ECOM168: Introduction to Bayesian model selection

Daniele Bianchi

School of Economics and Finance,
Queen Mary University of London

Outline

1. Introduction

2. Decision-theoretic foundations

3. A re-cap of Bayesian linear regressions

4. Zellner's g-prior

5. A primer on Bayesian model selection

Introduction

Introduction

In all areas of economic research data sets are increasing in both size and complexity.

High-dimensional models – either linear or non-linear – are more and more common in applied econometrics work.

The purpose of this module is to introduce Bayesian inference within the context of large-scale regression models based on:

- ↪ Hierarchical priors
- ↪ Shrinkage priors
- ↪ Variables selection

Introduction

Why Bayesian?

The benefits of Bayesian inference to approach high-dimensional problems are several.

↪ More accurate/intuitive quantification of model uncertainty.

↪ Parameters are random variables, i.e., they have their own probability density function.

↪ The use of a prior distribution provides a natural starting point to mitigate "weak" information in the likelihood.

↪ Regularisation/shrinkage is embedded in the prior specification, not in the loss function/likelihood.

The flexibility of the prior is particularly helpful within the context of financial data, i.e., small signal-to-noise ratio, and large-dimensional data where the risk of overfitting is higher.

Introduction

The aim of these lectures:

1. Re-cap the theoretical motivation behind Bayesian inference.
2. Explore the classes of priors and their mapping with conventional penalised regression methods.
3. Showcase the Bayesian paradigm as a natural framework to investigate sparsity patterns in the data.
4. Introduce the use of Bayesian model selection/averaging in the context of financial/economic applications.
5. Introduce Bayesian inference methods for generalised linear and non-linear models.

Decision-theoretic foundations

Foundations of Bayesian inference

The overarching purpose of inferential studies in Economics is to provide analytical tools to make a *decision* $d \in \mathcal{D}$.

By construction, decisions are conditional on some potentially unknown parameter θ:

$\hookrightarrow$ e.g., buying equity based on future returns, a policy intervention based on outcomes, etc.

This requires an *evaluation* criterion $L(\theta, d)$ a.k.a. a loss function.

$\hookrightarrow$ A penalty (or error) associated with the decision $d \in \mathcal{D}$ when the parameter takes the value $\theta \in \Theta$.

Foundations of Bayesian inference

Bayesian inference builds upon three key elements:

↪ A model to characterise the distribution of the observations $p\left(x|\theta\right)$

↪ A prior on the parameters $p\left(\theta\right)$

↪ A loss associated with the decision $L\left(\theta, d\right)$

Comment 1: The prior and the loss, and even sometimes the sampling distribution can come – partly or entirely – from subjective considerations.

Comment 2: One could leverage the intrinsic value of the model $p\left(x|\theta\right)$ to calibrate the prior.

↪ This is called ”objective” Bayesian inference (see Consonni et al., 2018 for a review).

↪ Caveat: the math gets way more complicated, and the impact on the loss function is not always clear (at least to me).

Foundations of Bayesian inference

The frequentist criticism often focus on the prior elicitation.

Caveat of this criticism:

- ↪ fail to take into account that constructing/evaluating the loss function could be just as complicated as deriving the prior distribution.
- ↪ presumes the expected loss is approximately equivalent to the average over i.i.d repetitions of the same experiment – mostly asymptotic inference.

But if we assume that we know the loss function and the properties of the experiment, this information could therefore be used more efficiently by building up a prior!!

Lindley (1957): a loss function and a prior are difficult to separate as implicitly they are part of the same overarching structure, and therefore should be analysed simultaneously.

Frequentist principle

Except for the most trivial settings, it is generally impossible to uniformly minimise (in $d \in \mathcal{D}$) the loss function $L\left(\theta, d\right)$ when $\theta \in \Theta$ is unknown.

The *frequentist* approach proposes to consider the average loss (or *risk*),

$$\begin{aligned} R\left(\theta, \delta\right) &= \mathbb{E}_{\theta}\left[L\left(\theta, \delta\left(x\right)\right)\right], \\ &= \int_{\mathcal{X}} L\left(\theta, \delta\left(x\right)\right) p\left(x|\theta\right) dx \end{aligned}$$

where $\delta\left(\cdot\right)$ is the decision rule (or *estimator*), so that $\delta\left(x\right)$ represents the *estimate* $\widehat{\theta}$ for the outcome $x \sim p\left(x|\theta\right)$.

Goal: select the best estimator by minimizing the risk function

Frequentist principle

Caveats associated with this approach:

1. The loss is averaged over different values of $x \propto p\left(x|\theta\right)$, i.e., x is not taken into account any further;

   $\hookrightarrow$ The risk criterion evaluates a decision rule/estimator "asymptotically".

2. The assumption of repeatability (a.k.a. iid) of experiments is not always grounded.

   $\hookrightarrow$ The same decision problem will not be necessarily met again and again.

3. $R\left(\theta, \delta\right)$ is a function of θ. At best, one can hope for a procedure δ_0 which minimize uniformly the risk, but such cases rarely occur unless the space of decisions is restricted.

Bayesian principle

Integrate over the parameter space Θ to get the posterior expected loss of $d \in \mathcal{D}$

$$\rho\left(p, d|x\right) = \mathbb{E}_p\left[L\left(\theta, d\right)|x\right] = \int_{\Theta} L\left(\theta, d\right) p\left(\theta|x\right) d\theta$$

The posterior expected loss is *conditional* on the observed value x. This is contrary to the frequentist approach which is *conditional* on θ.

$\hookrightarrow$ Recall that while x is observed, θ is unknown.

Assuming a prior $p\left(\theta\right)$, it is also possible to define the *integrated risk* as

$$r\left(p, \delta\right) = \mathbb{E}_p\left[R\left(\theta, \delta\right)\right] = \int_{\Theta}\int_{\mathcal{X}} L\left(\theta, \delta\left(x\right)\right) p\left(x|\theta\right) dx p\left(\theta\right) d\theta,$$

Bayes estimator

Theorem (Construction of Bayes estimator)

An estimator minimising $r(p,\delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes the posterior expected loss $r(p,\delta|x)$ since

$$r(p,\delta) = \int_{\mathcal{X}} \rho(p,\delta(x)|x)\, m(x)\, dx \qquad \textit{with} \qquad m(x) = \int_{\Theta} p(x|\theta)\, p(\theta)\, d\theta$$

Proof: the equality follows directly from Fubini's theorem since $L(\theta,\delta) \geq 0$,

$$\begin{aligned} r(p,\delta) &= \int_{\Theta}\int_{\mathcal{X}} L(\theta,\delta(x))\, p(x|\theta)\, dx p(\theta)\, d\theta, \\ &= \int_{\mathcal{X}}\int_{\Theta} L(\theta,\delta(x))\, p(x|\theta)\, p(\theta)\, d\theta dx, \\ &= \int_{\mathcal{X}} \underbrace{\int_{\Theta} L(\theta,\delta(x))\, p(\theta|x)\, d\theta}_{\rho(p,\delta(x)|x)} m(x)dx \quad \text{since} \quad m(x)p(\theta|x) = p(x|\theta)\, p(\theta) \end{aligned}$$

Bayes estimator

The result on the previous slide leads to the following definition of a Bayes estimator.

Definition (Bayes estimator)

A Bayes estimator associated with a prior distribution $p(\theta)$ and a loss function, is any estimator δ_p such that

$$\delta_p = \arg\min_{\delta} r(p, \delta)$$

The value $r(p, \delta_p)$ is called *Bayes risk*.

We can now define two key concepts of Bayes estimators:

- $\hookrightarrow$ Minimaxity
- $\hookrightarrow$ Admissibility

Minimaxity

Minimaxity is the frequentist "insurance" against the worst case scenario,

$$\inf_{\delta \in \mathcal{D}} \sup_{\theta} R(\theta, \delta) = \inf_{\delta \in \mathcal{D}} \sup_{\theta} \mathbb{E}_{\theta}\left[L(\theta, \delta(x))\right],$$

and a *minimax estimator* is any estimator δ_0 such that

$$\sup_{\theta} R(\theta, \delta_0) = \inf_{\delta \in \mathcal{D}} \sup_{\theta} R(\theta, \delta),$$

That is, is the minimizer of the worst-case frequentist risk. For a given prior, define the associated Bayes estimator δ_p,

$\hookrightarrow$ If $\sup_{\theta} R(\theta, \delta_p) = r(p, \delta_p)$, then δ_p can be shown to be minimax.

Admissibility

Admissibility: reduction of the set of acceptable estimators based on “local” properties.

Definition (Admissibile estimator)

An estimator δ_0 is *inadmissible* if there exists and estimator δ_1 such that, for every θ

$$R\left(\theta, \delta_0\right) \geq R\left(\theta, \delta_1\right),$$

and, for at least one θ_0

$$R\left(\theta_0, \delta_0\right) > R\left(\theta_0, \delta_1\right),$$

Otherwise δ_0 is admissible.

An estimator is admissibile with respect to a loss function if there is no other estimator that dominates it in terms of risk.

A Bayesian perspective on optimality

Why is important to think in terms of optimality with respect to a loss function?

$\hookrightarrow$ Consider a quadratic loss $L\left(\theta, d\right) = \left(\theta - d\right)^2$.

We can show that the Bayes estimator δ_p associated with the prior distribution $p\left(\theta\right)$ is the posterior expectation $\delta\left(x\right) = \mathbb{E}_p\left[\theta | x\right]$.

Proof: The first-order condition

$$\frac{\partial \mathbb{E}_p\left[\left(\theta - \delta\right)^2 | x\right]}{\partial \delta} = \mathbb{E}_p\left[\theta^2 | x\right] - 2\delta \mathbb{E}_p\left[\theta | x\right] + \delta^2 = 0$$

implies that $\delta = \mathbb{E}_p\left[\theta | x\right]$.

A Bayesian perspective on optimality

Equivalently, the optimality of the posterior mean estimate can be shown by decomposing the mean squared loss conditional on the prior,

$$\begin{aligned}\mathbb{E}_p\left[(\theta-\delta)^2 \,|x\right] &= \mathbb{E}_p\left[\theta^2|x\right] - 2\delta\mathbb{E}_p\left[\theta|x\right] + \delta^2 \pm \mathbb{E}_p\left[\theta|x\right]^2 \\ &= \underbrace{\left(\mathbb{E}_p\left[\theta^2|x\right] - \mathbb{E}_p\left[\theta|x\right]^2\right)}_{\text{Variance}} + \underbrace{\left(\mathbb{E}_p\left[\theta|x\right] - \delta\right)^2}_{\text{Bias}^2}\end{aligned}$$

The Bias is zero for $\delta = \mathbb{E}_p\left[\theta|x\right]$.

In reality, despite a zero bias, it is not always the case that the posterior mean always achieve the lowest mean-squared loss. There exists sub-optimal estimators that can achieve a lower squared loss,

↪ to the extent that a higher bias is compensated by a larger decrease in the variance.

↪ striking that balance is the ultimate goal of Bayesian model/variable selection methods.

A re-cap of Bayesian linear regressions

A re-cap of Bayesian linear regressions

Let $y \in \mathbb{R}^n$ be the vector of response variable and $X \in \mathbb{R}^{n \times p}$ the design matrix.

The regression model we build upon has the form;

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma^2 \mathbf{I}_n\right), \tag{1}$$

where $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ matrix of covariates and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the p-dimensional vector of regression parameters.

We need to postulate first a prior distribution for $\left(\beta, \sigma^2\right)$.

$\hookrightarrow$ Several alternatives: non-informative, conjugate, independent, informative (g-prior), hierarchical,...

Joint posterior distribution

A default prior for Eq.(1) is the natural conjugate prior:

$$\beta|\sigma^2 \sim N\left(\beta_0, \sigma^2\Sigma_0\right), \qquad \sigma^2 \sim \mathsf{IG}\left(v_0, s_0\right)$$

Under these priors the joint posterior distribution is defined as

$$p\left(\beta, \sigma^2|y, X\right) \propto \underbrace{\left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{2\sigma^2}\right\}}_{\text{Likelihood}} \underbrace{p\left(\beta|\sigma^2\Sigma_0\right) p\left(\sigma^2\right)}_{\text{Priors}},$$

$$\propto \left(\sigma^2\right)^{-\frac{n+p}{2}+v_0+1} \exp\left\{-\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{2\sigma^2} - \frac{s_0}{\sigma^2}\right\} \exp\left\{\frac{\left(\beta-\beta_0\right)^\top \Sigma_0^{-1}\left(\beta-\beta_0\right)}{2\sigma^2}\right\}$$

Full conditional distributions

From the linear regression model in Eq.(1), under the natural conjugate prior and the joint posterior $p\left(\beta, \sigma^2 | y, X\right)$, the full conditional posterior distributions take the form:

$$
\begin{aligned}
p\left(\beta | y, X, \sigma^2\right) &\propto N_p\left(\widehat{\Sigma}_n\left(\Sigma_0^{-1}\beta_0 + X^\top y\right), \sigma^2\widehat{\Sigma}_n\right),\\
p\left(\sigma^2 | y, X, \beta\right) &\propto \mathsf{IG}\left(v_0 + \frac{n+p}{2}, s_0 + \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{2}\right),
\end{aligned}
$$

where $\widehat{\Sigma}_n = \left(X^\top X + \Sigma_0^{-1}\right)^{-1}$ and $\boldsymbol{\varepsilon} = y - X\beta$.

So that the joint posterior $p\left(\beta, \sigma^2 | y\right)$ can be approximated via Gibbs sampler.

Digression: The Gibbs sampler

For a joint distribution $p\left(x_1, \ldots, x_p\right)$ with full conditionals $p\left(x_j | x_{-j}\right)$, the Gibbs sampler simulates successively from all conditionals modifying one component of X at a time.

Algorithm 1 Gibbs sampler

```
Start with an arbitrary value $x^{(0)} = \left(x_1^{(0)}, \ldots, x_p^{(0)}\right)$.
Iteration $t$: Given $\left(x_1^{(t-1)}, \ldots, x_p^{(t-1)}\right)$
  1. $x_1^{(t)}$ according to $p\left(x_1 | x_2^{(t-1)}, \ldots, x_p^{(t-1)}\right)$,
  2. $x_2^{(t)}$ according to $p\left(x_2 | x_1^{(t)}, x_3^{(t-1)}, \ldots, x_p^{(t-1)}\right)$,
        $\vdots$
  3. $x_p^{(t)}$ according to $p\left(x_p | x_1^{(t)}, \ldots, x_{p-1}^{(t-1)}\right)$,
```

Non-Gaussian error terms

One of the main advantages of the Bayesian paradigm is the flexibility to model non-Gaussian distributions as a simple hierarchical model.

$\hookrightarrow$ suppose we want to model a Student-t error term $\varepsilon \sim t_\nu$ for Eq.(1).

This would take the form of a scale mixture of normals $\varepsilon | \lambda \sim N\left(0, \lambda^{-1}\right)$ and $\lambda \sim G\left(\nu/2, \nu/2\right)$.

Assume for simplicity an independent prior structure as,

$$\beta \sim N\left(\beta_0, \Sigma_0\right), \qquad \sigma^2 \sim \mathsf{IG}\left(v_0, s_0\right), \qquad \lambda \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

Non-Gaussian error terms

Under $\beta_0 = 0$ the full conditional posterior distributions take the form:

$$p\left(\beta|y, X, \sigma^2, \lambda\right) \propto N\left(\widehat{\Sigma}_n\left(\sigma^{-2}X^\top \text{diag}\left(\lambda\right)y\right), \widehat{\Sigma}_n\right),$$

$$p\left(\sigma^2|y, X, \beta, \lambda\right) \propto \text{IG}\left(v_0 + \frac{n+p}{2}, s_0 + \frac{\boldsymbol{\varepsilon}^\top \text{diag}\left(\lambda\right)\boldsymbol{\varepsilon}}{2}\right),$$

$$p\left(\lambda_i|y, X, \beta, \sigma^2\right) \propto G\left(\frac{\nu+1}{2}, \frac{\nu+\sigma^{-2}\varepsilon_i^2}{2}\right), \qquad i = 1, \ldots, p$$

where $\widehat{\Sigma}_n = \left(\sigma^{-2}X^\top \text{diag}\left(\lambda\right)X + \Sigma_0^{-1}\right)^{-1}$, $\varepsilon_i = y_i - X_i\beta$, and $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \ldots, \varepsilon_n\right)^\top$.

So that the joint posterior $p\left(\beta, \sigma^2, \lambda|y\right)$ can be approximated via Gibbs sampler.

Key elements of interest for Bayesian model selection

Recall the general formulation of the Bayes theorem for a linear regression:[1]

$$p\left(\beta,\sigma^2|y\right)=\frac{p\left(y|\beta,\sigma^2\right)p\left(\beta,\sigma^2\right)}{p\left(y\right)}\propto\underbrace{p\left(y|\beta,\sigma^2\right)}_{\text{Likelihood}}\underbrace{p\left(\beta,\sigma^2\right)}_{\text{Priors}},$$

There are two key quantities of interest:

- The posterior $p\left(\beta,\sigma^2|y\right)$
- The marginal likelihood $p\left(y\right)$

While the joint posterior distribution is key for inference, the marginal likelihood is of paramount importance for model determination/assessment/selection.

[1] We omit X from the formulation for simplicity.

Marginal likelihood

In some cases, such as Eq.(1), the marginal likelihood is available in closed form:

$$p\,(y) \propto |\widetilde{X}^\top \widetilde{X}|^{-1/2} |\Sigma_0|^{-1/2} \left(s_0 + \frac{S^2}{2} \right)^{-(v_0+n/2)} \tag{2}$$

where $\widetilde{X} = \begin{bmatrix} X \\ \Sigma_0^{-1} \end{bmatrix} \in \mathbb{R}^{(n+p)\times p}$, $S^2 = y^\top H y$, with $H = I - X\widehat{\Sigma}_n X^\top$.

N.B., for complex models the marginal likelihood is not available in closed form and should be approximated numerically via Monte Carlo integration (more on Lecture 4).

This can be due to:

$\hookrightarrow$ Complex prior structure.

$\hookrightarrow$ Complex likelihood (model).

$\hookrightarrow$ Both complex prior and likelihood.

Marginal likelihood from information criteria

When the marginal likelihood cannot be easily obtained analytically or via numerical integration, a (computationally) straightforward strategy is to rely on information criteria.

For example, the Bayesian Information Criterion (BIC) is a first-order approximation to the marginal likelihood,

$$\log p\left(y\right) = \log p\left(y|\widetilde{\beta}, \widetilde{\sigma}^2\right) + \log p\left(\widetilde{\beta}, \widetilde{\sigma}^2\right) + \frac{p}{2}\log\left(2\pi\right) - \frac{p}{2}\log n \\ - \frac{1}{2}\log\left|J_n\left(\widetilde{\beta}, \widetilde{\sigma}^2\right)\right| + \mathcal{O}\left(n^{-1}\right),$$

where $J_n\left(\widetilde{\beta}, \widetilde{\sigma}^2\right)$ is the expected Fisher information evaluated at the mode $\left(\widetilde{\beta}, \widetilde{\sigma}^2\right)$. Removing any terms of order $\mathcal{O}\left(1\right)$ or less we obtain:

$$\log p\left(y\right) = \log p\left(y|\widehat{\beta}, \widehat{\sigma}^2\right) - \frac{p}{2}\log n + \mathcal{O}\left(1\right), \tag{3}$$

Marginal likelihood from information criteria

The approximation in Eq.(3) gives the basics for the BIC

$$BIC = -2\log p\left(y|\widehat{\beta}, \widehat{\sigma}^2\right) + p\log n,$$

with $\left(\widehat{\beta}, \widehat{\sigma}^2\right)$ the parameters point estimates.

An alternative popular criterion to approximate the marginal likelihood is the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002).

$$DIC = -4\mathbb{E}_{p(\beta,\sigma^2|y)}\left[\log p\left(y|\beta,\sigma^2\right)\right] + 2\log p\left(y|\widetilde{\beta}, \widetilde{\sigma}^2\right),$$

The DIC made by two components:

- The expectation of the data density w.r.t to the posterior (this can be evaluated numerically).
- The value of the data density evaluated at the posterior mode $\left(\widetilde{\beta}, \widetilde{\sigma}^2\right)$.

Marginal likelihood and model comparison

The marginal likelihood is at the core of model comparison and hypothesis testing within the context of Bayesian modeling.

Consider the case of two competing models, $\mathcal{M}_0$ and $\mathcal{M}_1$. Evidence in favour of the benchmark $\mathcal{M}_0$ corresponds to how good is its fit compared to the alternative $\mathcal{M}_1$.

This can be directly tested based on the Bayes factor,

$$BF_{01} = \frac{p\left(y|\mathcal{M}_0\right)}{p\left(y|\mathcal{M}_1\right)}, \tag{4}$$

The product of the prior odds $p\left(\mathcal{M}_0\right), p\left(\mathcal{M}_2\right)$ and the Bayes factor is the posterior odds,

$$PO_{01} = \frac{p\left(\mathcal{M}_0|y\right)}{p\left(\mathcal{M}_1|y\right)} = \frac{p\left(y|\mathcal{M}_0\right)}{p\left(y|\mathcal{M}_1\right)} \frac{p\left(\mathcal{M}_0\right)}{p\left(\mathcal{M}_1\right)},$$

If $p\left(\mathcal{M}_0\right) = p\left(\mathcal{M}_1\right) = 1/2$ the posterior odds is equal to the Bayes factor.

Marginal likelihood and model comparison

The Bayes factor in Eq.(4) is a primary tool for assessing the evidence in favour of a statistical model versus all competing alternatives.

Kass and Raftery (1995) provide a rule-of-thumb on how to interpret the statistical evidence against model 1 based on BF_{01}:

- ↪ For $BF_{01} > 3$ the evidence is substantial.
- ↪ For $BF_{01} > 10$ the evidence is strong.
- ↪ For $BF_{01} > 100$ the evidence is decisive.

Digression: The Savage-Dickey density ratio

For nested model comparisons, Verdinelli and Wasserman (1995) show that Bayes factors can be calculated using the Savage-Dickey density ration (SDDR) approach.

Consider two regression models as in Eq.(1). The first model $\mathcal{M}_0$ is unrestricted while the second model $\mathcal{M}_1$ imposes that $\beta_j = 0$ for some $j \in (1, p)$.[2]

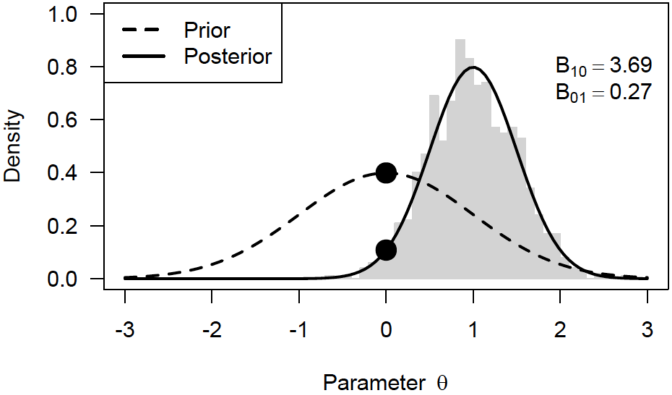
In this case the BF_{01} can be written as

$$BF_{01} = \frac{p\left(y|\mathcal{M}_0\right)}{p\left(y|\mathcal{M}_1\right)} = \frac{\int_{-\infty}^{\infty}\int_0^{\infty} p\left(y|\beta,\sigma^2,\mathcal{M}_0\right) p\left(\beta,\sigma^2|\mathcal{M}_0\right) d\beta d\sigma^2}{\int_0^{\infty} p\left(y|\beta^*,\sigma^2,\mathcal{M}_1\right) p\left(\beta^*,\sigma^2|\mathcal{M}_1\right) d\sigma^2} = \frac{\int_0^{\infty} p\left(\beta^*,\sigma^2|y,\mathcal{M}_1\right) d\sigma^2}{\int_0^{\infty} p\left(\beta^*,\sigma^2|\mathcal{M}_1\right) d\sigma^2},$$

That is, the SDDR is the ratio of the posterior and the prior of β under $\mathcal{M}_1$ evaluated at $\beta = \beta^*$. This is way easier to calculate that the marginal likelihoods.

[2] This is equivalent to test the null hypothesis $\mathcal{H}_0 : \beta_j = 0$ against the alternative $\mathcal{H}_1 : \beta_j \neq 0$.

Digression: The Savage-Dickey density ratio



The Savage-Dickey density ratio is defined as the ratio of posterior to prior density (solid and dashed line, respectively) at the critical value $\theta = 0$ (black points). The gray histogram shows the distribution of samples from the posterior distribution of the encompassing model, which are often required to approximate the marginal posterior density.

A word of caution on priors elicitation

Let $\mathcal{M}_\gamma$ denote a linear regression model which contain a subset $k < p$ of covariates $X_\gamma \in X$.

Given $p\left(y|\beta, \sigma^2\right)$ and $p\left(\beta, \sigma^2\right)$, the inferential target is the posterior model probability

$$p\left(\mathcal{M}_\gamma|y\right) \propto p\left(y|\mathcal{M}_\gamma\right) p\left(\mathcal{M}_\gamma\right), \tag{5}$$

where $p\left(y|\mathcal{M}_\gamma\right) = \int p\left(y|\beta_\gamma, \sigma^2\right) p\left(\beta_\gamma, \sigma^2\right) d\beta_\gamma d\sigma^2$, and $p\left(\mathcal{M}_\gamma\right)$ is the prior model probability, such that $\sum_\gamma^{2^p} p\left(\mathcal{M}_\gamma\right) = 1$.

Two main issues:

- $\hookrightarrow$ How do we set the priors $p\left(\beta_\gamma, \sigma^2\right)$?
- $\hookrightarrow$ How do we design the prior $p\left(\mathcal{M}_\gamma\right)$?

A word of caution on priors elicitation

With respect to $p\left(\beta_\gamma, \sigma^2\right)$ things can get quite complicated, especially when p is large. For instance, consider the natural conjugate prior

$$\beta_\gamma|\sigma^2 \sim N\left(0, \sigma^2\Sigma_0\right), \qquad \text{and} \qquad \sigma^2 \sim \mathsf{IG}\left(v_0, s_0\right),$$

Assume for simplicity $\Sigma_0 = \tau I_p$, with I_p the identity matrix.

Consider the case with $\mathcal{M}_0$ s.t. $X_\gamma = (x_1, x_2)$ and $\mathcal{M}_1$ s.t. $X_\gamma = x_1$. The posterior variance for β takes the form

$$\sigma^2\left(X_\gamma^\top X_\gamma + (\tau I_p)^{-1}\right)^{-1},$$

The impact of τ on x_1 will be identical on $\mathcal{M}_0$ and $\mathcal{M}_1$ only if $x_1 \perp x_2$!!

If $\operatorname{corr}(x_1, x_2) \neq 0$, the effect of τ on x_1 could change between the two models.

A word of caution on priors elicitation (cont'd)

In its simplest form, the prior under $\mathcal{M}_1$ and $\mathcal{M}_0$ are identical for the SDDR,

$$p\left(\beta, \sigma^2\right) = p\left(\beta, \sigma^2 | \beta_j = 0\right),$$

That is, the SDDR *"implicitly assumed that the common nuisance parameters fulfil exactly the same roles, whether they are part of $\mathcal{M}_0$ or $\mathcal{M}_1$"*.

This assumption is satisfied only if priors are independent, i.e., $p\left(\beta, \sigma^2\right) = p\left(\beta\right) p\left(\sigma^2\right)$.

$\hookrightarrow$ which is rarely the case for complex models.

In the following slides we are going to revisit an "empirical Bayes" prior which allows to set different priors for different, potentially nested, regression models.

Zellner’s g-prior

Zellner's g-prior

A default prior in linear regression models is the g-prior due to Zellner (1986),

$$\beta|\sigma^2 \sim N\left(0, g\sigma^2\left(X^\top X\right)^{-1}\right), \tag{6}$$

This somehow appears as a data-dependent prior through its dependence on X,

↪ this is not a genuine issue, since the model is conditional on X by construction.

The factor g can be interpreted as being inversely proportional to the amount of information available in the prior relative to the sample,

↪ for instance, for $g = n$, gives the prior the same weight as one observation of the sample.

The g-prior is often decomposed as a (conditional) Gaussian prior for β and an improper (Jeffreys) prior for σ^2, i.e., $p\left(\sigma^2\right) \propto \sigma^{-2}$.

Zellner's g-prior

Assuming $\beta_0 = 0$ and $p\left(\sigma^2\right) \propto \sigma^{-2}$, the joint prior distribution takes the form

$$p\left(\beta, \sigma^2\right) \propto \left(\sigma^2\right)^{-\frac{p+1}{2}} \exp\left(-\frac{1}{2g\sigma^2}\left(\beta^\top X^\top X\beta\right)\right)$$

The joint posterior distribution then takes the form

$$p\left(\beta, \sigma^2 | y\right) \propto \left(\sigma^2\right)^{-\frac{n+p+1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - X\beta\right)^\top \left(y - X\beta\right)\right) \\ \times \exp\left(-\frac{1}{2g\sigma^2}\left(\beta^\top X^\top X\beta\right)\right)$$

Zellner's g-prior

The conditional posterior distribution of the regression parameters β is[3],

$$\beta|\sigma^2, y, X, g \sim N\left(\widehat{\beta}, \widehat{\Sigma}\right)$$

where

$$\widehat{\beta} = \frac{g}{1+g}\left(X^\top X\right)^{-1} X^\top y, \qquad \widehat{\Sigma} = \frac{\sigma^2 g}{1+g}\left(X^\top X\right)^{-1},$$

Consequently, the posterior mean of the regression parameters is

$$\widehat{\beta} = \frac{g}{1+g}\widehat{\beta}^{OLS},$$

When $g \to \infty$ the posterior mean tends to the OLS estimate, while when $g \to 0$ the posterior contracts towards zero (hint: g is a shrinkage parameter!).

[3] Proof seen in class

Zellner's g-prior

The posterior distribution of σ^2 takes the form

$$\sigma^2 | y, X \sim \mathsf{IG}\left(\frac{n-1}{2}, s + \widehat{\beta}^\top X^\top X^\top \widehat{\beta} / (g+1)\right)$$

where $s = \left(y - X\widehat{\beta}\right)^\top \left(y - X\widehat{\beta}\right)$, corresponds to the (classical) residual sum of squares.

Since the prior $p\left(\beta, \sigma^2\right)$ changes for different choice of predictors X_γ, the g-prior is particularly suitable for model comparison.

Zellner's g-prior

Model comparison can be conducted using Bayes factors. In the case of linear models and under g-priors, those Bayes factors are available in closed form;

$$p\left(y|\beta_\gamma,\sigma^2\right)p\left(\beta_\gamma,\sigma^2\right) = \frac{|X_\gamma^\top X_\gamma|^{1/2}}{\left(2\pi\sigma^2\right)^{(n+p+1}/2} g^{p/2}\exp\left(-\frac{1}{2\sigma^2}\left(y-X_\gamma\beta_\gamma\right)^\top\left(y-X_\gamma\beta_\gamma\right)\right)\times$$
$$\exp\left(-\frac{1}{2g\sigma^2}\beta_\gamma^\top X_\gamma^\top X_\gamma\beta_\gamma\right),$$

Such that

$$p\left(y\right)\propto\left(g+1\right)^{-p/2}\left[s+\frac{\widehat{\beta}_\gamma X_\gamma^\top X_\gamma\widehat{\beta}_\gamma}{g+1}\right]^{-\frac{n-1}{2}},$$

represents the marginal likelihood.

Zellner's g-prior

Consider two different models $\mathcal{M}_j$ for $j = 1, 2$ with different sets of predictors $X_j \subset X$.

The Bayes factor under the g-prior is calculated as

$$BF_{12} = \frac{p\left(y|\mathcal{M}_1\right)}{p\left(y|\mathcal{M}_2\right)} = \frac{(g_1+1)^{-p_1/2}\left[s_1 + \widehat{\beta}_1^\top X_1^\top X_1 \widehat{\beta}_1/(g_1+1)\right]^{-(n-1)/2}}{(g_2+1)^{-p_2/2}\left[s + \widehat{\beta}_2^\top X_2^\top X_2 \widehat{\beta}_2/(g_2+1)\right]^{-(n-1)/2}}$$

with $p_j, \widehat{\beta}_j, s_j, g_j$ the number of predictors, the least squares projection, the value of s for $j = 1, 2$, and the scaling g respectively.

How to set g?

Summary of the g-prior structures often used in the literature

Structure	Comment
$g = n$	Unit information prior (Kass and Wasserman, 1995)
$g = \max\left(n, k^2\right)$	Benchmark prior (Fernandez et al., 2001)
$g = k^2$	Risk inflation criterion (Foster and George, 1994)
$g = \frac{1}{n}$	Opposite to Kass and Wasserman (1995)
$g = \frac{k}{n}$	More information to the prior as we add more regressors
$g = \sqrt{\frac{1}{n}}$	Smaller penalty for large models
$g = \sqrt{\frac{k}{n}}$	Fernandez et al. (2001)
$g = \frac{1}{k^2}$	Foster and George (1994)

How to set g?

Liang et al. (2008) suggest to place priors on g rather than use a fixed value.

Imposing a prior on $p\left(g\right)$ could still lead to efficient computations of the marginal likelihoods (see, e.g., Bayarri et al., 2012). Each of these papers study priors of the form

$$p\left(g\right) \propto g^{d}\left(g+b\right)^{-(a+c+d+1)},$$

with $a > 0, b > 0, c > -1$, and $d > -1$.

Specific configurations recommended in the literature include: $\{a = 1, b = 1, d = 0\}$ (see, Cui and George, 2008), $\{a = 1/2, b = 1, c = 0, d = 0\}$ (see, Liang et al., 2008), and $\{c = -3/4, d = (n-5)/2 - p_{\gamma}/2 + 3/4\}$ (see, Maruyama and George, 2011).

How to set g?

A subjective Bayes approach to set g does not guarantee model selection consistency.

To avoid the information paradox, we need the Bayes factor to diverge as $R^2 \to 0$. This is equivalent to

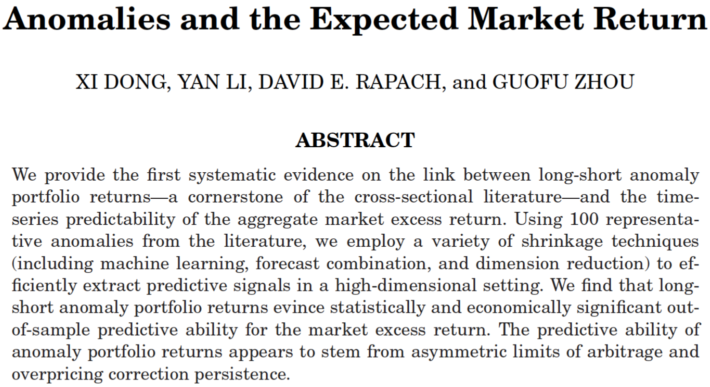
$$\int (1+g)^{(n-1-p)/2} \, p(g) \, dg = 0$$

One choice of $p(g)$ that satisfies this, is given by

$$p(g) \propto g^{-3/2} \exp\left(-n/2g\right), \qquad \text{that is} \qquad g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$$

This makes the computation way easier since g can be sampled from an inverse-Gamma.

Example: Anomalies and the expected market returns



Anomalies and the Expected Market Return

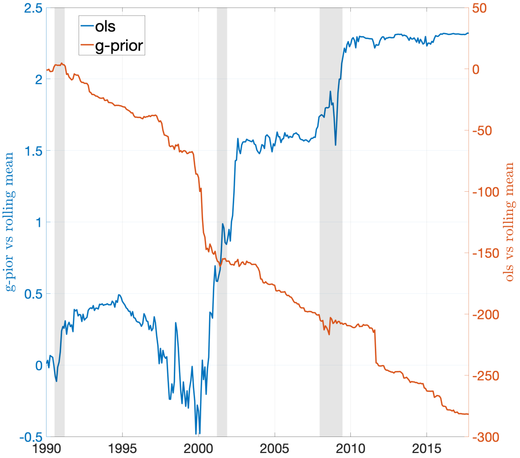
XI DONG, YAN LI, DAVID E. RAPACH, and GUOFU ZHOU

ABSTRACT

We provide the first systematic evidence on the link between long-short anomaly portfolio returns—a cornerstone of the cross-sectional literature—and the time-series predictability of the aggregate market excess return. Using 100 representative anomalies from the literature, we employ a variety of shrinkage techniques (including machine learning, forecast combination, and dimension reduction) to efficiently extract predictive signals in a high-dimensional setting. We find that long-short anomaly portfolio returns evince statistically and economically significant out-of-sample predictive ability for the market excess return. The predictive ability of anomaly portfolio returns appears to stem from asymmetric limits of arbitrage and overpricing correction persistence.

Goal: Forecasting one-month ahead equity premium (aka aggregate stock market returns minus risk free rate) based on current returns on a large set ($p = 100$) of anomaly portfolios. Sample period is from 1970-2017, monthly.

Example: Anomalies and the expected market returns



Cumulative squared error differentials $\Delta\text{CumSSE}_{t,s} = \sum_{\tau=\underline{t}}^{t}(\overline{e}_{\tau})^2 - \sum_{\tau=\underline{t}}^{t}(\widehat{e}_{\tau,s})^2$, where $\underline{t}$ is the first prediction time, $\overline{e}_{\tau} = \sum_{i=1}^{n}(y_{i\tau} - \overline{y}_{i\tau})$ and $\widehat{e}_{\tau,s} = \sum_{i=1}^{n}(y_{i\tau} - \widehat{y}_{i\tau}(\mathcal{M}_s))$ are the prediction error from the naive rolling mean benchmark and a given competing model $\mathcal{M}_s$ at time τ, respectively.

A primer on Bayesian model selection

Introduction to Bayesian model selection

The g-prior addresses the issue of setting a prior on different regression models that might be nested and have correlated variables.

Another important issue is to set a prior on the model space.

Recall that we can index all possible 2^p models using dummy variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$ signifying which predictors are included in the regression:

- $\hookrightarrow$ When $\gamma_j = 0$ a covariate is excluded from a model and when $\gamma_j = 1$ is included.
- $\hookrightarrow$ The model with no predictors is $\boldsymbol{\gamma} = (0, \ldots, 0)$, while the model with all predictors is $\boldsymbol{\gamma} = (1, \ldots, 1)$.

The advantage is that now instead of placing a prior on the model space, we can place a prior on the vector γ_j.

Bayesian model selection

Each of the 2^p models under comparison is associated with a binary indicator vector

$\hookrightarrow$ $\boldsymbol{\gamma} = (1, 0, 1, 0, 0, \ldots, 1, 0)$ clearly indicates which explanatory variables are relevant and which ones are not.

For instance, $q_\gamma = \mathbf{1}_p^\top \boldsymbol{\gamma}$ indicates the number of variables included in the model $\mathcal{M}_\gamma$.

We define as β_γ and X_γ the sub-vector of β and the sub-matrix of X containing only the components pertaining model $\mathcal{M}_\gamma$.

The model $\mathcal{M}_\gamma$ is thus defined as

$$y | \beta_\gamma, \sigma^2, \boldsymbol{\gamma} \sim N\left(X_\gamma \beta_\gamma, \sigma^2 \mathbf{I}_n\right),$$

Bayesian model selection with g-prior

Because the model space is potentially large, we cannot expect one to specify his/her own prior on every possible model $\mathcal{M}_\gamma$ in a completely and subjective manner.

A convenient approach is to derive *all* priors from a single global prior associated with the so-called *full* model that corresponds to $\boldsymbol{\gamma} = (1, \ldots, 1)$. The argument goes as follows:

For the full model, we use the g-prior as defined above:

$$\beta|\sigma^2 \sim N\left(\beta_0, g\sigma^2\left(X^\top X\right)^{-1}\right), \qquad \text{and} \qquad p\left(\sigma^2\right) \propto \sigma^{-2},$$

For each sub-model $\mathcal{M}_\gamma$, the prior distribution on β_γ conditional on σ^2 is fixed as

$$\beta_\gamma|\sigma^2, \boldsymbol{\gamma} \sim N\left(\beta_{0,\gamma}, g\sigma^2\left(X_\gamma^\top X_\gamma\right)^{-1}\right),$$

where $\beta_{0,\gamma} \in \beta$ is the subset of prior coefficients selected from $\boldsymbol{\gamma}$.

Bayesian model selection with g-prior

Although there are many possible choices of defining the prior on the model index $\boldsymbol{\gamma}$, the convention if often to opt for the uniform prior

$$p\left(\boldsymbol{\gamma}\right) \equiv 2^{-p},$$

This assumption on the prior of the model space makes the posterior distribution of $\boldsymbol{\gamma}$ fairly easy to compute; it is simply proportional to the marginal density of y in $\mathcal{M}_\gamma$.

$$\begin{aligned} p\left(\boldsymbol{\gamma}|y\right) &\propto p\left(y|\boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right) \propto p\left(y|\boldsymbol{\gamma}\right), \\ &\propto \left(g+1\right)^{-(q_{\boldsymbol{\gamma}}+1)/2} \left[y^\top y - \frac{g}{g+1} y^\top X_\gamma \left(X_\gamma^\top X_\gamma\right)^{-1} X_\gamma^\top y\right]^{-(n-1)/2} \end{aligned}$$

When the number of predictors is moderate, say less than 15 ($2^{15} = 32768$), the exact calculation can be undertaken.

Gibbs sampler to estimate γ

Let $\boldsymbol{\gamma}_{-j}\,(1 \leq j \leq p)$ be the vector $(\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$, the full conditional distribution $p\left(\gamma_j | y, \boldsymbol{\gamma}_{-j}\right)$ is proportional to $p\left(\boldsymbol{\gamma} | y\right)$.

A conventional Gibbs sampler takes the form:

Algorithm 2 Gibbs sampler for variable selection

```
Draw $\boldsymbol{\gamma}^{(0)}$ from the prior distribution
Iteration $t$: Given $\left(\gamma_1^{(t-1)}, \ldots, \gamma_p^{(t-1)}\right)$
    1. $\gamma_1^{(t)}$ according to $p\left(\gamma_1 | y, \boldsymbol{\gamma}_{-1}\right)$,
    2. $\gamma_2^{(t)}$ according to $p\left(\gamma_2 | y, \boldsymbol{\gamma}_{-2}\right)$,
       $\vdots$
    3. $\gamma_p^{(t)}$ according to $p\left(\gamma_p | y, \boldsymbol{\gamma}_{-p}\right)$,
```

Gibbs sampler to estimate γ

After a large number of iterations of this algorithm, its output can be used to approximate the posterior probabilities $p\left(\boldsymbol{\gamma}|y, X\right)$ by averaging the output of the Gibbs sampler,

$$\widehat{p}\left(\boldsymbol{\gamma}|y\right) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \boldsymbol{\gamma}^{(t)},$$

where T_0 first values are eliminated as *burn-in*.

The output can also be used to approximate the inclusion of a given variable $p\left(\gamma_j|y, X\right)$, as

$$\widehat{p}\left(\gamma_j|y\right) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \gamma_j^{(t)},$$

Choosing the prior on the model space

With respect to the priors over the model $\mathcal{M}_\gamma$, another popular starting point is

$$p\left(\boldsymbol{\gamma}|\phi\right) = \phi^{p}\left(1-\phi\right)^{p-p_\gamma},$$

where p_γ is the number of covariates in $\mathcal{M}_\gamma$, and the hyper-parameter $\phi \in (0,1)$ has the interpretation of the common probaility that a given variable is included.

Among the most popular default choices for ϕ are:

$\hookrightarrow$ $\phi = 0.5$, which assigns equal probability to each model, i.e., $p\left(\mathcal{M}_\gamma\right) = 1/2^p$.

$\hookrightarrow$ $\phi \sim U\left(0,1\right)$, giving equal probability to each possible number of covariates.

In general, fixed values of ϕ have been shown to perform poorly in controlling for the occurrence of spurious explanatory variables.

$\hookrightarrow$ Ley and Steel (2009) consider $\phi \sim \text{Beta}\left(1,b\right)$, i.e. a binomial-beta prior.

Reversible-jump Markov chain monte carlo

When p is small, the enumeration of all 2^p competing models is possible and the Bayesian paradigm selects the model having the largest evidence measured by the posterior,

$$\widehat{\gamma} = \arg\max_{\gamma} m\left(\gamma | y, X_\gamma\right),$$

However, when p is large, requires sampling from the posterior $m\left(\gamma | y, X_\gamma\right)$.

This can be implemented using reversible-jump (RJ) type algorithms (see Green, 1995).

Two key ingredients of the RJ strategy:

$\hookrightarrow$ the description of the possible moves from one model γ to another γ'.

$\hookrightarrow$ the determination of the acceptance probability for the move.

Reversible-jump Markov chain monte carlo

The description of the possible moves takes the form of a proposal distribution

$$\gamma^* \sim q\left(\gamma^* | \gamma\right)$$

In its simplified form, $q\left(\gamma^* | \gamma\right)$ can be,

- ↪ Addition of a regressor $\gamma_j^* = 1$ if $\gamma_j = 0$.
- ↪ Deletion of a regressor $\gamma_j^* = 0$ if $\gamma_j = 1$.

Then the sampled model indicator is accepted with a probability

$$\alpha\left(\gamma^* | \gamma\right) = \min\left\{1, \frac{m\left(\gamma^* | y, X_{\gamma^*}\right)}{m\left(\gamma | y, X_{\gamma}\right)}\right\} = \min\left\{1, BF_{\gamma^*, \gamma}\right\}$$

References

Bayarri, M., Berger, J., Forte, A., and Garcıa-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection1. *The Annals of Statistics*, 40(3):1550–1577.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior distributions for objective bayesian analysis. *Bayesian Analysis*.

Cui, W. and George, E. I. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of applied econometrics*, 24(4):651–674.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2):187–192.

Maruyama, Y. and George, E. I. (2011). Fully bayes factors with a generalized g-prior. *The Annals of Statistics*, 39(5):2740.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Verdinelli, I. and Wasserman, L. (1995). Computing bayes factors using a generalization of the savage-dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.