple applied example: election forecasting (Hibbs, 2008)



Above matchups are all listed as incumbent party's candidate vs. other party's candidate.
Income growth is a weighted measure over the four years preceding the election. Vote share excludes third parties.

Simple applied example: election forecasting (Hibbs, 2008)

- Fit to all 15 data points in Figure, the posterior mode $(\hat{a}, \hat{b}, \hat{\sigma})$ is (45.9, 3.2, 3.6).
- Although these data form a time series, we are treating them here as a simple regression problem.
- In our regression example, the log predictive probability density of the data is $\sum_{i=1}^{15} \log(\mathcal{N}(a + bx_i, \sigma^2))$, with an uncertainty induced by the posterior distribution $\pi(a, b, \sigma^2|y)$, which is a Normal-Inverseχ^2.

Simple applied example: election forecasting (Hibbs, 2008)

Let's manually compute the predictive information criteria:

- AIC The MLE is $(\hat{a}, \hat{b}, \hat{\sigma}) = (45.9, 3.2, 3.6)$. The estimated parameters are 3. Thus:

$$\text{AIC} = -2 \sum_{i=1}^{15} \log(\mathcal{N}(45.9 + 3.2x_i, 3.6^2)) + 2 \times 3 = 86.6.$$

Simple applied example: election forecasting (Hibbs, 2008)

- DIC The relevant formula is $p_{\text{DIC}} = 2(\log(p(y|\hat{\theta}_{\text{Bayes}}) - \text{E}_{\theta|y}[\log(p(y|\theta)]$. The second of these terms is invariant to reparameterization, we calculate it with S draws. The first term is not invariant:

$$\text{E}_{\theta|y}[\log(p(y|\theta)] = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{15}\log(\mathcal{N}(a^{(s)} + b^{(s)}x_i, (\sigma^{(s)})^2) = -42$$

$$\log(p(y|\hat{\theta}_{\text{Bayes}})) = \sum_{i=1}^{15}\log(\mathcal{N}(\text{E}(a|y) + \text{E}(b|y)x_i, (\text{E}(\sigma|y))^2) = -40.5,$$

which gives $p_{DIC} = 2(-40.5 - (-42.0)) = 3.0$. Finally:

$$\text{DIC} = -2\log(p(y|\hat{\theta}_{\text{Bayes}}) + 2p_{\text{DIC}} = -2 \times (42 + 1.5) = 87$$

Simple applied example: election forecasting (Hibbs, 2008)

- WAIC

$$\text{lppd} = \sum_{i=1}^{15} \log \left(\frac{1}{S} \sum_{s=1}^{S} \log(\mathcal{N}(a^{(s)} + b^{(s)} x_i, (\sigma^{(s)})^2) \right) = -40.9.$$

The effective number of parameters can be calculated as:

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{s=1}^{S} \log(\mathcal{N}(a^{(s)} + b^{(s)} x_i, (\sigma^{(s)})^2) = 2.7.$$

Thus:

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}} = 81.8 + 5.4 = 87.2.$$

Simple applied example: election forecasting (Hibbs, 2008)

- LOOIC We fit the model 15 times, leaving out a different data point each time. For each fit of the model, we sample S times from the posterior distribution of the parameters and compute the log predictive density. The cross-validated pointwise predictive accuracy is:

$$\text{lppd}_{\text{loo-cv}} = \sum_{l=1}^{15} \log \left(\frac{1}{S} \sum_{s=1}^{S} \log(\mathcal{N}(a^{(ls)} + b^{(ls)} x_l, (\sigma^{(ls)})^2) \right) = -43.8$$

Then:

$$\text{LOOIC} = -2\text{lppd}_{\text{loo-cv}} = 87.6$$

Finally, the effective number of parameters is:

$$p_{\text{loo-cv}} = \text{E}_{\text{lppd}} - \text{E}_{\text{lppd}_{\text{loo-cv}}} = 2.9$$

Simple applied example: election forecasting (Hibbs, 2008)

	Value	Eff. par.
AIC	86.6	3
DIC	87	3
WAIC	87.2	2.7
LOOIC	87.6	2.9

- Given that this model includes two linear coefficients and a variance parameter, these all look reasonable as an effective number of parameters.
- The four criteria tend to be similar due to the model simplicity: as the complexity grows, AIC and DIC tend to loose power in predictive accuracy.

Indice

Implementation in Stan

We illustrate how to write Stan code that computes and stores the pointwise log-likelihood using the eight schools example. The model is unchanged, we only need to store the pointwise log-likelihood (the `log_lik` object) in the `generated quantities` block:

```
...
generated quantities {
 vector[J] log_lik;
  for (j in 1:J){
     log_lik[j] = normal_lpdf(y[j]| theta[j], sigma[j]);
  }
}
```

The `loo` package

The `loo` R package provides the functions `loo()` and `waic()` for efficiently computing PSIS-LOO and WAIC for fitted Bayesian models using the methods described before.

These functions take as their argument an $S \times n$ loglikelihood matrix, where S is the size of the posterior sample (the number of retained draws) and n is the number of data points.

The `loo()` function returns PSIS-LOOIC and p_{LOO}. The `waic()` function computes the analogous functions for WAIC.

Using the loo package

```
y <- c(28,8,-3,7,-1,1,18,12)
sigma <- c(15,10,16,11,9,11,10,18)
J <- 8
data <- list(y = y, sigma=sigma, J = J)
fit_1 <- stan("8schools.stan",
              data = data, iter=200,
              cores = 4, chains =4)
#computing psis-looic
log_lik_1 <- extract_log_lik(fit_1)
loo_1 <- loo(log_lik_1)
print(loo_1)

         Estimate  SE
elpd_loo    -30.8 0.9
p_loo         1.3 0.3
looic        61.7 1.8
```

Eight schools example: model comparison

Model:

$$y_j \sim \mathcal{N}(\theta_j, \sigma_y^2)$$
$$\theta_j \sim \mathcal{N}(\mu, \tau^2)$$

Three possible priors (then, three models):

1. $\tau \propto 1$
2. $\tau^2 \sim \text{InvGamma}(0.001, 0.001)$
3. $\tau \sim \text{HalfCauchy}(0, 2.5)$

Let's compare the model through LOOIC and WAIC.

Eight schools example: model comparison. LOOIC

```
loo_diff <- compare(loo_1, loo_2, loo_3)
loo_diff
      elpd_diff se_diff elpd_loo p_loo looic
loo_2   0.0       0.0   -30.6      0.8  61.1
loo_3  -0.1       0.0   -30.6      0.8  61.2
loo_1  -0.4       0.3   -31.0      1.4  61.9
```

Model 2 (inverse gamma) is slightly favorite in terms of lower LOOIC. Model (1) reports the lowest LOOIC. Anyway, differences between models are quite negligible.

Eight schools example: model comparison. WAIC

```
waic_1 <- waic(log_lik_1)
waic_2 <- waic(log_lik_2)
waic_3 <- waic(log_lik_3)

waic_diff <- compare(waic_1, waic_2, waic_3)
waic_diff
```

```
       elpd_diff se_diff elpd_waic p_waic waic
waic_2   0.0       0.0   -30.6       0.8   61.1
waic_3   0.0       0.0   -30.6       0.8   61.2
waic_1  -0.3       0.3   -30.9       1.3   61.8
```

Model 2 (inverse gamma) is slightly favorite in terms of lower WAIC. The number of effective parameters, p_{WAIC}, is 0.8 for model 2 and 3, and 1.3 for model 1 (uniform prior).

Major League Soccer models: model comparison

Let's compute the PSIS-LOO for the Major League Soccer models:

```
library(LearnBayes)
data(soccergoals)
y <- soccergoals$goals
mls_data <- list(y=y, N=length(y))
mls_fit_1 <- stan('mls_gamma.stan', data =mls_data,
                  iter =500, cores = 4 )

mls_data <- list(y=y, N=length(y), mu=1, tau=0.5)
mls_fit_2 <- stan('mls_normal.stan', data =mls_data,
                  iter =500, cores = 4 )
mls_data <- list(y=y, N=length(y), mu=2, tau=0.5)
mls_fit_3 <- stan('mls_normal.stan', data =mls_data,
                  iter =500, cores = 4 )
mls_data <- list(y=y, N=length(y), mu=1, tau=2)
mls_fit_4 <- stan('mls_normal.stan', data =mls_data,
                  iter =500, cores = 4 )
```

Major League Soccer models

```
log_lik_1 <- extract_log_lik(mls_fit_1)
loo_1 <- loo(log_lik_1)
log_lik_2 <- extract_log_lik(mls_fit_2)
loo_2 <- loo(log_lik_2)
log_lik_3 <- extract_log_lik(mls_fit_3)
loo_3 <- loo(log_lik_3)
log_lik_4 <- extract_log_lik(mls_fit_4)
loo_4 <- loo(log_lik_4)

loo_diff <- compare(loo_1, loo_2, loo_3, loo_4)
```

```
      elpd_diff se_diff elpd_loo p_loo looic
loo_2   0.0       0.0   -53.2      0.7 106.3
loo_1   0.0       0.0   -53.2      0.7 106.4
loo_4  -0.1       0.2   -53.3      0.8 106.6
loo_3  -0.3       0.4   -53.5      0.8 107.0
```

Hibbs model: forecasting elections

```
data {
  int N;
  vector[N] y;
  vector[N] X;
}
parameters {
  real a;
  real b;
  real<lower=0> sigma;
}
model {
  target+= normal_lpdf(y|a+X*b, sigma);     // data model
  target+=-log(sigma);    // log prior for p(sigma) propto 1/sigma
}
generated quantities {
  vector[N] log_lik;           // pointwise log-likelihood
    for (n in 1:N)
        log_lik[n] = normal_lpdf(y[n]| a+X[n]*b, sigma);
}
```

Hibbs model: forecasting elections

```
log_lik_hibbs <- extract_log_lik(fit_hibbs)
loo_hibbs <- loo(log_lik_hibbs)
print(loo_hibbs)

         Estimate  SE
elpd_loo    -43.6 3.4
p_loo         2.7 1.0
looic        87.3 6.8

waic(log_lik_hibbs)

          Estimate  SE
elpd_waic    -43.5 3.4
p_waic         2.6 1.0
waic          87.0 6.7
```

Hibbs model: forecasting elections

- We retrieved the same results obtained analytically.
- If we had other covariates, we could add them in the model and compare the LOOIC and the WAIC of this extended model with those for the basic model.

Some considerations

In comparing nested models, the key questions of model comparison are typically: (1) is the improvement in fit large enough to justify the additional difficulty in fitting, and (2) is the prior distribution on the additional parameters reasonable?

The second scenario of model comparison is between two or more nonnested models-neither model generalizes the other. One might compare regressions that use different sets of predictors to fit the same data. In these settings, we are typically not interested in choosing one of the models-it would be better, both in substantive and predictive terms, to construct a larger model that includes both as special cases, including both sets of predictors and also potential interactions in a larger regression, possibly with an informative prior distribution if needed to control the estimation of all the extra parameters.

Some considerations

Formulas such as AIC, DIC, and WAIC fail in various examples: AIC does not work in settings with strong prior information, DIC gives nonsensical results when the posterior distribution is not well summarized by its mean, and WAIC relies on a data partition that would cause difficulties with structured models such as for spatial or network data. Cross-validation is appealing but can be computationally expensive and also is not always well defined in dependent data settings.

But there are times when it can be useful to compare highly dissimilar models, and, for that purpose, predictive comparisons can make sense. In addition, measures of effective numbers of parameters are appealing tools for understanding statistical procedures

Further readings

Further reading:

- Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016. Here the ▸ pdf
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5),1413–1432. Here the ▸ pdf