Can we make Objective Bayesian Model comparisons in a subjective Bayes world?

ΟΠΑ
AUEB

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Ioannis Ntzoufras

(ntzoufras@aueb.gr)

Based on a Joint work with: B. Liseo, G. Consonni & D. Fouskakis (Bayesian Analysis, 2018)

19–22 December 2018: Greek Stochastics κ', Athens

In this talk I will try the preposterous, the improbable, the unbelievable ...

In this talk I will try the preposterous, the improbable, the unbelievable ...

(a) To be objective within a subjective scientific theory!

In this talk I will try the preposterous, the improbable, the unbelievable ...

(a) To be objective within a subjective scientific theory!

(b) To fit the talk in the 30 minutes spot (incl. discussion).

In this talk I will try the preposterous, the improbable, the unbelievable ...

(a) To be objective within a subjective scientific theory!

(b) To fit the talk in the 30 minutes spot (incl. discussion).

Spoiler alert...

In this talk I will try the preposterous, the improbable, the unbelievable ...

(a) To be objective within a subjective scientific theory!

(b) To fit the talk in the 30 minutes spot (incl. discussion).

Spoiler alert...

Probably I will not be able to do a very good job with these two tasks

Objective & Bayes?

1. This is an “oxymoron” since Bayes is by definition subjective.
2. It is a “marketing” term for the implementation of the Bayesian methods under the absence of prior information; other (not so “attractive”) alternative is “Default Bayes” .
3. Even I do not like the term, I wrote a review paper in *Bayesian Analysis* in 2018 co-authored with G. Consonni, D. Fouskakis and B. Liseo.
4. O’Bayes has long tradition within ISBA (13 biannual meetings with over 100 participants per meeting $\Rightarrow$ so it is a real thing).
5. Research focuses on Default priors for inference, for Model comparisons, Prior combatibility across models, Bayesian Non Parametrics, Shrinkage methods for large p small n problems.

How can we specify Objective Bayes within model comparisons?

How can we specify Objective Bayes within model comparisons?

Any sensible Bayesian model comparison or selection procedure when no prior information is available.

How can we specify Objective Bayes within model comparisons?

Any sensible Bayesian model comparison or selection procedure when no prior information is available.

Why **ANY** procedure and not **THE** procedure?

How can we specify Objective Bayes within model comparisons?

Any sensible Bayesian model comparison or selection procedure when no prior information is available.

Why **ANY** procedure and not **THE** procedure?

Because of the “well-known” sensitivity of Bayes factors on priors with large prior variances which are considered uninformative within each model.

So even small changes in the prior distribution will lead to slightly different posterior solutions.

So what should we do?

So what should we do?

We use any procedure based on some basic principles which act as requirements for having sensible posterior procedures under no prior information.

So now we have available

So now we have available

- A bag of basic principles (which act as requirements for having good methods)

So now we have available

- A bag of basic principles (which act as requirements for having good methods)

- A toolbox of methods for prior construction (that lead to reasonable solutions)

So now we have available

- A bag of basic principles (which act as requirements for having good methods)
- A toolbox of methods for prior construction (that lead to reasonable solutions)

I will focus on the variable selection problem which is the most popular setup for model comparison & evaluations

Bayesian Model Comparison

Posterior Odds (PO) between models M_0 and M_1 is given by

$$PO_{01} \equiv \frac{\pi(M_0|\boldsymbol{y})}{\pi(M_1|\boldsymbol{y})} = \frac{m_0(\boldsymbol{y})}{m_1(\boldsymbol{y})} \times \frac{\pi(M_0)}{\pi(M_1)} = BF_{01} \times O_{01} \tag{1}$$

which is a function of the **Bayes Factor** (BF_{01}) and the **Prior Odds** (O_{01}).

In the above $m_\ell(\boldsymbol{y})$ is the marginal likelihood under model M_ℓ and $\pi(M_\ell)$ is the prior probability of model M_ℓ given by

$$m_\ell(\boldsymbol{y}) = \int f_\ell(\boldsymbol{y}|\boldsymbol{\theta}_\ell)\pi_\ell(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell, \tag{2}$$

where $f_\ell(\boldsymbol{y}|\boldsymbol{\theta}_\ell)$ is the likelihood under model M_ℓ with parameters $\boldsymbol{\theta}_\ell$ and $\pi_\ell(\boldsymbol{\theta}_\ell)$ is the prior distribution of model parameters given model M_ℓ.

The Lindley-Bartlett-Jeffreys Paradox

For a single model inference $\Rightarrow$ a highly diffuse prior on the model parameters is often used (to represent ignorance).

$\Rightarrow$ Posterior density takes the shape of the likelihood and is insensitive to the exact value of the prior density function.

For multiple models inference $\Rightarrow$ BFs (and POs) are quite sensitive to the choice of the prior variance of model parameters.

$\Rightarrow$ For nested models, we support the simplest model with the evidence increasing as the variance of the parameters increase ending up to support of more parsimonious model no matter what data we have.

$\Rightarrow$ Under this approach, the procedure is quite informative since the data do not contribute to the inference.

$\Rightarrow$ Improper priors cannot be used since the BFs depend on the undefined normalizing constants of the priors.

Principles for O'Bayes Model Comparisons

- Compatibility of priors.
- Validation of Bayesian approaches.
- Methods with good frequentist properties

  (FDR control - application in Quality control and clinical trials).
- Criteria for objective Bayesian model choice (Bayarri et al., 2012; Annals Stat.).

Compatibility of Priors

Priors of model parameters should be related across models, although in principle they need not be.

- Compatibility is necessary because of

  a) the Lindley-Bartlett-Jeffreys paradox

  b) same prior information should be dilluted in parameters of different dimension.

- Compatibility was initially proposed

  a) to lessen the sensitivity of model comparison to prior specifications, and

  b) to facilitate the task of multiple prior elicitations when several models are entertained.

Compatibility of Priors (cont'd)

Priors of model parameters should be related across models, although in principle they need not be.

- Compatibility is usually applied to nested models
- It can be extended to more general setups whenever we can identify a benchmark model (often the null model), which is nested into every other model under consideration.

Most usual approach/example of induced compatibility is achieved by using imaginary or historical data which specify the prior distributions of our actual analysis as posterior distributions based on the "a-priori" available data.

Validation of Bayesian approaches

An acceptable Bayesian procedure should correspond, at least asymptotically, to a prior which makes sense in the context where it is applied.

Examples:

- BIC $\Rightarrow$ unit information priors

  (also equivalent to variable selection Zellner's g-prior with $g = n$ in normal linear regression setup)

- Arithmetic Mean Intrinsic Bayes Factor (Berger & Pericchi, 1996; JASA) $\Rightarrow$ Intrinsic priors; see Cassela & Moreno (2006; JASA) for implementation in normal linear regression.

Methods with good frequentist properties

A popular alternative to the standard objective Bayes techniques is to use prior distributions that lead to good frequentist performances.

- This trend is especially notable in high-dimensional settings.
- The control of False Discovery Rate (FDR) is tailored to multiple comparisons in order to control the problem of multiplicity.
- Applications in Clinical trials and Quality Control

Criteria for objective Bayesian model choice

(Bayarri et al., 2012; Annals Stat.).

- Propriety;
- Model Selection Consistency; Information consistency; intrinsic consistency;
- Predictive matching;
- Measurement Invariance; Group Invariance

Criteria for objective Bayesian model choice

C1: The basic criterion - Propriety

The prior of each model specific parameter, conditionally on the common ones, $\pi(\boldsymbol{\theta}_{\ell\backslash 0}|\boldsymbol{\theta}_0, M_\ell)$, should be proper.

Basic requirement in order to be able to calculate Bayes factors and posterior model probabilities.

Criteria for objective Bayesian model choice

C2: Model Selection Consistency

If a true model exists M_{TRUE} and it is within the models we consider $M_{TRUE} \in \mathcal{M}$, then
the posterior probability of $M_{TRUE} \to 1$ when $n \to \infty$.

- Crucial requirement in order to have a sensible model comparison procedure.
- It ensures that for suitably large sample, the true model will be identified.
- It does not ensures that the method will give a reasonable solution for any specific problem with a specific sample size n.
- We need also to investigate the rate of convergence to M_{TRUE}.

Criteria for objective Bayesian model choice

C3: Information consistency

For any sequence of datasets with the same sample size n such that
[Likelihood ratio (LR) of M_ℓ vs. M_0] $\rightarrow \infty \Rightarrow$ BF $\rightarrow \infty$.

- It mean that as the signal of the systematic model componenent gets stronger, then we a-posteriori support more and more this model.

- Used by Liang $et\ al.$ (2008) to support the introduction of mixtures of g-priors vs. the traditional g-priors.

- Questionable — Zellner strongly objected the need of information consistency for fixed n leading to certain decisions for fixed n.

Criteria for objective Bayesian model choice

C4: Intrinsic consistency

When $n \longrightarrow \infty$, the prior should converge to a limiting proper prior, free of dataset based characteristics such as the sample size.

Criteria for objective Bayesian model choice

C5: Predictive matching

For all samples of a *minimal size*, one should not be able to discriminate between two models, that is BF=1 (exact matching) or BF$\approx$1 (approximate matching).

- Crucial criterion/requirement.
- The minimal sample size n^* is defined as the smallest sample size we can obtain a proper posterior.

Criteria for objective Bayesian model choice

C6: Measurement Invariance

Proposed solutions should not be affected by changes of measurement units.

- Reasonable but not essential to get a sensible solution.

Criteria for objective Bayesian model choice

C7: Group Invariance

Group invariance criterion can be understood as a formalization of the Jeffreys' requirement that the prior for a non-null parameter should be "centered at the simplest model."

- The *group invariance criterion (C7)* states that if models M_ℓ and M_0 are invariant under a group of transformations G_0, then the conditional priors $\pi(\boldsymbol{\theta}_{\ell\setminus 0}|\boldsymbol{\theta}_0, M_\ell)$ should be chosen in such a way that the conditional marginal distribution $f(\boldsymbol{y}|\boldsymbol{\theta}_0, M_\ell)$ is also invariant under G_0.
- Can be used to find priors on common parameters.
- Local prior approach as opposed to non-local prior approach.
- Reasonable under one perspective but not essential to get a sensible solution.

Tools for O'Bayes Model Comparisons

- The Unit information approach.
- Training Samples $\Rightarrow$ Intrinsic Bayes Factors $\Rightarrow$ Intrinsic Priors.
- Imaginary Data.
  - Fixed Imaginary Data $\Rightarrow$ power prior $\Rightarrow$ g-prior $\&$ its mixtures.
  - Random Imaginary Data $\Rightarrow$ Expected posterior prior $\&$ Power-EPPs.
- Emprirical Bayes approaches.
- Non-local prior approaches.

The Unit information approach

Informal Definition: a unit information prior (UIP) has an information content equivalent to a sample of size one.

- UIP introduced by Kass & Wasserman (1995; JASA).

$\Rightarrow$ -2log(BF) $\approx$ BIC for $n \longrightarrow \infty$.

- UIP provides a Bayesian interpretation for the BIC model selection procedure.

The Unit information approach (cont'd)

$$\boldsymbol{\theta}_\ell | M_\ell \sim \mathrm{N}_{d_\ell}\left(\boldsymbol{\mu}_{\theta_\ell},\, n\left[\mathcal{J}_\ell^n(\boldsymbol{\mu}_{\theta_\ell})\right]^{-1}\right), \qquad (3)$$

where $\mathcal{J}_\ell^n(\cdot)$ is the negative of the Hessian matrix of the log-likelihood.

Simplified approach: Empirical UIP

1. Run the full model with flat priors
2. Use the posterior means and variances to specify prior parameters for model selection using independent priors.
3. All prior variances are multiplied by n to ensure information equivalent to one data-point.

The posterior model probabilities under this approach can be used as an initial yardstick for comparisons with other objective Bayes approaches; see Ntzoufras (2009) for examples.

The Unit information approach (cont'd)

UIP can be obtained by using imaginary/prior data and the power-prior approach (will be discussed soon).

Under this setting, the prior mean $\boldsymbol{\mu}_{\theta_\ell}$

Examples of priors based on the unit information principle:

- Zellner's g-prior for normal linear regression.
- The g-prior of Ntzoufras $et\ al.$ (2003) for binary response models
- The prior of Overstall & Forster (2010) for GLMMs
- The g-prior extension for GLMs of Bove & Held (2011), and
- The priors used by Ntzoufras & Tarantola (2008) for graphical models for contingency tables.

UIPs can be easily used for i.i.d. observations. Not obvious how to extend for non i.i.d. cases icluding hierarchical models where we might need to incorporate the use of Efficient sample size.

Using Training Samples

Using Training samples to build priors and Validation samples for model evaluation
$\Rightarrow$ Partial Bayes Factors $\Rightarrow$ Intrinsic Bayes Factors (IBFs).

- Under the IBF approach (Berger & Pericchi, 1996l; JASA), *minimal training samples* are used to "train" our prior.
- *Minimal training samples* $\Rightarrow$ samples "as small as possible, subject to yielding proper posteriors".
- Disadvantage: Need to consider all possible sub-samples with minimal sample size, and then take averages (or other summaries). This can be computationally costly.
- This lead to the use of *intrinsic priors.*

Using Training Samples (cont'd)

Related Approach:

Using Fractions of likelihood to train priors and the rest for model evaluation

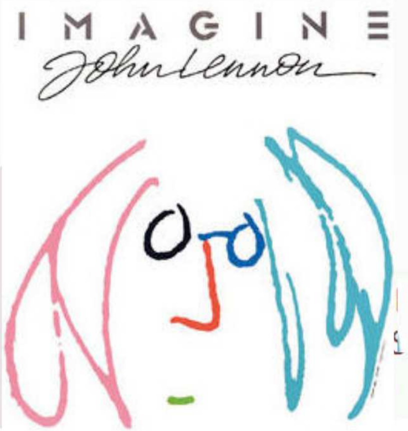
⇒ Fractional Bayes Factors (FBFs).

- FBF was introduced by O'Hagan (1995) but he claims that it is not an O'Bayes techniques – Please do not tell him...
- Epic rivalry between IBF and FBF especially in the 6th Valencia Meeting (1998).
- It does not require training samples but fractions of the likelihood to be used for training the prior.

Imagining Data

We train our priors via a thought experiment with an appropriate dataset.

- Traced back to the work of Good in 1950s.
- Main pathway ⇒ imaginary dataset fully supports the null hypothesis in nested model comparisons (local approach; C7 criterion).
- In order to make them minimally informative, they are combined with the notions of
  - minimal training samples, or
  - Unit information approaches

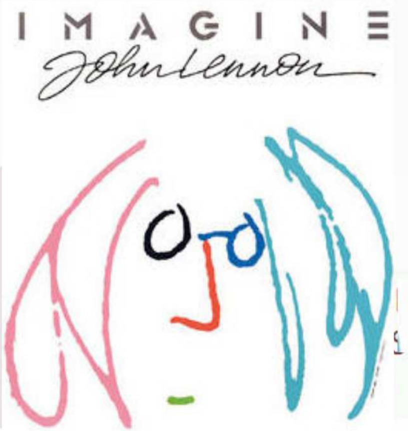


Imagining Data (cont'd)

- **Fixed imaginary observations**
  - **Power prior**: Raises the likelihood to a power but by using historical/imaginary data.

    Similar to fractional approach but not using the actual data.

    Use power parameter $= 1 \Rightarrow$ Unit information prior/approach.
  - Fixed Imaginary Data $\Rightarrow$ power prior $\Rightarrow$ g-**prior** $\&$ its **mixtures**.
- **Random Imaginary Data**: Generate imaginary data from a null model
  - **Expected posterior prior (EPP)** $\&$ **Power-EPPs**.



Empirical Bayes approaches (Isn’t it the Devil?)

We train our priors for some parameters via the actual data (or part of them).

- **Past:** One of the biggest sins of Bayesians.
- **But** it can provide sensible solutions ⇒ More and more people are using such approaches under prior ignorance.
- **Main criticism**: double use of the data which violates a basic principle of Bayesian theory.
- This can be mitigated by combining EB with other ideas described in the previous section, such as the unit information principle, in order to minimize the re-use of the data especially in cases when the sample size is not large.

Empirical Bayes approaches (cont'd)

Examples of Using Empirical Bayes

EB methods in model selection usually focus on the specification of the prior for a small number of parameters, typically those causing the sensitivity of the Bayes factor, for example:

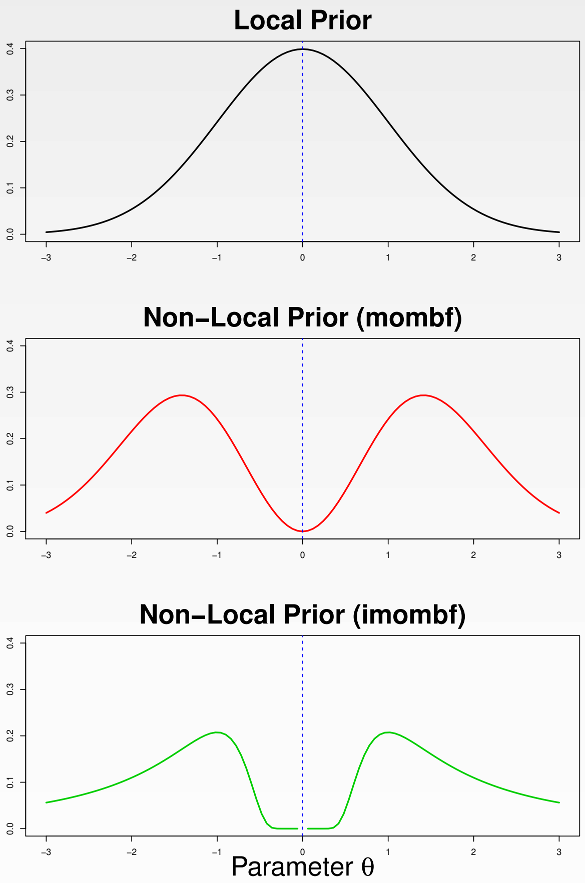
(a) g in the g-prior (George and Forster, 2000; Liang $et\ al.$ 2008).

(b) the prior inclusion probability (George and Forster, 2000; Scott and Berger 2010).

(c) the shrinkage parameter in lasso (Yuan & Lin, 2005).

Non-local prior approaches

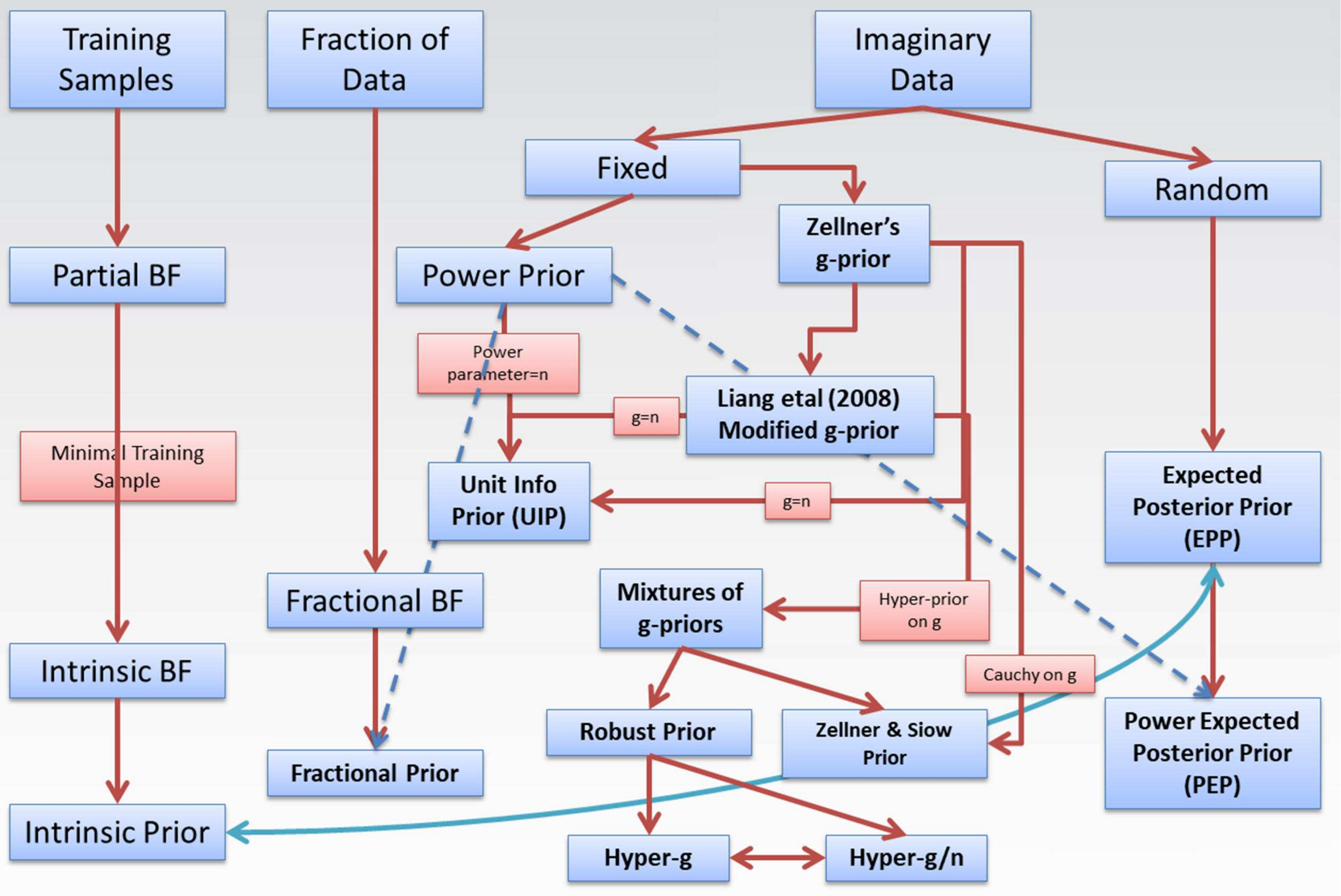
Non-local priors assign zero or deminishing probabilities in intervals of parameters values supported by the alternative models.



- Introduced by Johnson & Rossel (2010) in order to improve convergence rates in favor of the true null hypothesis.
- They are in contrast to the mainstream local prior approaches who assign priors centered at the parameter values specified by the null model.
- Example of such priors are the **moment prior** and the **inverse moment prior** (Johnson & Rossel, 2010).

O’Bayes Variable Selection in Regression

Priors for Model Coefficients: It’s too complicated!



Priors on the Model Space

- Not a lot of work because focus is given in the priors for parameters.
- But priors on the model space are important because they
  - They control sparsity and the weight of parsimony.
  - They can be used to account for the problem of multiplicity.

Priors on the Model Space (cont'd)

Past: Uniform Prior on the model space

⇒ Bernoulli(1/2) for each covariate inclusion in variable selection.

1. Naive choice — Seems non informative but it is not.
2. Does not account for model complexity/preference.
3. Does not account for multiple comparisons.
4. Is totally informative in terms of dimensionality ⟶ in variable selection with 1000 covariates it will a-priori support models with average dimension of 500 covariates!

Priors on the Model Space (cont'd)

Now: Other alternatives (variable selection)

Beta-binomial prior in variable selection: Assumes that every covariate inclusion indicator a-priori follows a Bernoulli(p) with p following a Beta hyperprior (George & Forster 2010; Scott & Berger 2010)

- The parameters of the beta control the sparsity.
- Default choice: $p \sim Uniform(0, 1) \Rightarrow$ uniform (also) on the model dimension.
- Good for sparse problems.
- Leads to overfitting when the proportion of important covariates is high (not so often in practice).

Other hyper-priors have been also proposed in the $n << p$ literature.

Priors on the Model Space (cont'd)

Now: Other alternatives (general comparisons)

- The Joint prior specification approach of Ntzoufras $et\ al.$ (2008; Stat.Sci.).
  - Adjusts prior model probabilities to eliminate Lindleys Paradox.
  - Actually flat prior can be used within each model.
  - A prior complexity penalty must be specified.
- Prior adjustment based on Minimim KL Divergence by Villa & Walker (2015; Scan.J.Stat.)
  - Adjusts prior model probabilities based on the minimum KL divergence of each model compared to other. The greater the distance, the higher the "*worth*" of the model.
  - In variable selection we need to specify a complexity function

Concluding Comments: Can we do O'Bayes model comparisons?

Purely objective model comparisons are not possible.

In every model comparison, we need to make subjective choices which will influence specific properties of our model selection procedure such as

1. Informativeness within model
2. Sparsity
3. Dimension complexity and model parsimony

What can we do within O'Bayes framework?

Reasonable model comparisons that follow some generally acceptable principles, rules and criteria.

Default Popular Choices for Variable Selection in Regression: Zellner's g-prior with $g = n$, Hyper-g prior, beta-uniform prior on model space, BIC & EBIC as approximate techniques.

Default Popular R packages: BAS, BayesVarSel, mombf (for non-local).

Phew! Did I manage to say all this in 25 or 30 minutes?

Don’t go! There is more…

- **Statistics**

- Statistics

- Greeks

- Statistics
- Greeks
- Italians

- Statistics
- Greeks
- Italians
- And other nationalities (English, Austrian, Spanish, Americans)

- Statistics
- Greeks
- Italians
- And other nationalities (English, Austrian, Spanish, Americans)
- In an Island!

- Statistics
- Greeks
- Italians
- And other nationalities (English, Austrian, Spanish, Americans)
- In an Island! It’s not a joke...

- **Statistics**
- **Greeks**
- **Italians**
- **And other nationalities (English, Austrian, Spanish, Americans)**
- **In an Island!** It’s not a joke... It’s a conference

- **Statistics**
- **Greeks**
- **Italians**
- **And other nationalities (English, Austrian, Spanish, Americans)**
- **In an Island!** It’s not a joke... It’s a conference
- **WHAT Can you Ask for more in your Academic life?**

5th Meeting on Statistics

Friday 6 – Sunday 8 September 2019
Aegina

Statistics5@Aegina`

Do not miss it!