DECLARATORIA SULLA TESI DI DOTTORATO
Da inserire come prima pagina della tesi

Il/la sottoscritto/a

COGNOME Cozzi

NOME Mattia

Matr. 1094879

Titolo della tesi:

Hierarchical Models and Information Measures

Dottorato di ricerca in Statistica

Ciclo XXI

Tutor del dottorando Prof. Pietro Muliere

Anno di discussione 2010

DICHIARA

sotto la sua responsabilità di essere a conoscenza:

1) che, ai sensi del D.P.R. 28.12.2000, N. 445, le dichiarazioni mendaci, la falsità negli atti e l'uso di atti falsi sono puniti ai sensi del codice penale e delle Leggi speciali in materia, e che nel caso ricorressero dette ipotesi, decade fin dall'inizio e senza necessità di nessuna formalità dai benefici previsti dalla presente declaratoria e da quella sull'embargo;

2) che l'Università ha l'obbligo, ai sensi dell'art. 6, comma 11, del Decreto Ministeriale 30 aprile 1999 prot. n. 224/1999, di curare il deposito di copia della tesi finale presso le Biblioteche Nazionali Centrali di Roma e Firenze, dove sarà consentita la consultabilità, fatto salvo l'eventuale embargo legato alla necessità di tutelare i diritti di enti esterni terzi e di sfruttamento industriale/commerciale dei contenuti della tesi;

3) che il Servizio Biblioteca Bocconi archivierà la tesi nel proprio Archivio istituzionale ad Accesso Aperto e che consentirà unicamente la consultabilità on-line del testo completo (fatto salvo l'eventuale embargo);

4) che per l'archiviazione presso la Biblioteca Bocconi, l'Università richiede che la tesi sia consegnata dal dottorando alla Società NORMADEC (operante in nome e per conto dell'Università) tramite procedura on-line con contenuto non modificabile e che la Società Normadec indicherà in ogni piè di pagina le seguenti informazioni:

- tesi di dottorato*)*
  Hierarchical Models and Information Measures ;

- di *Cozzi Mattia* ;
- discussa presso l'Università commerciale Luigi Bocconi – Milano nell'anno 2010*;*
- La tesi è tutelata dalla normativa sul diritto d'autore (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche). Sono comunque fatti salvi i diritti dell'Università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte;
- **solo nel caso sia stata sottoscritta apposita altra dichiarazione con richiesta di embargo**: La tesi è soggetta ad embargo della durata di ..... mesi (indicare durata embargo);

5) che la copia della tesi depositata presso la NORMADEC tramite procedura on-line è del tutto identica a quelle consegnate/inviate ai Commissari e a qualsiasi altra copia depositata negli Uffici dell'Ateneo in forma cartacea o digitale e che di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;

6) che il contenuto e l'organizzazione della tesi è opera originale realizzata dal sottoscritto e non compromette in alcun modo i diritti di terzi (legge 22 aprile 1941, n.633 e successive integrazioni e modifiche), ivi compresi quelli relativi alla sicurezza dei dati personali; che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura, civile, amministrativa o penale e sarà dal sottoscritto tenuta indenne da qualsiasi richiesta o rivendicazione da parte di terzi;

7) **scegliere l'ipotesi 7a o 7b indicate di seguito**:

7a) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati; non è oggetto di eventuali registrazioni di tipo brevettale o di tutela, e quindi non è soggetta a embargo;

Data, 3 Febbraio 2010

F.to (indicare nome e cognome) Mattia Cozzi

Introduction

This work is meant to explore some applications of information-theoretical concepts in Statistics. As it will emerge through the chapters, "'capturing the intangible concept of Information"', as an article by E. Soofi titles, is quite a difficult task. Information is a rich concept, and it shares common ground with other important words in Statistics, like those of Uncertainty, Dependence and Utility. The difficulty in taming this concept for statistical purposes is also evident from the several definitions of Information that have been put forward in the course of the history of Statistics: surely it is not needed to recall the famous and widely used Fisher Information, but the Minimum Discrimination Information proposed and developed by Kullback or the concepts of Entropy and Mutual Information introduced in Probability and Statistics by Shannon - who is considered, with Wiener, the founder of Information Theory - may not be so widely known to the statistical audience. Last but not least it is not possible not to mention a huge field of research like that stimulated by the Maximum Entropy Principle, whose 'father' is Jaynes and whose 'progeny' is numerous and flourishing. As these few considerations will have probably made clear, when dealing with Information theory and Statistics, we are not faced with a monolitic block of well established theory: bridges between the two disciplines were built in the recent past, are being built nowadays and they will surely keep on being built in the future. This durable confront has contributed to the emerging of new approaches to old problems and it has stimulated research in new fields, but most of all it has been and still is an unvaluable tool in understanding what Information really is: the final definition

of this fundamental quantity in Statistics could not emerge but from a mosaic of linked fragments, highlighting its multidimensional and everchanging face.

In this work attention will be focussed on a particular problem: the choice among different hierarchical models via information measures. As it will be made clear later, this topic brings down together different research fields, and precisely: comparison of experiments, hierarchical model theory and analysis of dependence.

A bayesian perspective will be adopted throughout most of the analysis, so that the topic treated could be also named bayesian comparison of hierarchical experiments. Alternatively it could also be considered as an attempt to study the information flows over particular "channels", as it is usually called a statistical model in Information Theory.

The first chapter will present some basic concepts in Information Theory, their main properties, together with their interpretations in statistical terms, needed in the following chapters to better understand the approach adopted for the problem at hand.

A second chapter will be devoted to the exposition of hierarchical models, with some previously obtained results about the information flows along them.

We will then introduce what is the main problem treated in this work - the Allocation problem - which is basically an experimental design problem.

It will clearly emerge how its structure is typically hierarchical: allocations, as explained in the following, are hierarchical models, and the Allocation problem is thus a matter of choosing between these hierarchical structures.

In line with the Information-Theoretical approach embraced here, we propose to evaluate relative merits of these models by the - opportunely defined - amount of information they are able to provide about the parameters of interest: the optimal model will be the one proving to be more informative with respect to some chosen measure of Information.

In particular, throughout this work the assumed measure is Mutual Information.

It seems however that previous results about information flows in hierarchical models

cannot be directly applied in this framework: briefly, they generally assume a fixed hierarchical model why our aim is to compare different models.

We will then deal with some specific 2-level hierarchical models: a Normal model and a model with Bernoulli observables.

Even with these basic specifications, computational tools become necessary to evaluate the corresponding Information quantities, and some partial conclusions will be derived.

Finally we will briefly face the problem of choosing a model once observations are fixed, that is the traditional model selection issue. With the concepts and conclusions from the previous chapters, a question seems natural, that is, whether models can be chosen coherently on the basis of the same criterion used for experimental design. Modifications of the criterion prove to be necessary. But our efforts will only be a proposal which still has to be investigated.

Contents

Chapter 1

Main Information-theoretical concepts

It is very easy but also almost useless - since obviuos - to state that the ultimate goal of a statistical analysis is to gain information about the world, or, more precisely, about a more or less well-precised phenomenon of interest. Usually the statistician performs an experiment whose outcome will increase his knowledge about the phenomenon.

It is much more difficult to evaluate precisely, if ever possible, how much the statistician has gained by the experiment. This attempt requires an accurate, coherent and possibly operative definition of Information.

In the following, the main concepts of Information Theory will be presented as possible tools to formulate an answer to the question raised above.

Now recognized as a branch of Probability Theory, this discipline was born approximately in the middle of the 20th century as an effort to provide a formal, mathematical framework to treat problems of signal transmission along Communication channels. And since then it has undergone a rapid and tumultuous development, stimulated by the interest of mathematicians and communication engineers, involved in the creation of many of the media network we use nowadays.

Notwithstanding its fast evolution in less more than 50 years, it is quite noticeable that almost all of the ideas and concepts at the basis of the modern Theory of Information were originally introduced by the fundational works of Shannon [59] and, almost at the same time, of Wiener [66], in the treatment of automatic control.

However, in no case, its applicability remain limited to communication problems, in strict sense; in fact, some primitive attempts to study human languages through its concepts were already present in the first papers on the subject [59], even if they were intended as illustrative examples and attention was focussed on the transmission problem, showing, or at least suggesting, its versatile nature.

Moreover the abstract setting used in the original formulation and the acknoledgement of the statistical nature of the process of communication has made Information Theory applicable outside the original field where it was developed.

Nowadays Information Theory exhibits links with many other disciplines, other with respect to Communication Theory and Probability Theory, such as Computer Science, Mathematics, Physics, even Economics, and obviously Statistics [13].

At this point a premise is necessary: by means of mathematical reasoning, it has been possible to arrive at a precise and coherent definition of the quantity of Information generated by a statistical source (a random variable in information-theoretical jargon) or transmitted along a communication channel (an analog of a statistical model): in other terms, a measure of the **amount** of information contained in a message has been constructed.

As Renyi [53] points out quite simply but vividly, from a mathematical point of view, the answers to the questions "'Do you like cheese?"' and "'Would you marry me?"' provide essentially the same amount of Information, as it will be clear later. But who could assign the same value to them? To be plain, a mathematical theory capable of capturing the

quality of Information is much more difficult to develop, if ever possible, than the present one, dealing 'only' with its quantity.

For these reasons, in the following, expressions like 'facing an amount of uncertainty' or 'gaining an amount of information' will frequently be abbreviated to 'facing an uncertainty of' or 'gaining ... units information'.

1.1 A measure of Uncertainty and Information

In the beginning, we will adopt a framework consisting of discrete random variables's (r.v., in the following) assuming a finite number of values only.
Even if initially this choice has the drawback of limiting the range of applicability for the conceptual tools, it is actually the original environment in which entropy was defined and it makes it easier to grasp intuition into the subject. It will be made clear later for which concepts the generalization to the infinite and continuous cases can take place properly and where instead some caution is needed.

Observation 1.1.1 *For the purposes at hand, it will be sufficient to represent a r.v. X assuming the values $(x_1, \ldots, x_n)$ just with the vector of the corresponding probabilities* $\mathbf{p} = (p_1, \ldots, p_n)\Big(p_i \geq 0,\, i = 1, \ldots, n,\, \sum_{i=1}^{n} p_i = 1\Big)$*, where obviously $p_i = P(X = x_i)$, i.e. with its distribution.*

Definition 1.1.1 *The* ***entropy*** *of a r.v. X assuming the values $(x_1, \ldots, x_n)$ according to the probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ $\Big(p_i \geq 0,\ i = 1, \ldots, n,$ $\sum_{i=1}^{n} p_i = 1\Big)$, or the* ***entropy of the distribution*** $\mathbf{p}$ *is defined as*

$$H(X) = H(\mathbf{p}) = -\sum_{i=1}^{n} p_i log_2 p_i = \sum_{i=1}^{n} p_i log_2 \frac{1}{p_i} \tag{1.1}$$

where the convention $p\ log_2 \frac{1}{p} = 0$ whenever $p = 0$ is adopted by continuity.

Observation 1.1.2 *For its basic role in Information Theory, the* ***entropy of a binary r.v.*** *with probability of 'success' p is usually denoted with a simplified notation, i.e.*

$$H(p, 1-p) = H(p) \tag{1.2}$$

As it can be easily seen this is a concave and symmetric function of p, assuming its maximum value at $p = \frac{1}{2}$.

The adoption of entropy has emerged both since its occurence as a central quantity in the solution of many specific and fundamental problems in IT and since it satisfies some postulates, i.e. it exhibits some properties, that seem intuitively reasonable to require for a measure of Information and Uncertainty. Renyi [52] summarizes the two reasons in what he calls respectively the *pragmatic approach* and the *axiomatic approach*; notwithstanding its importance for concrete and, at the same time, fundational aspects - regarding, for example, Coding - we will not go into details with the former but just scratch its surface with an example, later.

Instead we will follow the latter approach and present the following theorem by Khinchin [36] that formalizes more precisely a result already obtained by Shannon in his seminal paper [59].

Theorem 1.1.1 *Suppose the function $H(p_1, \ldots, p_n)$ is defined for any integer $n \geq 1$ and for any $(p_1, \ldots, p_n)$ with $p_i \geq 0$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} p_i = 1$, it is continuous in every p_i for $i = 1, \ldots, n$, and further it satisfies the following postulates: for every n,*

1. $H(\frac{1}{n}, \ldots, \frac{1}{n}) = max_{(p_1,\ldots,p_n)} H(p_1, \ldots, p_n)$
2. $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$
3. $H((X, Y)) = H(X) + \sum_{i=1}^{n} p_i H(Y|X = x_i) = H(X) + H(Y|X)$
   where (X, Y) is the random vector that assumes values in the set $\{(x_i, y_j), i = 1, \ldots, n, j = 1, \ldots, m\}$ with joint probabilities $\pi = (\pi_{ij})_{ij}$, $p_i = \sum_{j=1}^{m} \pi_{ij}$ the marginal probabilities of X, $H(X) = H(p_1, \ldots, p_n)$, and, for every $i = 1, \ldots, n$, $H(Y|X = x_i)$ is the function H evaluated at $(q_{1|i}, \ldots, q_{m|i})$, the conditional distribution of Y given $X = x_i$

Then

$$H(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i log_2 p_i \tag{1.3}$$

To be more adherent to the notation of the theorem, condition 3 should have been formulated as

$$\begin{aligned} H(\pi_{11}, \ldots, \pi_{n1}, \pi_{12}, \ldots, \pi_{n2}, \ldots, \pi_{1m}, \ldots, \pi_{nm}) = \\ &= H(p_1, \ldots, p_n) + \sum_{i=1}^{n} p_i H(q_{1|i}, \ldots, q_{m|i}) \\ &= H(\mathbf{p}) + \sum_{i=1}^{n} p_i H(\mathbf{q}_i) \end{aligned}$$

but the originally adopted expression appears to be more suggestive, leading immediately to an easy interpretation of the postulates, without loosing too much mathematical precision.

In the expression of H, K is an arbitrary constant, essentially defining the unit of measurement: if $K = 1$, $\log_e 2$ $(\ln 2)$, $\log_{10} 2$ respectively, Information is measured in **bits**

(*binary digits*), **nats** (*natural digits*) or *decimal digits* (or **Hartleys**)[30].

So conversion from one unit of measurement to another can simply be made by multiplication by an opportune constant.

Since conditional entropy emerged as an almost natural byproduct in the preceding theorem, it calls for a precise

Definition 1.1.2 *Consider a random vector (X, Y) with joint probabilities $\pi = (\pi_{ij})_{ij}$, $\mathbf{p} = (p_1, \ldots, p_n)$ marginal distribution of X and, for every $i = 1, \ldots, n$, $\mathbf{q_i} = (q_{1|i}, \ldots, q_{m|i}) = (\frac{\pi_{i1}}{p_i}, \ldots, \frac{\pi_{im}}{p_i})$ is the conditional distribution of Y given $X = x_i$. Then the* ***conditional entropy of Y given X*** *is*

$$H(Y|X) = -\sum_{i=1}^{n} p_i \Big\{ \sum_{j=1}^{m} q_{j|i} \log_2 q_{j|i} \Big\} = \sum_{i=1}^{n} p_i H(Y|X = x_i) = \sum_{i=1}^{n} p_i H(\mathbf{q}_i) \qquad (1.4)$$

Note that $H(Y|X)$ is the expected value of a r.v. assuming value $H(Y|X = x_i)$ with probability p_i, $i = 1, \ldots, n$.

Obviously, inverting the roles of X and Y, we could have defined the conditional entropy of X given Y.

Some comments on the preceding postulates could make them closer to intuition.

- *Probabilistic Entropy* - First of all, there is a somehow implicit postulate imposing H to be a function of the probabilities only, i.e. of the distribution of the r.v.. It makes Entropy a measure of *probabilistic dispersion*. It states an egalitarian evaluation of the possible outcomes of the r.v: specific and relative features of the outcomes, others than their probabilities of occurence (for example, real values attached to outcomes or relative distances betweeen them, if defined) are assumed irrelevant. Thus it qualifies Entropy as a way of expressing dispersion for variables for which Variance is not defined - qualitative r.v.'s, for example - and it calls for some thinking about the difference between the concepts of Uncertainty and Dispersion.

- *Continuity* - It is the more technical one and it is assumed for mathematical convenience: it makes the resulting function a nicer object to be manipulated.

- *Equiprobable events lead to Max Uncertainty* - Postulate 1 formalizes the simple idea that any unbalance among the probabilities of the outcomes obvioulsy gives information about which ones are more likely to happen. This can be made clear by an example. Consider the following three (dichotomous) distributions, on an hypothetical $\{0, 1\}$ sample space:

$$\mathbf{p} = (0.01,\ 0.99)\ \mathbf{q} = (0.25,\ 0.75)\ \mathbf{r} = (0.5,\ 0.5)$$

  Anyone attributes greater uncertainty as we move through these: according to $\mathbf{p}$, it is almost sure that a '1' will be observed, while basing our judgement on $\mathbf{q}$ the same result seems still the favorite one but we do not feel as safe as before in putting a stake on it; finally, if $\mathbf{r}$ is considered, our guess is blind.
  To go back to the general case, postulate 1 just makes precise the fact that the more the probabilities of a distribution assume similar values the less information we have to discriminate the more likely outcomes, and the greater our uncertainty is.

- *Impossible-event Indifference* - Postulate 2 translates into mathematical terms the reasonable consideration that adding impossible events to the scheme does not modify our evaluation of the situation: mere existence of some other outcomes about whose non-occurence we are certain does not alter our opinion about the uncertainty we are facing.

- *Additivity* - Finally postulate 3 states a sort of coherence between two ways of computing the uncertainty in a random entity, one that directly considers the probabilities of the final outcomes and the other one that looks more deeply in the "'process"' that leads to them.

A simple example may clarify the concept. Suppose we are first faced with the outcomes (x_1, x_2) with probabilities $(\frac{1}{4}, \frac{3}{4})$, and if x_1 occurs with the "'second stage"' oucomes (y_1, y_2) with probabilities $(\frac{1}{2}, \frac{1}{2})$ while if x_2 occurs with (y_3, y_4) and probabilities $(\frac{1}{3}, \frac{2}{3})$. In turn we could be faced directly with the following one: (y_1, y_2, y_3, y_4) with $(\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2})$. Since the probabilities of the final outcomes are the same in both schemes, the postulate requires the value of the uncertainty to be the same in both cases: that is, it must be independent of the way the final outcomes are reached.

Observation 1.1.3 *There is a precise way in which Entropy quantifies the above mentioned 'probabilistic dispersion' of a r.v. It refers to what is called '**Almost Equipartition Property**' (AEP): AEP states that, when observing long sequences of outcomes from a r.v. (i.e. for large sample size, N large), we are almost certain to observe a sequence from the '**Typical set**', a set whose elements are characterized by means of H(X), the entropy value of the source. These outcomes all have almost the same probability of occurence, and their probabilities almost sum up to unity.*

1.1.1 Properties of Entropy and Conditional Entropy

Some very useful properties of the quantities defined above can be easily verified. We state them here for future reference.

Theorem 1.1.2 *For a random vector (X, Y), with X and Y assuming with positive probability n and m values respectively, we have*

1. $0 \leq H(X) \leq \log_2 n$
2. $H(X) = 0$ *iff* X *assumes with probability 1 a single value, i.e.* $p_i = 1$ *for some* $i = 1, \ldots, n$ *and* $p_j = 0$ *for* $j \neq i$
3. $H(X) = \log_2 n$ *iff* $\mathbf{p} = (\frac{1}{n}, \ldots, \frac{1}{n})$

4. $0 \leq H(Y|X) \leq H(Y) \leq \log_2 m$

5. $H(Y|X) = H(Y)$ *if* $X \perp Y$

6. *If* $Y = f(X)$ *for some one-to-one function* f *on* $\mathcal{X}$*, then* $H(Y) = H(X)$.

Thus **Uncertainty**, as measured by Entropy, is **always non-negative**: we can face a positive degree of uncertainty ($H > 0$) or, at best, we can be certain and face no uncertainty($H = 0$).

This latter case actually occurs if there is **no randomness** in the mechanism producing the oucomes: only one result is possible.

If we are certain that one among n outcomes will occur in an experiment, we can face at most an uncertainty of $\log_2 n$ when all the events are **equiprobable**: no one is more likely than any other. How to guess the resulting one?

Furthermore and notably, learning the outcome of another experiment X will, on average, reduce the uncertainty we feel about Y, in which we are interested. More synthetically, **conditioning reduces uncertainty** (in entropic terms). The condition '**on average**' is necessary; it is *not true* that

$$H(Y|X = x_i) \leq H(Y)$$

for every $i = 1, \ldots, n$. To be convinced about the truthfullness of this statement just consider the following

Example 1.1.1 *(A surprising event)*
Suppose that the distribution of Y *is* $\boldsymbol{p} = (0.05, 0.05, 0.9)$ *and that after observing* $X = x_1$ *it is updated to* $\boldsymbol{p}^* = (0.5, 0.5, 0)$*: we obtain* $1 = H(\boldsymbol{p}^*) \geq H(\boldsymbol{p}) = 0.569$*, that is* $1 = H(Y|X = x_1) \geq H(Y) = 0.569$*. Thus uncertainty increased after learning the value of* X*, and we ended up as if the experiment had made us more confused.*

Since $H(Y|X)$ is the mean value of the r.v assuming value $H(Y|X = x_i)$ with probability p_i $i = 1, \ldots, n$, by the property of internality of the mean function, what we can state is that there is *at least one i - one outcome - for which uncertainty is reduced*. And this single case or the cases in which entropy is reduced are *of significant weight in probabilistic terms*.

The entropy of a r.v. is unaffected by the knowledge of the outcomes of another r.v., if they are **indepedent**.

Property 6 establishes **invariance** of Entropy with respect to one-to-one transformations.

Finally let us state the following result - whose proof we provide to highlight some basic tecniques in IT - that provides a valuable tool to look at entropic relations between collections of r.v.'s.

Lemma 1.1.1 ***Chain rule for Entropy***

Let $\mathbf{X} = (X_1, \ldots, X_p)$ *be a random vector of discrete r.v.'s, then*

$$H(\mathbf{X}) = H(X_1, \ldots, X_p) = \sum_{k=1}^{p} H(X_k|X_1, \ldots, X_{k-1}) \tag{1.5}$$

where obviulsy $H(X_1|X_0) = H(X_1)$

Before going through the proof, let us make the simple

Observation 1.1.4 *Once Entropy has been accepted as a measure of information, the interpretation of the addends in (1.5) is quite spontaneous: each one is the additional information content provided by the 'conditioned' r.v., taken into account the amount already available in terms of the 'conditioning' r.v.'s. The fact that* $H(X_2|X_1) \leq H(X_2)$ *suggests the idea that, once* X_1 *is known, the informational contribute of* X_2 *is not completely new: part of it was already included in* X_1*. And the shared content is exactly* $I(X_1, X_2)$*, the mutual information between the two r.v.'s. This kind of reasoning is very*

familiar in Statistics: think about correlation between two r.v.'s as a shared explicative power.

Proof A simple proof by induction starts by recalling that, by theorem 1.1.1, Entropy already satisfies the relation of the theorem with $p = 2$, that is for the marginal distribution of (X_1, X_2), for example. That is just the Additivity property required to Entropy. Consider now the random vector $(X_1, \ldots, X_p, X_{p+1})$ and suppose that (1.5) is true for p; we have

$$\begin{aligned} H((X_1, \ldots, X_p), X_{p+1}) &= H(X_1, \ldots, X_p) + H(X_{p+1}|X_1, \ldots, X_p) \\ &= \sum_{k=1}^{p} H(X_k|X_1, \ldots, X_{k-1}) + H(X_{p+1}|X_1, \ldots, X_p) \\ &= \sum_{k=1}^{p+1} H(X_k|X_1, \ldots, X_{k-1}) \end{aligned} \tag{1.6}$$

where the first equality follows by additivity and the fact that $(X_1, \ldots, X_p)$ can just be interpreted as a r.v. with outcomes $(x_1, \ldots, x_p)$, and the second one is a consequence of the hypothesis of its validity for p.

A more direct proof using the definition of entropy procedes as follows:

$$\begin{aligned} &H(X_1, \ldots, X_p) = \\ &= -\sum_{x_1,\ldots,x_p} p_{(X_1,\ldots,X_p)}(x_1, \ldots, x_p) \log_2 p_{(X_1,\ldots,X_p)}(x_1, \ldots, x_p) \\ &= -\sum_{x_1,\ldots,x_p} p_{(X_1,\ldots,X_p)}(x_1, \ldots, x_p) \log_2 \prod_{k=1}^{p} p_{(X_k|X^{k-1})} p(x_k|x^{k-1}) \\ &= -\sum_{x_1,\ldots,x_p} p_{(X_1,\ldots,X_p)}(x_1, \ldots, x_p) \sum_{k=1}^{p} \log_2 p_{(X_k|X^{k-1})} p(x_k|x^{k-1}) \\ &= \sum_{k=1}^{p} \Big\{ -\sum_{x_1,\ldots,x_p} p_{(X_1,\ldots,X_p)}(x_k, \ldots, x_p) \log_2 p_{(X_k|X^{k-1})} p(x_k|x^{k-1}) \Big\} \\ &= \sum_{k=1}^{p} \Big\{ -\sum_{x_1,\ldots,x_k} p_{(X_k|X^{k-1})}(x_k|x^{k-1}) \log_2 p_{(X_k|X^{k-1})} p(x_k|x^{k-1}) \Big\} \end{aligned} \tag{1.7}$$

$$= \sum_{k=1}^{p} H(X_k|X^{k-1}) \tag{1.8}$$

where the second equality follows from a simple factorization of the joint distribution, and the last but one equality from marginalization. The last one comes from the definition of conditional entropy. For ease of notation, we write $X^k = (X^1, \ldots, X^k)$.

$\diamond$

1.1.2 Uncertainty and Information

We defined Entropy, characterized it and stated some of its important properties. Postulates and properties help understand why such a quantity can be considered as a measure of Uncertainty.

The reasons that lead to interpret Entropy also as a measure of Information can be summarized in the common sense principle, stated in the following

Observation 1.1.5 ***Anything able to reduce our uncertainty about the phenomenon under consideration has provided us with some Information***

According to this perspective, Uncertainty and Information are nothing else but two sides of the same story: once we manage to measure Uncertainty and a fortiori **changes of Uncertainty between conditions - uncertainty differentials** - we have also find a way to measure the amount of information we have gained from any received additional piece of Information.

If we consider a r.v. X and an experiment Y somehow related to it, if we accept to measure uncertainty by Entropy, then we are faced with two quantities:

- $H(X)$: the amount of uncertainty in X **before** observing the result of the experiment Y
- $H(X|Y)$: the average amount of uncertainty in X **after** observing the result of the experiment Y

We can then conclude that by performing the experiment we obtain **on average a non-negative** - by property 4 of theorem 1.1.2 - **reduction of Uncertainty** equal to $H(X) - H(X|Y)$.

Consequently, adopting the viewpoint of Observation 1.1.5, we can state that from the experiment we have gained an amount of Information equal to

$$I_X(Y) = H(X) - H(X|Y) \geq 0$$

adopting the temporary notation $I_X(Y)$ for the information about X provided by Y.

Example 1.1.2 *(Partial Information)*
The simplest case of the above situation can be formalized as follows. X takes value in the set $\mathcal{X} = \{x_1, \ldots, x_n\}$ with probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ and we can observe the oucome of a binary r.v. $Y = I_E(X)$, the indicator function of E, a subset of $\mathcal{X}$. For ease of notation we can just assume that $E = \{x_1, \ldots, x_m\}$, $m < n$. In this case, after the experiment, we learn only whether the value assumed by X belongs to E or not: we are given only partial information about the outcome of X.
Synthetically

$$\{ \underbrace{x_1, \ldots, x_m}_{E, Y=1}, \underbrace{x_{m+1}, \ldots, x_n}_{E^c, Y=0} \}$$

Thus, after the experiment, we can be faced only with the two conditional distributions of $X|Y = 1$ and $X|Y = 0$, $\mathbf{p_1} = (\frac{p_1}{p_E}, \ldots, \frac{p_m}{p_E}, 0, \ldots, 0)$ and $\mathbf{p_o} = (0, \ldots, 0, \frac{p_{m+1}}{1-p_E}, \ldots, \frac{p_n}{1-p_E})$,

with probabilities p_E and $1-p_E$ respectively, $p_E = \sum_{i=1}^{m} p_i$.

$$
\begin{aligned}
H(X|Y) &= \\
&= p_E H(X|Y=1) + (1-p_E)H(X|Y=0) \\
&= p_E\Big\{ -\sum_{i=1}^{m} \frac{p_i}{p_E} \log_2 \frac{p_i}{p_E} \Big\} + (1-p_E)\Big\{ -\sum_{i=m+1}^{n} \frac{p_i}{1-p_E} \log_2 \frac{p_i}{1-p_E} \Big\} \\
&= -\sum_{i=1}^{m} p_i[\log_2 p_i - \log_2 p_E] - \sum_{i=m+1}^{n} p_i[\log_2 p_i - \log_2(1-p_E)] \qquad (1.9) \\
&= -\sum_{i=1}^{n} p_i \log_2 p_i - \Big\{ -\Big(\sum_{i=1}^{m} p_i\Big) \log_2 p_E - \Big(\sum_{i=m+1}^{n} p_i\Big) \log_2(1-p_E) \Big\} \\
&= H(X) - \Big(-p_E \log_2 p_E - (1-p_E)\log_2(1-p_E) \Big) \\
&= H(X) - H(p_E) = H(X) - H(Y)
\end{aligned}
$$

where $H(p)$ was defined in Observation 1.1.2.

Finally we obtain

$$I_X(Y) = H(X) - H(X|Y) = H(X) - [H(X) - H(Y)] = H(Y) = H(p_E)$$

Note that, always by Observation 1.1.2, the closer p_E to $\frac{1}{2}$, the higher is reduction in uncertainty about X. The higher the entropy of the r.v. which we can observe the greater the information we gain.

1.2 Mutual Information and Conditional Mutual Information

Obviously the reasoning of the example in the previuos section can be extended to more general situations - that is, to others than dichotomous experiments. And we are in fact in the position to give the definition of a quantity of fundamental relevance in this work, and in IT generally.

Definition 1.2.1 *Let (X, Y) be a random vector with distribution $\pi = (\pi_{ij}, i = 1, \ldots, n, j = 1, \ldots, m)$ and marginals $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_m)$. The **Mutual Information** between the r.v.'s X and Y is*

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} \log_2 \frac{\pi_{ij}}{p_i q_j} \end{aligned} \tag{1.10}$$

The equivalence of the three expressions can be easily verified.

In fact,

$$\begin{aligned} I(X,Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} \log_2 \frac{\pi_{ij}}{p_i q_j} = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} \log_2 \frac{p_i q_{j|i}}{p_i q_j} = \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} \Big[\log_2 q_{j|i} - \log_2 q_j \Big] \\ &= -\sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \pi_{ij} \Big) \log_2 q_j - \sum_{i=1}^{n} p_i \Big\{ -\sum_{j=1}^{m} q_{j|i} \log_2 q_{j|i} \Big\} \\ &= -\sum_{j=1}^{m} q_j \log_2 q_j - \sum_{i=1}^{n} p_i H(Y|X = x_i) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{1.11}$$

where $q_{j|i} = P(Y = y_j | X = x_i)$. And obviously, by symmetry of the expression for I(X,Y) in its components, also the second equality follows.

Observation 1.2.1 *Note that, even though generally $H(X) \neq H(Y)$, the observation of either of the two r.v.'s has the same effect, in terms of the reduction in uncertainty, on the other one. For this reason, Mutual Information can be interpreted as a common, **shared information content** of the two r.v.'s, whence the term 'mutual' indeed.*
There is no need then to denote in different ways, $I_X(Y)$ and $I_Y(X)$, the 'uncertainty differentials' previously encountered.

Some elementary properties and relations of Mutual Information with other information-theoretical quantities are collected in the following

Theorem 1.2.1 *Let $I(X,Y)$ be the Mutual Information between the r.v.'s X and Y. Then*

1. *$I(X,Y) \geq 0$ and $I(X,Y) = 0$ iff $X \perp Y$*
2. *$I(X,Y) = I(Y,X)$*
3. *$I(X,X) = H(X)$*
4. *$I(X,Y) \leq min\{H(X), H(Y)\}$ with equality iff a r.v. is a function of the other*
5. *$H(X,Y) = H(X) + H(Y) - I(X,Y)$*
6. *If $V = f(X)$ and $Z = f(Y)$ for one-to-one functions f and g on $\mathcal{X}$ and $\mathcal{Y}$ respectively, then $I(X,Y) = I(V,Z)$ (Invariance)*

Proof

The first part of property 1 can be assessed simply by recalling that $I(X,Y) = H(X) - H(X|Y)$ and $H(X|Y) \leq H(X)$ (conditioning reduces entropy).

Otherwise it can be verified directly. Note first that the function $u \log_2 u$ is convex on $u > 0$; then, by an application of Jensen's inequality, with $u = \pi_{ij}/p_i q_j$,

$$\begin{aligned} I(X,Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p_i q_j \Big(\frac{\pi_{ij}}{p_i q_j}\Big) \log_2 \Big(\frac{\pi_{ij}}{p_i q_j}\Big) \\ &\geq \Big\{ \sum_{i=1}^{n}\sum_{j=1}^{m} p_i q_j \Big(\frac{\pi_{ij}}{p_i q_j}\Big)\Big\} \log_2 \Big\{ \sum_{i=1}^{n}\sum_{j=1}^{m} p_i q_j \Big(\frac{\pi_{ij}}{p_i q_j}\Big)\Big\} \qquad (1.12) \\ &= \Big\{ \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij}\Big\} \log_2 \Big\{ \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij}\Big\} = 0 \end{aligned}$$

Given that $H(X|Y) = H(X)$ if $X \perp Y$, the 'if' implication of the second part is verified. Property 2 follows obviously by an easy examination of the symmetric expression defining

I(X,Y).

As to what regards property 3, we have $I(X,X) = H(X) - H(X|X) = H(X) - 0 = H(X)$ since for every i, $H(X|X = x_i) = 0$: the distribution of $X|X = x_i$ is clearly degenerated on the point $\{x_i\}$.

Since $H(X|Y) \geq 0$, $I(X,Y) = H(X) - H(X|Y) \leq H(X)$ and analogously $I(X,Y) \leq H(Y)$, whence property 4.

By the Additivity property of Entropy, $H(X,Y) = H(X) + H(Y|X)$ while $H(Y|X) = H(Y) - I(X,Y)$ follows from the definition of Mutual Information: thus $H(X,Y) = H(X) + [H(Y) - I(X,Y)] = H(X) + H(Y) - I(X,Y)$.

Finally property 6 can be proved just by observing that, in this discrete case, one-to-one transformations are simply a relabelling of the outcomes: they leave joint and marginal probabilities unchanged. For example, p_i is now attached to the outcome $v_i = f(x_i)$ and p_{ij} to $(v_i, z_j) = (f(x_i), g(y_j))$, but neither value has been modified, nor their correspondence since still $p_i = \sum_{j=1}^{m} p_{ij}$.

$\diamond$

Observation 1.2.2 *Property 3 of Theorem 1.2.1 helps clarify further the* ***interpretation of Entropy as a measure of Information****. Literally it tells us that the information we obtain by observing the value of X is, on average, equal to its entropy $H(X)$: observation of X removes uncertainty completely, and, since before observation we faced an amount of uncertainty equal to $H(X)$, we must have gained an equal amount of information.*

As Renyi has significantly pointed out [53], just like potential energy can be transformed into kinetic energy so uncertainty, that is potential information*, can be transformed into real information by observation. In fact Entropy is also referred to as* **self-information***.*

Observation 1.2.3 *(Functional Dependence)*

Consider property 4 of Theorem 1.2.1 and suppose that $y = f(x)$ for some deterministic

function f. Then, since $H(Y|X = x_i) = H(\delta_{f(x_i)}) = 0$ for every i, it is easy to conclude that

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(Y) \leq H(X)$$

so that $\min\{H(X), H(Y)\} = H(Y)$ *and*

$$H(X,Y) = H(X) + H(Y|X) = H(X)$$

Thus entropic relations agree with intuition:

- all the information in the vector (X,Y) is already contained in X;

- any 'transformation' applied to the values of X results in an information loss, except in the case where the function is one-to-one;

*- **mutual information is maximal when a functional dependence between the r.v.'s exists**;*

- further $H(X) > H(Y)$ excludes X being a deterministic function of Y; more pragmatically, if $|\mathcal{X}| = n > m = |\mathcal{Y}|$ with $p_i > 0$ for every i, then X cannot be a function of Y.

Very often we will be interested in measuring the mutual information between r.v.'s when some additional knowledge is already available, mostly in the form of the values of other r.v.'s.

This can be done according to the following

Definition 1.2.2 *Consider r.v.'s X, Y and Z, with joint pmf $p_{X,Y,Z}$. The **Conditional Mutual Information** between X and Y given Z, $I(X,Y|Z)$, is defined as*

$$\begin{aligned} I(X,Y|Z) &= \sum_{z\in\mathcal{Z}} p_Z(z) I(X,Y|Z=z) \\ &= \sum_{z\in\mathcal{Z}} p_Z(z) \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y|Z}(x,y|z) \log_2 \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z) p_{Y|Z}(y|z)} \\ &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z) \end{aligned} \tag{1.13}$$

Equivalence of the different expressions can be easily verified by opportune factorizations and marginalizations.

Observation 1.2.4 *Conditional Mutual Information enjoys essentially the same* ***properties*** *of its unconditional version as specified by theorem 1.2.1: it is just needed to add the conditioning to all the statements. Thus for example, property 1 becomes:* $I(X,Y|Z) \geq 0$ *and* $I(X,Y|Z) = 0$ *iff* $X \perp Y|Z$*. And so on...*

In synthesis, Conditional Mutual Information is simply a weighted average of mutual information values in different experimental conditions, with weights being equal to the probabilities of occurence of the conditions.
Easy interpretations suggest to identify Z with some confounding factor or some source of variation common to the considered variables, but some caution is needed. Consider the following simple

Example 1.2.1 *(Conditional independence vs induced dependence)*

1. *Let* X*,* Y *and* Z *be r.v.'s such that* $X \perp Y|Z$*; in general, we will not have* $X \perp Y$ *- they will not be marginally independent. So* $I(X,Y) > 0$ *while* $I(X,Y|Z) = 0$ *since* $I(X,Y|Z=z) = 0$ *for any* $z \in \mathcal{Z}$*.*

2. *On the other side, suppose* X_1 *and* X_2 *represent the outcomes of the independent tossing of two fair coins where 'head'=1 and 'tail'=0; define* $S = X_1 + X_2$*.*
   By definition $I(X_1, X_2) = 0$*, but it is easy to see that* $I(X_1, X_2|S) > 0$*.*

While in the first case, Z *can be naturally viewed as a common and unique source of variation so that its knowledge removes dependence between the r.v.'s, in the second one the opposite effect emerges -* S *makes them more dependent - and it is more difficult to give a precise connotation of its role.*

Introducing the concept of Conditional Mutual Information makes it possible to analyse relations among 'complex' objects - information between random vectors, for example - by means of decompositions in simpler constituents.

Theorem 1.2.2 ***(Chain rule for Mutual Information)***
Let X and $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be a r.v. and a random vector respectively. Then their Mutual Information can be expressed as follows

$$I(X, \mathbf{Y}) = \sum_{k=1}^{p} I(X, Y_k | Y_1, \ldots, Y_{k-1}) = \sum_{k=1}^{p} I(X, Y_k | \mathbf{Y}^{k-1}) \tag{1.14}$$

with $\mathbf{Y}^0 = \emptyset$.

Proof We will show the validity of the theorem for the case $p = 2$. The general statement will subsequently follow by an induction argument on p.

$$\begin{aligned}
&I(X, \mathbf{Y}) = \\
&= I(X, (Y_1, Y_2)) = \sum_{x,y_1,y_2} p_{X,Y_1,Y_2}(x, y_1, y_2) \log_2 \frac{p_{X,Y_1,Y_2}(x, y_1, y_2)}{p_X(x) p_{Y_1,Y_2}(y_1, y_2)} \\
&= \sum_{x,y_1,y_2} p_{X,Y_1,Y_2}(x, y_1, y_2) \log_2 \frac{p_{X,Y_1}(x, y_1) p_{Y_2|X,Y_1}(y_2|x, y_1)}{p_X(x) p_{Y_1}(y_1) p_{Y_2|Y_1}(y_2|y_1)} \\
&= \sum_{x,y_1,y_2} p_{X,Y_1,Y_2}(x, y_1, y_2) \Big\{ \log_2 \frac{p_{X,Y_1}(x, y_1)}{p_X(x) p_{Y_1}(y_1)} + \log_2 \frac{p_{Y_2|X,Y_1}(y_2|x, y_1)}{p_{Y_2|Y_1}(y_2|y_1)} \Big\} \\
&= \sum_{x,y_1} p_{X,Y_1,Y_2}(x, y_1, y_2) \log_2 \frac{p_{X,Y_1}(x, y_1)}{p_X(x) p_{Y_1}(y_1)} \\
&+ \sum_{x,y_1,y_2} p_{X,Y_1,Y_2}(x, y_1, y_2) \log_2 \frac{p_{Y_2|X,Y_1}(y_2|x, y_1)}{p_{Y_2|Y_1}(y_2|y_1)} \underbrace{\frac{p_{X|Y_1}(x|y_1)}{p_{X|Y_1}(x|y_1)}}_{1} \\
&= \sum_{x,y_1} p_{X,Y_1}(x, y_1) \log_2 \frac{p_{X,Y_1}(x, y_1)}{p_X(x) p_{Y_1}(y_1)} \\
&+ \sum_{y_1} p_{Y_1}(y_1) \sum_{x,y_2} p_{X,Y_2}(x, y_2|y_1) \log_2 \frac{p_{X,Y_2|Y_1}(x, y_2|y_1)}{p_{Y_2|Y_1}(y_2|y_1) p_{X|Y_1}(x|y_1)} \\
&= I(X, Y_1) + I(X, Y_2|Y_1)
\end{aligned} \tag{1.15}$$

◇

Note that the decomposition of the theorem is just one of the $p!$ possible since so many are the factorizations of the joint distribution $p_{X,Y_1,\ldots,Y_p}$ with respect to the Y's.

Using the previous decompositions, and assuming a specific type of dependence between the considered r.v.'s it is possible to state the following useful

Theorem 1.2.3 ***(Data Processing Inequality)***
Suppose that X, Y and Z form a Markov chain in this order,

$$X \leftrightarrow Y \leftrightarrow Z^1$$

then

$$I(Y, Z) \geq I(X, Z) \tag{1.16}$$

Proof By the chain rule for Mutual Information, we can write the two decompositions

$$\begin{aligned} I((X, Y), Z) &= I(X, Z) + I(Y, Z|X) \\ &= I(Y, Z) + I(X, Z|Y) = I(Y, Z) \end{aligned} \tag{1.17}$$

since $X \perp Z|Y$, and consequently $I(X, Z|Y) = 0$.
Observing that $I(Y, Z|X) \geq 0$ completes the proof.

◇

Remark 1.2.1 *If in the preceding theorem, we fix $X = T(\mathbf{X})$, $Y = \mathbf{X}$ and $Z = \Theta$, where Θ is the parameter in the distribution of the sample data $\mathbf{X}$ and $T(\mathbf{X})$ is any statistic computed on data, for which $T(\mathbf{X}) \leftrightarrow \mathbf{X} \leftrightarrow \Theta$ is clearly verified, the conclusion is*

[1] The double arrow indicates the fact that the chain can be read in both directions while the sequence of the r.v.'s - the "order" mentioned in the theorem - and the consequent conditional independence relations remain the same.

$I(\mathbf{X}, \Theta) \geq I(T(\mathbf{X}), \Theta)$: ***no data manipulation, however cleverly designed, can provide us more information than the amount originally contained in the sample***.

Coherently one could call a statistic T such that $I(T, \Theta) = I(\mathbf{X}, \Theta)$ - for which no information loss occurs - ***sufficient (with respect to Mutual Information)***.

It is not difficult to show that a statistic sufficient in the classical sense is also sufficient with respect to Mutual Information.

We close this section with an example we hope will give a more 'pragmatic' meaning to the concepts introduced so far.

Example 1.2.2 *(A Game of Information)*

The following game is a formalized version of an hungarian traditional game named "Barkochba" (see [53] pp.5-8, for example) substantially analogous to the popular "Twenty question" parlour game ([13], pp.5-6).

Suppose a friend chooses at random (equal selection probabilities) an object from a set containing 2^n ones. For simplicity, they can just be labelled $\{1, \ldots, N\}$ where $N = 2^n$. The game ends when you find out which object has been selected. The only tool you are given is the possibility of posing questions with yes-no anwer to your friend. The goal is clearly to minimize the number of questions needed to identify the object.

Which questions should you ask? Is there a best way of posing the questions?

Possible strategies can be, for example, that of asking "'Has object i been selected?, starting from $i = 1$ and following the list up to $i = N$; otherwise you could ask if $i \geq k_t$ or $i < k_t$ for a certain sequence of suitably chosen constants ($k_t \in \{1, \ldots, N\}, t = 1 \ldots, n$).

Essentially your goal is that of reducing to 0 the uncertainty you are faced with. And in this case, computing the entropy of the random experiment X represented by the choice of

the object, we arrive at

$$H\Big(\Big(\frac{1}{2^n},\dots,\frac{1}{2^n}\Big)\Big) = -\sum_{i=1}^{N}\frac{1}{2^n}log_2\frac{1}{2^n} = log_2 2^n = n \;\; bits \tag{1.18}$$

Now, the answer to a yes-no question is clearly nothing different from the observation of the value of an indicator funtion of some subset E of the choice set: selecting a question is exactly equivalent to selecting such a subset, and there are 2^n of them. Which one to choose?

Recalling Example 1.1.2 and its conclusions, to maximize the information gain we will have from the answer, we should choose a subset E for which p_E is as close as possible to $1/2$. In our present situation we are lucky, since we can obtain a set with probability exactly equal to $1/2$ just by collecting half of the objects (indeed, we can find a discrete number of these 'optimal' sets...).

Thus $H(p_E) = H(1/2) = 1$ bit and, after the first question, we are left with $H(X) - H(1/2) = n - 1$ bits of uncertainty, whatever the answer.

Pursuing this strategy at each stage, it is easy to conclude that you will need exactly n questions to find the selected object.

It is possible to show that any other strategy will need at least n questions, on average, to guide you to the prescribed goal, so that the one described above is optimal.

The reasoning followed above is not at all limited to cases of equiprobability: it will work with any distribution used to select the object. In each case, the entropy of such distribution can then be thougth of as the minimum number of binary questions needed to discover it.

Frequently however it will not be so simple as before to find the right questions to pose: collecting objects with probabilities adding exactly to $1/2$ will not be possible in general, so we should content ourselves with suboptimal solutions. To be true, in cases different from equal selection probabilities the entropy of such 'selection distribution' will be lower than n: we are facing a lower uncertainty and a lower number of questions will be needed, so

that these situations are more favourable than the one considered here. They can be defined 'suboptimal' meaning that generally the lower bound can only be approached but not reached, contrary to the higher lower bound of this example that can be actually achieved. Further while in this example choice of a question (subset) at a certain stage has no effect on the possibility of forming the best partitioning at the successive stages, this will not be true in the case in which unequal selection probabilities are adopted: partitioning at earlier stages will influence the possibility of 'optimal' partioning at the later ones.

Computational difficulties aside, entropy characterizes the difficulty of the game and establishes a lower bound for players.

Finally note that when the game ends you are left with an n-tuple of '0's and '1's: without loss of generality, if, at each stage, the set E is assumed to be the first half of the 'list' of remaining objects, as originally ordered, this n-tuple provides you with a map to spot the object inside the whole list: for example, an n-tuple beginning with $(1, 1, 0, \dots)$ tells you to concentrate attention on the first half of the 'list', then to narrow your search to the first quarter of the list and then on the second eighth of it (second half of the first quarter), and so on.

Such n-tuples can be easily interpreted as labels *for the objects, or more properly as* ***codewords****: in fact, no great effort is needed to change slightly the viewpoint on the game and see it as a* ***communication game****, instead of an information one.*

Some random mechanism, X, selects an object from a set and your friend has to help you understand it - communicate it to you - with the only aid of yes-no questions and with the shortest number of them: that is, your friend will send you a coderword identifying the selected object. Suppose to play this game repeatedly: what is the shortest average length of such 'binary strings' with a particular chosen code? Again the answer is $H(X)$. Thus even the best designed code that you and your friend would agree upon cannot beat the lower bound represented by entropy.

1.3 Differential Entropy and Mutual Information

The quantities defined in the preceding sections can be extended to embrace discrete r.v.'s assuming infinite values or continuous r.v.'s. Their properties partly differ from those stated for the finite case, and some clarifications will be needed.

1.3.1 Countable case

No great change in the definition of Entropy is needed.

Definition 1.3.1 *Let X be a discrete r.v. taking value in the set $\mathcal{X} = \{x_1, x_2, \dots\}$ with pmf $p(x)$, $x \in \mathcal{X}$. The entropy of X is defined as*

$$H(X) = -\sum_{i=1}^{\infty} p(x_i) \log_2 p(x_i) = E[-\log_2 p(X)] \tag{1.19}$$

All of the properties stated in Theorem 1.1.2 remain valid, with the obvious exception of the upper bound in 1: the **entropy of a countably discrete r.v.** is not necessarily finite. It **can be infinite** obviously depending on the character of the series in the preceding definition.

Observation 1.3.1 *Note that the possibility of an infinite entropy agrees with what suggested by intuition: if $p_X(x) > 0$ for all but finitely many x's $\in \mathcal{X}$, we are back to the case of the preceding section and we would not really need the more general definition: entropy would have a finite value, since the upper bound in property 1 of Theorem 1.1.2 would obviously be valid with $n = |\{x : x \in \mathcal{X}, p_X(x) > 0\}|$.*

Thus the above definition becomes really necessary when $|\mathcal{X}| = \infty$ and $p_X(x) > 0$, $\forall x \in \mathcal{X}$. If you recall Example 1.2.2 on the 'yes-no question game', in this case we have an infinite number of object that can be selected with positive probability and an infinite number of such questions could actually be needed to spot the selected one.

When, even in this case, $H(X) < \infty$, we are reassured that, with probability one, we will

stop asking questions - and find out the object - in a finite time: it could take us a long sequence of 0's and 1's but a finite sequence only.

1.3.2 Continuous case

A first attempt at defining entropy for continuous r.v.'s could resort to 'discretization' (or quantization) and application of definition 1.1.1, or more probably 1.3.1, to the 'discretized' version.

Let X be a continuous r.v. with pdf $f_X(x) > 0$ on (a, b), $-\infty < a < b < \infty$, and, for a partition $a = x_0 < x_1 < x_2 < \cdots < x_n = b$ with $y_j \in (x_{j-1}, x_j)$, define the distribution of the discrete r.v. $X^{(n)}$ as

$$p_j = P(X^{(n)} = y_j) \qquad p_j = \int_{x_{j-1}}^{x_j} f_X(x)dx \tag{1.20}$$

Then we can define

$$H(X^{(n)}) = -\sum_{j=1}^{n} p_j \log_2 p_j \tag{1.21}$$

and take the limit as n tends to ∞.

We can write

$$\begin{aligned} H(X^{(n)}) &= -\sum_{j=1}^{n} p_j \log_2 f_X(y_j) + \sum_{j=1}^{n} p_j \log_2 \Big(\frac{f_X(y_j)}{p_j}\Big) \\ &= -\sum_{j=1}^{n} \frac{p_j}{\Delta_j}\Delta_j \log_2 f_X(y_j) + \sum_{j=1}^{n} \frac{p_j}{\Delta_j}\Delta_j \log_2 \Big(\frac{f_X(y_j)}{\frac{p_j}{\Delta_j}\Delta_j}\Big) \\ &\to -\int_a^b f_X(x) \log_2 f_X(x)dx + \lim_{n\to\infty} \sum_{j=1}^{n} \frac{p_j}{\Delta_j}\Delta_j \log_2 \Big(\frac{f_X(y_j)}{\frac{p_j}{\Delta_j}}\frac{1}{\Delta_j}\Big) \\ &= \infty \end{aligned} \tag{1.22}$$

$\Delta_j = x_j - x_{j-1}$, since $\frac{p_j}{\Delta_j} = \frac{1}{\Delta_j}\int_{x_{j-1}}^{x_j} f_X(x)dx \to f_X(y_j)$ as $n \to \infty$ $(\Delta_j \to 0)$ but $\frac{1}{\Delta_j} \to \infty$.

Thus definition by means of a limiting argument provides no help, except for the observation that the preceding expression going to infinity is generally due to the second term:

thus if we partition the interval in subintervals of the same length $\frac{1}{n}$, the second term will be $\log_2 n$, and

$$H(X^{(n)}) - \log_2 n \to -\int_a^b f_X(x) \log_2 f_X(x) dx \tag{1.23}$$

as $n \to \infty$.

This observation is the reason for the following

Definition 1.3.2 *Let X be an absolutely continuous r.v. with probability density function $f_X(x)$ with respect to Lebesgue measure* [2] *, $x \in \Re$. Then the **entropy of** X is defined as*

$$h(X) = -\int_{-\infty}^{+\infty} f_X(x) \log_2 f_X(x) dx = E[-log_2 f_X(X)] \tag{1.24}$$

A simple example will prove how the differential entropy h just defined is a somehow different entity from the discrete entropy H previuosly introduced, and it will also underline that its interpretation is not so straightforward as that for discrete entropy.

Example 1.3.1 *(Uniform distribution)*
Let X be r.v. with $f_X(x) = \frac{1}{a}$ for $0 < x < a$, $a > 0$ and 0 otherwise. Then

$$h_a = -\int_{-\infty}^{+\infty} \frac{1}{a} \log_2 \frac{1}{a} dx = \log_2 a \int_0^a \frac{1}{a} dx = \log_2 a \tag{1.25}$$

So if $a < 1$, $h_a(X) < 0$.

Thus differential entropy **can be negative** and this contrasts with intuition: a discrete-entropy value of 0 meant no uncertainty. How to interpret a negative one? Can we be more than certain?
In any case, it remains **true that smaller values of entropy correpond to lower uncertainty**, as common sense would suggest: in Example 1.3.1, as a becomes smaller our uncertainty about the possible values assumed by X is constantly decreasing - the

[2] The definition could be given even with respect to any other space, $\mathcal{X}$, and dominating σ-finite measure μ, i.e. $h(X) = -\int_{\mathcal{X}} f_X(x) \log_2 f_X(x) d\mu(x)$, thus encompassing as special cases the discrete one with counting measure as dominating or the continuous one here introduced.

interval which they can possibly belong to is becoming narrow and narrow - and this is translated in the value of h_a becoming lower and lower, even if taking on negative values.

Consider now the following

Example 1.3.2 *(Uniform distribution, continued)*
Define $Y = bX$, $b > 0$. Clearly Y has a uniform distribution on the interval $(0, ab)$ and, by (2.18), we then have

$$h(Y) = h_{ab} = \log_2 a + \log_2 b = h_a + h_b = h(X) + h_b \tag{1.26}$$

i.e. $h(X) \neq h(Y)$.
But we used a one-to-one function to tranform from X to Y: thus differential entropy is not invariant with respect to one-to-one tranformations.

Observation 1.3.2 *The **lack-of-invariance** issue emerging from the preceding example can be stated more generally. Let X be a r.v. with pdf $f_X(x)$ and define $Y = g(X)$ with $g(\cdot)$ one-to one function. Then*

$$\begin{aligned}
h(Y) &= \\
&= -\int_{-\infty}^{+\infty} f_Y(y) \log_2 f_Y(y) dx \\
&= -\int_{-\infty}^{+\infty} f_X(g^{-1}(y)) \Big| \frac{dg^{-1}(y)}{dy} \Big| \log_2 \Big\{ f_X(g^{-1}(y)) \Big| \frac{dg^{-1}(y)}{dy} \Big| \Big\} dy \\
&= -\int_{-\infty}^{+\infty} f_X(g^{-1}(y)) \Big| \frac{dg^{-1}(y)}{dy} \Big| \Big\{ \log_2 f_X(g^{-1}(y)) + \log_2 \Big| \frac{dg^{-1}(y)}{dy} \Big| \Big\} dy \\
&= -\int_{-\infty}^{+\infty} f_X(x) \log_2 f_X(x) dx - \int_{-\infty}^{+\infty} f_Y(y) \log_2 \Big| \frac{dg^{-1}(y)}{dy} \Big| dy \\
&= h(X) - E\Big[\log_2 |J(Y)| \Big]
\end{aligned} \tag{1.27}$$

where g^{-1} is the inverse of g and $J(y) = \frac{dg^{-1}(y)}{dy}$ is the Jacobian of the transformation.

Only a weaker form of invariance is still valid: **invariance with respect to translations**. In this case, in fact, $J(y) = 1$, $\forall y \in \mathcal{Y}$ and the second term on the right side of (2.20)

vanishes.

Contrary to the discrete case, differential entropy is thus influenced by the underlying space on which the distribution is assigned: this is intuitively equal to state that values of the pdf are no more the only relevant quantities taken into account in the computation, but values assumed by the r.v. are important too.

Having these provisos been made, it is natural to extend the Information measures defined in the preceding section to the present context.

Definition 1.3.3 *Let (X, Y, Z) be a random vector with pdf $f_{X,Y,Z}(x, y, z)$ with respect to Lebesgue measure on $\Re^3$. The* ***joint differential entropy*** *of (X, Y), the* ***conditional differential entropy*** *of Y given X, the* ***Mutual Information*** *between X and Y and the* ***Conditional Mutual Information*** *between X and Y given Z are respectively*

$$
\begin{aligned}
h(X,Y) &= -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log_2 f_{X,Y}(x,y)dxdy \\
h(Y|X) &= -\int_{-\infty}^{+\infty} f_X(x)\int_{-\infty}^{+\infty} f_{Y|X}(y|x)\log_2 f_{Y|X}(y|x)dxdy \\
I(X,Y) &= h(X) - h(X|Y) = h(Y) - h(Y|X) \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log_2 \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy \qquad (1.28) \\
I(X,Y|Z) &= h(X|Z) - h(X|Y,Z) = h(Y|Z) - h(Y|X,Z) \\
&= \int_{-\infty}^{+\infty} f_Z(z) \\
&\times \Big\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y|Z}(x,y|z)\log_2 \frac{f_{X,Y|Z}(x,y|z)}{f_{X|Z}(x|z)f_{Y|Z}(y|z)}dxdy\Big\}dz
\end{aligned}
$$

with obviuos meaning of the notation and assuming that all integrals are defined.

It is clear that the above definition is in no way limited to bivariate random vectors: to obtain the general versions it is only needed to assume X and Y themselves to be random vectors.

By Example 1.3.1 it is clear that even joint and conditional entropies can assume negative values; they can even be infinite, so no general upper bound can be provided.

Observation 1.3.3 *As to what regards the* ***possibility of*** *differential entropy* ***being infinite****, we could resort to the same reasoning adopted for the countably discrete case: we have an uncountably infinite set of real numbers to select from and an infinite number of questions would probably be necessary to reach this goal. Unfortunately this line of thought would lead us to conclude that entropy of 'actually continuous' r.v.'s (those for which $f_X(x) > 0$, for all x in some non-empty interval, for example) is always infinite,*

since infinite is the amount of information needed to know exactly a real number (We need to specify an infinite sequence of decimal digits). And indeed this conclusion was reached by the limiting process described at the beginning of the section.
But this is clearly in contrast with the cases - almost all of those encountered in practice - in which $h(X)$ is finite: consider example 1.3.1 for any value of a!
Differential entropy being infinite is a different condition than (discrete) entropy being infinite: discrete entropy of an absolutely continuous r.v. is always infinite. Differential entropy of such a r.v. can be and frequently is finite.
The fact is that differential entropy actually measures the 'probabilistic dispersion' of some object - continuous r.v. - intrinsecally different from discrete r.v., or in other terms it measures dispersion with respect to different dimensions. For a deep discussion of these issues, see the works of Renyi ([3], [50], [51]).

But some properties of the discrete versions remain valid.

Theorem 1.3.1 *Let (X, Y) be a random vector. Then*

1. *Conditioning (*still*) reduces entropy*
   $h(Y|X) \leq h(Y)$

2. *Additivity*
   $h(X, Y) = h(X) + h(Y|X)$

3. *Chain rule*
   $h(X_1, \ldots, X_p) = \sum_{i=1}^{p} h(X_i|X^{i-1})$

4. $h(X, Y) = h(X) + h(Y) - I(X, Y)$

Proof. Property 2 is an immediate consequence of the definition of $h(X, Y)$ and of a natural factorization of the joint density of X and Y. Property 3 follows from property 2 and the definition of mutual information. As to what regards property 1 it will be a

consequence of non-negativity of mutual information, proved in the next theorem. $\diamond$

One property that is missing from the above list is the identification of a distribution maximixing entropy, or, as it is usually called, a **maximum entropy distribution**. An obvious reason for this fact is that we can face different spaces with different 'properties': it is not possible to find a distribution able to maximise entropy for every underlying space, whatever this would be. Moreover, for the entropy maximization problem to be well defined, unbounded spaces like the half real axis, or the whole real axis, ask for the introduction of additional constraints - i.e. additional information - with respect to the only 'integrate to one' constraint. The uniform distribution on an infinite space is no more a probability distribution.
A few simple examples will clarify the subject.

Example 1.3.3 *(Maximum entropy distribution on a finite interval)*
Suppose we are looking for a distribution maximising (differential) entropy and the only information we are given is that its support is the interval (a, b), $a < b$, of real numbers. We can construct the Lagrangean function and maximise it under the only available constraint that the distribution must concentrate all its mass in the assigned interval.

$$\mathcal{L}(f, \lambda) = -\int_{-\infty}^{+\infty} f(x) \log_2 f(x) dx - \lambda\Big(\int_a^b f(x)dx - 1\Big) \tag{1.29}$$

$$\frac{\partial \mathcal{L}}{\partial f(x)} = -\log_2 f(x)^* - 1 - \lambda^* = 0 \tag{1.30}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \int_a^b f^*(x)dx - 1 = 0 \tag{1.31}$$

From the first condition, it is evident that $f^*(x) = 2^{-(1+\lambda^*)} = c$, with c a constant. Thus, the second one determines the value $c = 1/(b-a)$.
The desired distribution is then the uniform distribution on the prescribed interval: $f^*(x) = 1/(b-a)$ for $a < x < b$, and 0 otherwise.

The conclusions of Example 1.3.3 may be misleading: actually we ended up right with the uniform distribution as we did with the discrete case. But consider the following

Example 1.3.4 *(Maximum entropy distribution on the positive real axis)*
The problem is the same as in Example 1.3.3 but with $a = 0$, $b = +\infty$ and the additional constraint on the expected value of the r.v. whose distribution we are seeking.
More precisely, for $\theta > 0$,

$$\begin{aligned}
\mathcal{L}(f, \lambda, \mu) &= -\int_{-\infty}^{+\infty} f(x) \ln f(x) dx - \lambda\Big(\int_0^{+\infty} f(x)dx - 1\Big) \\
&- \mu\Big(\int_0^{+\infty} x f(x)dx - \theta\Big) \\
\frac{\partial \mathcal{L}}{\partial f(x)} &= -\ln f^*(x) - 1 - \lambda^* - \mu^* x = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \int_0^{+\infty} f^*(x)dx - 1 = 0 \\
\frac{\partial \mathcal{L}}{\partial \mu} &= \int_0^{+\infty} x f^*(x)dx - \theta = 0
\end{aligned} \tag{1.32}$$

Again, from the first condition, we obtain $f^*(x) = e^{-(1+\lambda^*)-\mu^* x}$, and from the second one,

$$1 = \int_0^{+\infty} f^*(x)dx = \int_0^{+\infty} e^{-(1+\lambda^*)-\mu^* x} dx = \frac{e^{-(1+\lambda^*)}}{\mu^*} \tag{1.32}$$

or $(1 + \lambda^*) = -\ln \mu^*$. Thus $f^*(x) = e^{\ln \mu^* - \mu^* x} = \mu^* e^{-\mu^* x}$, that is the density of an exponential distribution. The constraint on the expected value finally leads to $\mu^* = \frac{1}{\theta}$ and $f^*(x) = \theta^{-1} e^{-\frac{x}{\theta}}$.
Few calculations will lead to the conclusion that the problem is not well posed if the additional constraint on the mean value is not introduced.

Finally another characterization of Normality derives from the following

Example 1.3.5 *(Maximum entropy distribution on the real line)*
A distribution on the real line, with density $f(x)$, is sought such that it maximises entropy

and has second moment equal to $\sigma^2 > 0$. Proceding as before we write

$$\begin{aligned}
\mathcal{L}(f,\lambda,\mu) &= -\int_{-\infty}^{+\infty} f(x)\ln f(x)dx - \lambda\Big(\int_{-\infty}^{+\infty} f(x)dx - 1\Big) \\
&- \mu\Big(\int_{-\infty}^{+\infty} x^2 f(x)dx - \sigma^2\Big) \\
\frac{\partial\mathcal{L}}{\partial f(x)} &= -\ln f^*(x) - 1 - \lambda^* - \mu^* x^2 = 0 \\
\frac{\partial\mathcal{L}}{\partial\lambda} &= \int_{-\infty}^{+\infty} f^*(x)dx - 1 = 0 \\
\frac{\partial\mathcal{L}}{\partial\mu} &= \int_{-\infty}^{+\infty} x^2 f^*(x)dx - \sigma^2 = 0
\end{aligned}$$

We can conclude that

$$f^*(x) = e^{-(1+\lambda^*)-\mu^* x^2} \tag{1.33}$$

$$1 = e^{-(1+\lambda^*)}\int_{-\infty}^{+\infty} e^{-\mu^* x^2}dx \tag{1.34}$$

$$e^{(1+\lambda^*)} = \sqrt{\pi/\mu^*} \tag{1.35}$$

$$(1+\lambda^*) = \ln\sqrt{\pi/\mu^*} \tag{1.36}$$

$$f^*(x) = e^{-\ln\sqrt{\pi/\mu^*}-\mu^* x^2} = \frac{1}{\sqrt{\pi/\mu^*}}e^{-\mu^* x^2} \tag{1.37}$$

$$\tag{1.38}$$

It is easy now to see it must be that $\mu^* = \frac{1}{2\sigma^2}$.

A further constraint could be introduced to impose a value of the mean different from 0. Otherwise we could just recall that entropy is invariant with respect to translation: so the distribution on the real line maximising entropy, under fixed mean and second central moment, is the normal distribution just obtained opportunely translated.

Actually these examples are just the tip of an iceberg we will not be able to explore - the Maximum Entropy Principle and its applications.

The fundational papers in this framework are those of Jaynes - [32] and [34] - while a

recent synthesis of the Maximum Entropy Principle is still provided by Jaynes [35]. Essentially, and doing no justice to the richness of this field of research, this approach tries to provide an answer to the questions 'What is the distribution that assumes only pieces of information specified by the considered moments and no additional pieces?' and 'How to update the current distribution about a phenomenon under the light shed by new information?'.
Maximum Entropy Methods can also be used to elicit prior distributions, for example.

The properties collected in the following theorem help explain why Mutual Information is a more appropriate concept of Information to deal with in connection with continuous r.v.'s.

Theorem 1.3.2 *Let (X, Y) be a random vector and $I(X, Y)$ be the Mutual Information between X and Y as defined by (2.21). Then*

1. *$I(X, Y) \geq 0$ and $I(X, Y) = 0$ iff $X \perp Y$*
2. *$I(X, Y) = I(Y, X)$*
3. *$h(X, Y) = h(X) + h(Y) - I(X, Y)$*
4. *If $V = g(X)$ and $Z = h(Y)$ for one-to-one functions g and h on $\mathcal{X}$ and $\mathcal{Y}$ respectively, then $I(X, Y) = I(V, Z)$ (Invariance)*

Proof We will demonstrate the validity of 1 and 4; properties 2 and 3 are obvious. By convexity of $u \log u$ with $u = \frac{f_{X,Y}}{f_X \cdot f_Y}$ and Jensen's inequality, we see that

$$\begin{aligned} I(X,Y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log_2 \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_X(x)f_Y(y)\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\log_2 \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy \\ &\geq \Big\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_X(x)f_Y(y)\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy\Big\} \\ &\quad \times \log_2\Big\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_X(x)f_Y(y)\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy\Big\} \end{aligned} \tag{1.39}$$

$$\begin{aligned} &= \Big\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy\Big\}\log_2\Big\{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy\Big\} \\ &= 1\log_2 1 = 0 \end{aligned} \tag{1.40}$$

As to what regards property 4, by the change of variable formula

$$\begin{aligned} I(X,Y) &= \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log_2 \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(g^{-1}(v),h^{-1}(z))|J|\log_2 \frac{f_{X,Y}(g^{-1}(v),h^{-1}(z))}{f_X(g^{-1}(v))f_Y(h^{-1}(z))}dvdz \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(g^{-1}(v),h^{-1}(z))\Big(\frac{\partial g^{-1}(v)}{\partial v}\cdot\frac{\partial h^{-1}(z)}{\partial z}\Big) \\ &\times \log_2 \frac{f_{X,Y}(g^{-1}(v),h^{-1}(z))}{f_X(g^{-1}(v))f_Y(h^{-1}(z))}dvdz \end{aligned} \tag{1.41}$$

$$
\begin{aligned}
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{V,Z}(v,z)\log_2 \frac{f_{X,Y}(g^{-1}(v),h^{-1}(z))\frac{\partial g^{-1}(v)}{\partial v}\cdot\frac{\partial h^{-1}(z)}{\partial z}}{f_X(g^{-1}(v))\frac{\partial g^{-1}(v)}{\partial v}f_Y(h^{-1}(z))\frac{\partial h^{-1}(z)}{\partial z}}dvdz \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{V,Z}(v,z)\log_2 \frac{f_{V,Z}(v,z)}{f_V(v)f_Z(z)}dvdz \qquad (1.42)\\
&= I(V,Z)
\end{aligned}
$$

where $|J| = \frac{\partial g^{-1}(v)}{\partial v}\cdot\frac{\partial h^{-1}(z)}{\partial z}$, the Jacobian of the tranformation $(v,z) = F(x,y) = (g(x),h(y))$, since $J = \left(\frac{\partial g^{-1}(v)}{\partial v},0;\ 0,\frac{\partial h^{-1}(z)}{\partial z}\right)$. $\diamond$

Thus Mutual Information maintains the properties of what we usually consider a measure of the information a r.v. provides about another one: on the average, knowledge of the value of X provides information about Y or at most provides no information at all. But it cannot provide 'negative information'.

X provides about Y as much information as that provided by Y on X.

Even if entropies can be negative, the joint entropy of two r.v.'s is always smaller than or at most equal to the sum of their individual entropies, by non negativity of mutual information: any form of dependence reduces the uncertainty of the system.

Finally one-to-one tranformations leave mutual information unchanged, as one would expect, since the original variables, and their relationship of dependence, can always be 'recontructed' from the tranformed ones.

Observation 1.3.4 *The last property has an important implication with respect to Bayesian analysis. As it will be done in a major part of this work, a Bayesian may desire to measure the amount of information about a parameter Θ gained from the observation of some r.v. X, $I(X,\Theta)$. The* ***choice of a parametrization*** *for the statistical model is usually done on the basis of mathematical convenience: a parametrization originated from the problem at hand may be subsequently changed to help in the statistical analysis. We could then move from Θ to η.*

Intuition would suggest that observation should lead us to the same information gain

whatever form (parametrization) we choose to express our uncertainty about the unknown elements of the problem: this is precisely what happens if we measure this information gain via Mutual Information, since property 4, in the present context, states $I(\Theta, X) = I(\eta, X)$.

Finally Mutual Information does not run into the 'discretisation' problem pointed out for Entropy.

In fact, if we consider a random vector (X, Y) assuming value in the set $(a, b) \times (c, d)$ with pdf $f_{X,Y}(x, y)$ and apply the discretisation with partitionings $a = x_0 < x_1 < \cdots < x_n = b$, $c = y_0 < y_1 < \cdots < y_n = d$, with $\Delta_i = x_i - x_{i-1}$ and $\Delta_j = y_j - y_{j-1}$ to obtain the discrete random vector $(X^{(n)}, Y^{(n)})$ with law

$$p_{ij} = \int_{x_{i-1}}^{x_i} \int_{y_{j-1}}^{y_j} f_{X,Y}(x, y)dxdy \tag{1.43}$$

$$p_{i.} = \int_{x_{i-1}}^{x_i} \int_{c}^{d} f_{X,Y}(x, y)dxdy \tag{1.44}$$

$$p_{.j} = \int_{a}^{b} \int_{y_{j-1}}^{y_j} f_{X,Y}(x, y)dxdy \tag{1.45}$$

it is easy to see that

$$\begin{aligned} I(X^{(n)}, Y^{(n)}) &= \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_{i.}p_{.j}} \\ &= \sum_{i,j} \frac{p_{ij}}{\Delta_i \Delta_j} \log_2 \left\{ \frac{\frac{p_{ij}}{\Delta_i \Delta_j}}{\frac{p_{i.}}{\Delta_i} \frac{p_{.j}}{\Delta_j}} \right\} \Delta_i \Delta_j \\ &\to \int_a^b \int_c^d f_{X,Y}(x, y) \log_2 \frac{f_{X,Y}(x, y)}{f_X(x) f_Y(y)} dxdy \end{aligned} \tag{1.46}$$

as $n \to \infty$, for the same reasons indicated in the entropy case.

Thus Mutual Information in the continuous case still possesses the interpretation and meaning it showed in the discrete one.

In conclusion, it must be noticed that Mutual Information is a special case of a fundamental quantity in Information Theory and Statistics, named in different ways: cross-entropy, relative entropy or Kullback-Leibler divergence.

Definition 1.3.4 *Let f and g be pdf's with respect to a σ-finite measure μ on the same space $\mathcal{X}$, such that $P << Q$* [3] *where P and Q are the distributions associated with f and g respectively. The* ***Kullback-Leibler divergence (cross-entropy or relative entropy)*** *between f and g is defined as*

$$\begin{aligned} KL(f|g) = KL(P|Q) &= \int_{\mathcal{X}} f(x) \log_2 \frac{f(x)}{g(x)} d\mu(x) \\ &= \int_{\mathcal{X}} \log_2 \Big(\frac{dP}{dQ}(x)\Big) dP(x) = \int_{\mathcal{X}} \log_2(P_Q(x)) dP(x) \end{aligned} \tag{1.47}$$

where $P_Q = \frac{dP}{dQ}$ is the Radon-Nykodym derivative of P with respect to Q. Equivalence of the different expressions follows easily by noting that $f/g = \frac{dP/d\mu}{dQ/d\mu} = dP/dQ$.
Furthermore $KL(f|g) \geq 0$ with equality iff $f = g$ (P=Q).

From the above definition and comparison with the expression for mutual information in (2.25), it can be concluded that

$$\begin{aligned} I(X,Y) &= KL(f_{X,Y}|f_X f_Y) \\ &= E\big[KL(f_{Y|X}(\cdot|X)|f_Y)\big] \\ &= E\big[KL(f_{X|Y}(\cdot|Y)|f_X)\big] \end{aligned} \tag{1.48}$$

where the expectations in the second and third lines above are with respect to the marginal distributions of X and Y respectively.

Even if Kullback-Leibler divergence, as it is known, is **not a proper distance** - since, in general, it is not symmetric and it does not satisfy the Triangular Inequality - nonetheless it can be interpreted as **a measure of how far apart** two distributions are from one another, of how differently they spread their probability mass on the underlying space.

[3] "<<" is used here to denote absolute continuity of P with respect to Q, that is, $P(A) = 0$ whenever $Q(A) = 0$, for every measurable set A.

In this light, the first expression in (1.49) tells us that high values of mutual information are the consequence of the joint density function concentrating mass somehow differently than what independence would suggests: thus **high mutual information** characterizes situations far from independence, supporting the view of a **strong degree of dependence** among the r.v.'s. See [1],[2] and [60].

The second and third expressions for I(X,Y) in (1.48) provide further insight; it is easy to reformulate the second one [4] as

$$\begin{aligned} I(X,Y) &= E\big[KL(f_{Y|X}(\cdot|X)|f_Y)\big] \\ &= \int_{\mathcal{X}} f_X(x)\Big\{KL\Big(f_{Y|X}(y|x)|f_Y(y)\Big)\Big\}dx \end{aligned} \tag{1.49}$$

$$= \int_{\mathcal{X}} f_X(x)\Big\{\int_{\mathcal{Y}} f_{Y|X}(y|x)\log_2\frac{f_{Y|X}(y|x)}{f_Y(y)}dy\Big\}dx \tag{1.50}$$

(1.48) expresses mutual information as an *average* - with respect to the distribution of X - *of Kullback-Leibler divergence between the conditional distributions of Y given $X = x$ and the marginal distribution of Y*: in synthesis, I(X,Y) expresses how much, on average, knowledge of X is able to move the distribution of Y away from its marginal, i.e. *how strong is the influence of X on the statistical behaviour of Y.*

Low values of I(X,Y) mean that $f_{Y|X}(y|x)$ is pretty 'close', similar to $f_Y(y)$ whatever the value x, assumed by X: knowledge of X conveys poor information about Y. And vice versa in the case of high I(X,Y).

[4] the same can obviously be done with the third since it is just a matter of factorization

Chapter 2

Information and Hierarchical models

The choice between different hierarchical models according to Information measures is the main topic of this chapter.

More precisely we should say that we are interested in an a priori choice of hierarchical models, with a priori meaning before taking any observation.
So our main concern will be that of experimental design, in the particular framework of hierarchical experiments, and not of model choice with respect to a given set of data.
Note that, in evaluation of performances of decision procedure, this essentially justifies averaging over the distribution of the observables even in the Bayesian setting, that will almost always be assumed in the following: observations are still to be drawn so they represent unknown quantities. This topic is an instance of a choice-among-experiments problem, and consequently a brief introduction to the subject of comparison of experiments is provided, together with its specification to the problem of allocation of observations.
This latter has an intrinsically hierachical structure, so we are led to the presentation of main information-theoretical results on this particular class of models.

2.1 Hierarchical models

In the Bayesian framework, the idea of probabilistic reasoning of a hierarchical type was initially discussed by Good [26] as a way to express prior information in a more realistic and pragmatic fashion, in the effort to construct a theory of partially-ordered probabilities.

Indeed Good observes that a point evaluation of a probability (e.g. P(rain tomorrow)=0.304657) is hardly realistic, and consequently each probability statement has some uncertainty attached to it: this clearly leads directly to the probability of a probability value lying in a certain interval between 0 and 1: that is, to a probability on a probability. Obviously there is no need to end the probability-chain at the 'second level' just mentioned, and other levels can be added to refine our probability elicitation.

Thus the formulation of these statistical models is such that inclusion of prior information can be accomplished more consciously by the decision maker.

In his subsequent work (e.g. [27]) Good gave other contributions to the topic, providing a review of this methodology at the beginning of the '80 [28].

From the introduction of the first statistical models with a hierarchial structure to the present-day applications, the original idea has grown enormously in scope with the formulation of highly structured models with complex chain-type relations, but their essence can be captured by the following

Definition 2.1.1 *A Bayesian statistical model, $\{f(x|\theta), \pi(\theta)\}$, for the data X, is hierarchical if there exist $k \in \aleph$ and random variables (or vectors) Θ_j with $j = 1, \ldots, k$ such that*

$$X|\Theta = \theta \sim f(x|\theta) \quad ; \quad \Theta_j|\Theta_{j+1} = \theta_{j+1} \sim \pi_j(\theta_j|\theta_{j+1}) \quad j = 0, \ldots, k-1 \tag{2.1}$$

with $\Theta_0 \equiv \Theta$ and $\Theta_k \equiv \theta_k$ known constant(s).

k represents the number of levels or stages.

That is, if the prior distribution $\pi(\theta)$ [1]*, via the decomposition in conditional distributions* π_j*, admits the representation*

$$\pi(\theta) = \int \pi_0(\theta|\theta_1)\pi_1(\theta_1|\theta_2)\cdots\pi_{k-1}(\theta_{k-1}|\theta_k)d\theta_1\cdots\theta_{k-1} \tag{2.2}$$

Remark 2.1.1 *A part from the hierarchical representation of the prior distribution, expression (2.2) shows that hierarchical models are in full agreement with the Bayesian paradigm, and consequently they enjoy the optimality properties typical of the Bayesian approach.*

Remark 2.1.2 *The above representation of this class of models may be misleading: the density for the observables* $f(x|\theta)$ *need not be assumed as the density of one observation. On the contrary it stands for the density of the whole set of observables, and a huge variety of statistical models can be formulated by specyfing different conditional probability relations -* $f(x|\theta)$ *- between the vetor-valued* X *and* Θ *(and between* Θ_j *and* Θ_{j+1} *at higher levels* j*,* $j = 0, \ldots, k$*,* $\pi_j(\theta_i|\theta_i)$*).*

To make the arguments of the above remark more concrete, it is frequently possible to decompose the prior information, available to the decision maker, into a structural part and a subjective part: the former usually regards the procedures or processes that seem to have generated the data, e.g. functional relationships between variables or parameters in a model or chosen design in an experiment, while the latter gives the statistician the opportunity to express his opinion about the quantities remaining unknown in the model, e.g. plausible intensity of the functional relationships mentioned above or parameters in the assumed disturbance distribution.

[1] Note that more formally we should write $\pi(\theta|\theta_k)$ for the prior on model parameters (and $\pi(\theta_j|\theta_k)$ for the j-th level marginal), making explicit the conditioning on hyperparameters at the highest level, θ_k. Anyway, for ease of notation, we will omit the conditioning since dependence on some θ_k is implied once the order of the hierarchical model has been specified.

Usually the structural information is transferred into the lower levels of the models - those closer to observables - while the subjective part is expressed into the higher levels.
The possibility of expressing chain-type probability evaluations (e.g. a probability evaluation of the plausibility of another probability statement) is at the origin of the great flexibility of hierarchical models in representing complex prior opinions in a much clearer way than usual statistical models. As mentioned, such a property was the motivation for their introduction by Good, and it is also strongly emphasized by Lindley and Smith [43],and Smith [61] in their seminal papers on Bayesian hierarchical linear models.
We present here their original formalization since, notwithstanding the copious development of generalizations and applications of it, it still remains a very useful tool in statistical analysis, it admits closed-form expressions for posterior distributions and it exhibits a relatively straighforward interpretation of how different pieces of prior structural information, taken into account during the modelling process, lead to alternative summary of data and inferences.
The mentioned flexibility of hierarchical models can thus be appreciated - the same general linear structure encompasses different potential situations - and it gives a very transparent, though heuristic, view of information fluxes in this model class.
Furthermore, in the following we will deal explicitly with a particular case of the general linear model.

We will limit our attention to the 2-level hierarchical model under the assumption of Normality.

Definition 2.1.2 *A 2-level hierarchical linear model for the n-vector of observables* $\mathbf{X}$ *is defined by*

- $\mathbf{X}|\mathbf{\Theta} = \theta \sim \mathcal{N}_n(\mathbf{A}\theta, \mathbf{\Sigma})$
- $\mathbf{\Theta}|\mathbf{\Theta_1} = \theta_\mathbf{1} \sim \mathcal{N}_p(\mathbf{A_1}\theta_\mathbf{1}, \mathbf{\Sigma_1})$

- $\mathbf{\Theta_1}|\mathbf{\Theta_2} = \theta_\mathbf{2} \sim \mathcal{N}_{p_1}(\mathbf{A_2}\theta_\mathbf{2}, \mathbf{\Sigma_2})$

where $\mathbf{A}, \mathbf{A_1}$ *and* $\mathbf{A_2}$ *are known "design" matrices of dimensions* $(n \times p)$, $(p \times p_1)$ *and* $(p_1 \times p_2)$ *respectively, and* $\mathbf{\Sigma}, \mathbf{\Sigma_1}$ *and* $\mathbf{\Sigma_2}$ *are known covariance matrices of opportune dimensions.*

The flexibility of this structure can be illustrated by two examples presented originally by Lindley and Smith [43].

Example 2.1.1 Exchangeable multiple regression equations.
Prior information could suggest a partitioning of $\mathbf{X}$ *as* $[\mathbf{X_1}, \ldots, \mathbf{X_m}]$, *where each subvector* $\mathbf{X_h}$, *of possibly different dimension* n_h $(h = 1, \ldots, m)$, *is composed by observations coming from the same unit, for example.*
In this case $\sum_h n_h = n$, $\mathbf{A}$, *an* $(n \times p)$ *matrix, has the block diagonal representation* $\mathbf{A} = diag(\mathbf{B_1}, \ldots, \mathbf{B_m})$ *with* $\mathbf{B_h}$ $(n_h \times q)$, *and* $\mathbf{\Theta} = (\mathbf{\Delta_1}, \ldots, \mathbf{\Delta_m})$ *is a* $(q \cdot m =)$ p*-dimensional vector, with each* $\mathbf{\Delta}_h$, q*-dimensional.*
Concisely,

- $\mathbf{X_h}|\mathbf{\Theta} = \theta \sim \mathcal{N}_{n_h}(\mathbf{B_h}\delta_h, \mathbf{C_h}) \quad h = 1, \ldots, m$
- $\mathbf{\Delta}_h|\mathbf{\Theta_1} = \theta_\mathbf{1} \sim \mathcal{N}_q(\theta_\mathbf{1}, \mathbf{\Sigma_1}) \quad h = 1, \ldots, m$

Clearly a second level is usually added specifying a prior distribution for Θ_1 *conditional on* Θ_2*: this can be a conjugate normal prior or a non-informative prior, according to the available information.*
So we are considering m *multiple regressions with the same covariates, whose parameter vectors,* Δ_h $h = 1, \ldots, m$, *are assumed exchangeables.*

If the weights of different covariates, in the columns of $\mathbf{A}$, can be considered similar with respect to the influence on the response, an assumption of exchangeability between parameters could be reasonable.

Example 2.1.2 Exchangeable parameters.

The distributional assumptions for the observables $\mathbf{X}$ *given* $\mathbf{\Theta}$ *remains essentially the same as the general model in definition 3.1.2.*

It is at the first stage of the hierarchy that exchangeability betweeen parameters is introduced via the conditional prior

$$\mathbf{\Theta}|\Theta_1 = \theta_1 \sim \mathcal{N}_p(\theta_1 \cdot \mathbf{1}_p, \mathbf{\Sigma_1}) \tag{2.3}$$

where $\mathbf{1}_p$ *is a p-dimensional vector of ones.*

Note that Θ_1 *is a random variable representing the common mean of the regression coefficients, the diagonal elements of* $\mathbf{\Sigma_1}$ *- the variances of the components* $\Theta^{(i)}$*,* $i = 1, \ldots, p$*, of* $\mathbf{\Theta}$ *- express how much representative* θ_1 *is considered as a value for each coefficient, and finally the covariances in* $\mathbf{\Sigma_1}$ *quantify the degree of similarity between each pairs of coefficients.*

A second stage prior then expresses opinion, if any, about the plausible value of this common mean, and the intensity of this opinion.

Remark 2.1.3 *Note how the two preceding examples differ from a standard, non-hierarchical, Bayesian analysis of the same problems.*

In example 2.1.1 a standard Bayesian analysis of the linear model would assume that each regression equation had the same parameter vector - $\mathbf{\Delta_h} = \mathbf{\Delta}$ $h = 1, \ldots, m$ *- and then would have specified a prior on it.*

- $\mathbf{X_h}|\mathbf{\Theta} = \theta \sim \mathcal{N}_{n_h}(\mathbf{B_h}\delta, \mathbf{C_h}) \quad h = 1, \ldots, m$
  (or more compactly: $\mathbf{X}|\mathbf{\Theta} = \theta \sim \mathcal{N}_n(\mathbf{A}\delta, \mathbf{\Sigma}) \quad h = 1, \ldots, m$ *with* $\mathbf{A}' = [B_1', \ldots, B_m']$*)*
- $\mathbf{\Delta}|\mathbf{\Theta_1} = \theta_\mathbf{1} \sim \mathcal{N}_q(\theta_\mathbf{1}, \mathbf{\Sigma_1})$

Instead of postulating equality of regression parameter vectors, the hierarchical formulation only assumes exchangeability, that is some sort of similarity between them: with this

respect, it is the covariance matrix $\mathbf{\Sigma_1}$ *that quantifies the extension of this similarity. Indeed letting* $\mathbf{\Sigma_1} \to \mathbf{0}$*, the first level prior tends to a degenerate distribution on* $\theta_\mathbf{1}$*: we are then back to the non-hierachical model.*

Through the introduction of an additional level the assumptions of a standard analysis have been weakened, in some way, from equality to closeness: generally this happens on the basis of structural prior information, whose believed validity can be quantified and introduced into the analysis through the choice of $\mathbf{\Sigma_1}$*. This also suggests that care must be paid in the choice of this parameter since it seems to be able to greatly influence the analysis, changing the model, in some continuous way, from a standard one to a hierarchical one, finally to one in which all regression equations are treated separately (*$\mathbf{\Sigma_1}$ *with large elements).*

Finally, in the hierarchical model, the 2nd level prior plays essentially a role analogous to the prior in the standard model: it expresses more subjective information about plausible values for parameters. However in this formulation it has been "moved up" one stage in the hierarchy: the "purely subjective" judgments have been "moved farther from observations".

This could suggest that these judments could have smaller "informative power" and influences on the inferences drawn from the analysis.

In example 3.1.2 the reasoning is almost identical: instead of assuming exact equality of all regression coefficients - which is quite a strong, hardly justifiable opinion - they are considered similar in magnitude, with the 2nd stage prior expressing the order of this magnitude.

In a quite informal fashion, the preceding examples and remarks have introduced the topic of prior information representation through and information fluxes in hierarchical models.

This topic has already been formalized and partially studied at the beginning of the '80,

and the results are presented in the next subsection.

Even if the exposition in this subsection has regarded only linear hierachical models under the assumption of normality, it must be noticed that obviuously this class is by no means limited to this case, and the general definition 3.1.1 covers a huge variety of models that do not assume normality nor linearity.

We limited attention to the linear normal subclass since it is characterized by a certain intuitive interpretability, and since some of the results that will be presented can be placed in this framework.

Indeed the literature on the theory and applications of hierarchical models is quite extensive. Some references are Gelman et al. [19] and Congdon [11] for the general theory, while some recent applications can be found in Clark and Gelfand [9] for the environmental sciences or in Rossi et al. [56] for marketing.

2.1.1 Hierarchical Models and Information Measures

Following the suggestions of Lindley [41] on how to measure the amount of information provided by an experiment, at the beginning of the '80s Goel and DeGroot [24] and Goel [22] obtained some results about the behaviour of the information, contained in a set of observations, with respect to parameters at different levels in a general hierarchical model. Their approach was taken up by Colin [10] at the end of the same decade, when he generalized their findings in some directions.

The main features of this approach are given by

- the specific measures of information,
- a very useful property possessed by hierarchical models.

In the following remark, we make explicit a fact that can be immediately recovered from definition 2.1.1, equation 2.1.

Remark 2.1.4 ***A k-level hierarchical model is a Markov chain*** *(of order 1 in the formalization of Definition 2.1.1) since* $\mathbf{\Theta}_j|\{\mathbf{\Theta}_l \quad l > j\} \sim \mathbf{\Theta}_j|\mathbf{\Theta}_{j+1}$*: the distribution of the parameters at the* j*-th level conditioned on the parameters on all higher levels only depends on those belonging to the immediately upper one; it can synthetically be written as*

$$\mathbf{\Theta_k} \leftrightarrow \mathbf{\Theta_{k-1}} \leftrightarrow \cdots \leftrightarrow \mathbf{\Theta_1} \leftrightarrow \mathbf{\Theta} \leftrightarrow \mathbf{X} \tag{2.4}$$

In this framework, Markovity concretely means that we learn about parameters at the j-th level only through the information we have been able to gain about $(j-1)$-th level parameters, since $\mathbf{\Theta}_j|\mathbf{X}, \mathbf{\Theta}_{j-1} \sim \mathbf{\Theta}_j|\mathbf{\Theta}_{j-1}$.

The measures of information they considered are examples of two general classes.

Utility & Probability-based measures of Information

The first, studied by DeGroot [15] [17], has not Information-theoretical considerations as its fundational point: it instead moves from an Utility argument.
An original formalization of these type of measures can be found in Raiffa and Schlaifer [49]:

- the *conditional value of sample information* (**C.V.S.I.**) is defined as the difference between expected terminal utilities (with respect to the posterior distribution) of the posterior Bayes action and prior Bayes action (increase in expected utility afforded by a particular observation $\mathbf{x}$);
- the *expected value of sample information* (**E.V.S.I.**) is then defined simply as the expected value of C.V.S.I. with respect to the marginal distribution of $\mathbf{X}$ (expected increase in utility afforded by experiment $\mathbf{X}$).

In the analysis of DeGroot and Goel, Utility considerations enter mainly in the effort to answer the question

"How to quantify the **utility** provided by a distribution?".

where the considered distribution is typically a prior or posterior for the parameters.
In turn, this utility is perceived as inversely proportional to the uncertainty implied by the distribution, and the goal then becomes

"How to quantify the **uncertainty** expressed by a distribution?".

Clearly the conclusion follows that the Information provided by a statistical experiment is equal to (some monotone non-decreasing function of) the expected reduction in Uncertainty it can lead to.
In particular DeGroot proved [15], [24] that under some mild and reasonable requirements - for example, non-negativity of expected Information from an experiment - any concave function can measure uncertainty. This argument can be formalized

Definition 2.1.3 *Let Ξ be a convex family of probability distributions π for a parameter Θ. A real-valued measurable function $U(\pi)$ defined on Ξ is called an* ***Uncertainty function*** *if it is concave.*

and directly

Definition 2.1.4 *For a given Uncertainty function U, the conditional value of sample Information provided by the observation $\mathbf{x}$, $I(\mathbf{x}, \pi; U)$, (C.V.S.I.) is given by*

$$I(\mathbf{x}, \pi; U) = U(\pi) - U(\pi(\cdot|\mathbf{x})) \tag{2.5}$$

while the expected value of sample Information provided by the experiment $\mathbf{X}$, $I(\mathbf{X}, \pi; U)$, (E.V.S.I.)is

$$I(\mathbf{X}, \pi; U) = U(\pi) - E[U(\pi(\cdot|\mathbf{X}))] \tag{2.6}$$

where expectation is taken with respect to the marginal distribution of $\mathbf{X}$ induced by π.

Example 2.1.3 *Choosing* $U_1(\pi) = var(\Theta)$, $U_2(\pi) = \log\{var(\Theta)\}$, $U_3(\pi) = -\int \pi(\theta)\log\pi(\theta)d\theta$, *one obtains*

$$\begin{aligned} I(\mathbf{X}, \pi; U_1) &= var(\Theta) - E\left[var(\Theta|\mathbf{X})\right] \\ I(\mathbf{X}, \pi; U_2) &= E\left[\log\left\{\frac{var(\Theta)}{var(\Theta|\mathbf{X})}\right\}\right] \\ I(\mathbf{X}, \pi; U_3) &= E\left[\log\frac{\pi(\theta|\mathbf{X})}{\pi(\theta)}\right] = I(\mathbf{X}, \Theta) \end{aligned} \tag{2.7}$$

From the previuos example we can then conclude that Mutual Information belongs to this class of Information Measures.

Probability-based measures of Information

The main results of Goel and DeGroot, however, refer to measures belonging to the second class mentioned above: this latter include those sharing a general characterization as **distances** (or pseudo-distances) **between distributions**.
As in the preceding section, a distinction can be made between

- *conditional amount of sample information* (C.A.S.I.) as some chosen distance between the posterior distribution for a specific sample point $\mathbf{x}$ and the prior;
- *expected amount of sample information* (E.A.S.I.) as the expected distance between posterior and prior value (expected C.A.S.I.)

Remark 2.1.5 *It is clear that from each specific distance function, chosen to define a C.A.S.I. measure, a corresponding E.A.S.I measure directly follows.*

Remark 2.1.6 *The terms "value" and "amount" are used to distinguish an utilitarian evaluation of information from one lacking of this character and measuring only an "objective" quantity: this should vaguely parallel the distinction between decision problems and "purely inferential" questions [4], where indeed only information is sought and no*

specific utility function is available.

However, as discussed by DeGroot [17], this boundary could not be so neat as perceived: choosing a distance function looks very much like choosing a utility function, and the possibility of quantifying information via some arbitrary concave (uncertainty) function seems to suggest that information cannot be evaluated unless some contingent however-abstractly-defined goal has been elicited.

There are indeed many available choices but we will mention only three examples.

Example 2.1.4 *In this framework, the* ***Kullback-Leibler Information*** *for discrimination (or relative entropy) between $\pi(\cdot|\mathbf{x})$ and $\pi(\cdot)$ given by definition in 1.3.4 assumes the expression:*

$$KL(\pi(\cdot|\mathbf{x}) \mid \pi(\cdot)) = \int \pi(\theta|\mathbf{x}) \log \frac{\pi(\theta|\mathbf{x})}{\pi(\theta)} \, d\theta \tag{2.8}$$

so that $I(\mathbf{X}, \pi; KL) = E\left[KL(\pi(\cdot|\mathbf{X}) \mid \pi(\cdot))\right] = I(\Theta, \mathbf{X})$.

Remark 2.1.7 *Together examples 2.1.3 (with U_3) and 2.1.4 highlight how Mutual Information can be characterized both as an expected utility measure and average distance measure. Among the ones considered here, it is the only one sharing both characterizations.*

An additional interpretation of Mutual Information is in terms of cumulative relative entropy risk, as discussed by Haussler and Opper [31], for example.

Example 2.1.5 *Another entire class of Information measures arises by consideration of* ***Renyi's entropy functions****, which, for $\alpha \neq 1$, are defined by*

$$f_\alpha(\pi(\cdot|\mathbf{x}), \pi(\cdot)) = \frac{1}{\alpha - 1} \log \int \pi(\theta|\mathbf{x})^\alpha \pi(\theta)^{1-\alpha} d\theta \tag{2.9}$$

and the corresponding information measure is clearly given by $I_\alpha(\mathbf{X}, \Theta) = E[f_\alpha(\pi(\cdot|\mathbf{X}), \pi(\cdot))]$.

Remark 2.1.8 *Note that for $\alpha \to 1$, we have*

$$f_\alpha(\pi(\cdot|\mathbf{x}), \pi(\cdot)) \to KL(\pi(\cdot|\mathbf{x}) \mid \pi(\cdot))$$

so that we recover the Kullback-Leibler Information as a limiting case of Renyi's entropies.

All the preceding measures are indeed special cases of one very general class of divergences between probability distributions, introduced by Csiszars [14]

Definition 2.1.5 *The f-divergence between the probability densities $\pi(\cdot|\mathbf{x})$ and $\pi(\cdot)$ is defined by*

$$D_f(\pi(\cdot|\mathbf{x}), \pi(\cdot)) = \int \pi(\theta) f\left(\frac{\pi(\theta|\mathbf{x})}{\pi(\theta)}\right) d\theta \tag{2.10}$$

where f is an arbirtary convex function defined on $(0, +\infty)$.
The corresponding information measure is denoted
$I_f(\mathbf{X}, \Theta) = E[D_f(\pi(\cdot|\mathbf{X}), \pi(\cdot))]$.

In the preceding definition, choosing $f(x) = x \log x$ leads to the Kullback-Leibler divergence, while choosing $f(x) = x^\alpha$ leads to the so-called "α"-divergences of Csiszars, or to the Renyi's entropies after a monotone transformation.

If we now focus attention on hierarchical models, it becomes immediately clear that we can measure the amount of information provided by data $\mathbf{x}$ about the parameters at the j-th level, $\mathbf{\Theta}_j$, computing the f-divergence between $\pi_j(\cdot \mid \mathbf{x})$ and $\pi_j(\cdot)$.

$$D_f(j; \mathbf{x}) \equiv D_f(\pi_j(\cdot|\mathbf{x}), \pi_j(\cdot)) = \int \pi_j(\theta_j) f\left(\frac{\pi_j(\theta_j|\mathbf{x})}{\pi_j(\theta_j)}\right) d\theta_j \tag{2.11}$$

Note that, according to the notation adopted in definition 2.1.1, $\pi_j(\cdot)$ is the marginal prior distribution of $\mathbf{\Theta}_j$, obtained as

$$\pi_j(\theta_j) = \int \pi_j(\theta_j|\theta_{j+1})\pi_1(\theta_{j+1}|\theta_{j+2}) \cdots \pi_{k-1}(\theta_{k-1}|\theta_k) d\theta_{j+1} \cdots \theta_{k-1} \tag{2.12}$$

In the spirit of this discussion, we consider "more informative" those data which "change more significantly our prior opinion", leading to a posterior distribution, $\pi(\cdot \mid \mathbf{x})$, farther away, in the f-divergence sense, from the prior, $\pi(\cdot)$. More surprising experimental results are attached more informative content.

It is for this broad class that Goel and DeGroot proved the following

Theorem 2.1.1 *For any convex function on* $(0, +\infty)$*,* $D_f(j; \mathbf{x})$ *is a decreasing function of* j*,* $j = 1, \ldots, k$*.*

So the influence of observations on posterior distributions becomes smaller and smaller as we proceed to higher levels of the hierarchy.
Averaging with respect to the marginal distribution of $\mathbf{X}$ immediately leads to

Theorem 2.1.2 *For any convex function on* $(0, +\infty)$*,* $I_f(\mathbf{X}, \mathbf{\Theta}_j)$ *is a decreasing function of* j*,* $j = 1, \ldots, k$*.*

On average, we learn less about parameters at high stages of the model than we learn about parameters at low stages.
Notice how this confirms a general intuition: if the outcome of the experiment $\mathbf{X}$ is intepreted as a message carrying information about unknown quantities, its original content becomes diluted as it goes through the stages of the model.

If now we focus attention on the specific case of Mutual Information, it can be easily seen how the previous result immediately follows from Markovity and the Data Processing Inequality (Theorem 1.2.3).
From Markovity we have that, for $j = 1, \ldots, k-1$, $\mathbf{X} \leftrightarrow \mathbf{\Theta}_j \leftrightarrow \mathbf{\Theta}_{j+1}$ and consequently $\mathbf{X} \perp \mathbf{\Theta}_{j+1} | \mathbf{\Theta}_j$, whence

$$\begin{aligned} I((\mathbf{\Theta}_j, \mathbf{\Theta}_{j+1}), \mathbf{X}) &= I(\mathbf{\Theta}_{j+1}, \mathbf{X}) + I(\mathbf{\Theta}_j, \mathbf{X} | \mathbf{\Theta}_{j+1}) \\ &= I(\mathbf{\Theta}_j, \mathbf{X}) + I(\mathbf{\Theta}_{j+1}, \mathbf{X} | \mathbf{\Theta}_j) = I(\mathbf{\Theta}_j, \mathbf{X}) \end{aligned} \tag{2.13}$$

since $I(\boldsymbol{\Theta}_{j+1}, \mathbf{X}|\boldsymbol{\Theta}_j) = 0$.

Thus $I(\boldsymbol{\Theta}_j, \mathbf{X}) \geq I(\boldsymbol{\Theta}_{j+1}, \mathbf{X})$, $j = 1, \ldots, k-1$, by non-negativity of Mutual Information.

Remark 2.1.9 *A by-product of the above discussion is the following equality*

$$I((\boldsymbol{\Theta}_j, \boldsymbol{\Theta}_{j+1}), \mathbf{X}) = I(\boldsymbol{\Theta}_j, \mathbf{X}) \tag{2.14}$$

which easily generalizes to

$$I((\boldsymbol{\Theta}, \boldsymbol{\Theta}_1^k), \mathbf{X}) = I(\boldsymbol{\Theta}, \mathbf{X}) \tag{2.15}$$

where $\boldsymbol{\Theta}_1^k \equiv (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_k)$.

The former says that in the amount of information provided by $\mathbf{X}$ *about the parameter* $\boldsymbol{\Theta}_j$ *is already included that provided about* $\boldsymbol{\Theta}_{j+1}$*: this strenghten the idea that we learn about* $\boldsymbol{\Theta}_{j+1}$ *only through what we have been able to learn about* $\boldsymbol{\Theta}_j$*.*

The latter states something even more drastic: the hierarchical formalization of the prior - which seems to suggests that we can learn more since we can learn something about each of many parameters - adds nothing to the global amount of information we can extract from data: this is somehow fixed by the marginal prior distribution for $\boldsymbol{\Theta}$*. At most it can provide a more useful decomposition (in terms of interpretability, for example) of the gained information.*

There is finally a different way of expressing the diminishing impact of observations on posterior distributions: instead of comparing posterior and prior distributions, one can compare - that is, measure the distance between - two different posterior distributions, $\pi_j(\cdot|\mathbf{x})$ and $\pi_j(\cdot|\mathbf{x}^*)$, corresponding to different data points $\mathbf{x}$ and $\mathbf{x}^*$, and study this distance as a function of the level j in the hierarchy.

In the light of previuosly presented results, one can expect that different observations lead to less "distant" posteriors at higher levels than at lower levels. This statement is confirmed and made precise by the following theorem of Goel [22].

Theorem 2.1.3 *For any convex function $f(\cdot)$ on $(0, +\infty)$, the f-divergence between posterior distributions at level j corresponding to different observations $\mathbf{x}$ and $\mathbf{x}^*$, $\pi_j(\cdot|\mathbf{x})$ and $\pi_j(\cdot|\mathbf{x}^*)$,*

$$D_f(j; \mathbf{x}, \mathbf{x}^*) \equiv D_f(\pi_j(\cdot \mid \mathbf{x}), \pi_j(\cdot \mid \mathbf{x}^*)) = \int \pi_j(\theta_j \mid \mathbf{x}^*) f\left(\frac{\pi_j(\theta_j \mid \mathbf{x})}{\pi_j(\theta_j \mid \mathbf{x}^*)}\right) d\theta_j \qquad (2.16)$$

is a decreasing function of j.

Clearly in this case a derived E.A.S.I. measure is difficult to imagine, and it does not seem to have a specific useful meaning.

Remark 2.1.10 *For sake of completeness, it must be clarified that, even if the previuos theorems are valid for many information measures, it is not possible to extend their validity to **all** information measures: examples are given in Goel and DeGroot [24] in which the amount of information gained about level 1 is smaller than the amount about level 2. We note in passing that these examples pertain information measures derived from uncertainty functions, and not from divergences between probability distributions: we are not aware of results proving that consideration of divergences only can rule out such counterintuitive situations, even if the general results of Goel and DeGroot seem to cover a huge variety of divergence measures. However we have not studied the topic further.*

2.2 Comparison of Hierarchical Models

The results presented in the preceding section assumed a fixed hierarchical model and established some relations - essentially a kind of ordering - among information quantities defined for the involved parameters.

In this section our aim will be the comparison of different hierarchical structures, according to one of the information measures discussed in Chapter 1 and in the previous section:

Mutual Information.
Unfortunately it seems that the properties leading to the theorems of Goel and DeGroot - Markovity and the use of Jensen's Inequality, made possible by concavity of $f(\cdot)$ - provide no help of comparable use.
The goal of such comparisons and the extent to which they are meaningful will be clarified in the following.

As in the most part of this work, here the approach will be Bayesian. Nonetheless, in the next subsection, we briefly present the classical decision-theoretical approach since it is the one followed by DeGroot in one of the seminal papers in this area of studies [16].
This introduction to the field named 'Comparison of Experiments' is really reductive, but nor complete silence nor a full presentation appeared to be viable choices: the former because the present work was partially born as a consequence of the ideas of the above mentioned paper of DeGroot; the latter since the classical theory is very technical, and presenting it in an exhaustive way would have taken too much time and space.
We hope that this short introduction will attract the attention of the readers to a very valuable field, and to its basic problems that are at the foundations of Statistics.
At the same time, it will let us motivate the choice to perform a Bayesian analysis of this class of problems, e.g. a Bayesian comparison of experiments. Indeed it can be said that, in some cases, the classical theory is 'incomplete', in the sense of not being able to provide a definite judgment in a comparison between two experiments: some experiments cannot be compared.

2.2.1 Comparison of Experiments

The field of Comparison of Experiments was initiated by Blackwell in the middle of the past century [5],[6], following previous work and suggestions of Bohnenblust, Shapley,

Sherman [7]. Thanks to the contributions of many researchers, among which LeCam [38] who gave it much of its present form, it evolved rapidly into a quite rich mathematical discipline, an extensive compendium of which is represented by the treatise of Torgersen [64] wholly dedicated to the subject.

In the Comparison-of-Experiments framework the statistician is faced with the problem of chosing which experiment to perform among some possible ones, and his goal is to choose the best one.

In this setting an experiment is just a set of probability measures on the same probability space, each of them typically representing the behaviour of some observable in a particular state of the world. For example, as usual in Statistics, a parametric family of distributions $\{P_\theta, \theta \in \Theta\}$ on some probability space $(\mathcal{X}, \mathcal{A})$ can be interpreted as an experiment with outcomes in $\mathcal{X}$, with possible states of the world indexed by $\Theta \subseteq \Re^k$.

Performing the experiment $\mathcal{E}$ actually means observing the value of the r.v. X with distribution P_θ.

And chosing an experiment means chosing between families of probability distributions, according with some criterion of optimality, before observing any outcome or value. What ties together different experiments is clearly their common set of possible states of the world: in the parametric case, they all share the same parameter space Θ and thus the probability distributions in each experiment (read: family) are indexed by the same e.g. real-valued index θ.

The optimality criterion proposed by the founders of the subject is very strong.

Consider just two experiments $\mathcal{E} = \{(\mathcal{X}, \mathcal{A}), P_\theta, \theta \in \Theta\}$ and $\mathcal{F} = \{(\mathcal{Y}, \mathcal{B}), Q_\theta, \theta \in \Theta\}$. For a decision space $\mathcal{D}$ and a loss function $L : \Theta \times \mathcal{D} \to \Re^+$, the risk function of a (possibly randomized) decision rule ρ_x based on experiment $\mathcal{E}$ is defined as usual as

$$R(\theta, \rho) = \int_{\mathcal{X}} \int_{\mathcal{D}} L(\theta, d) d\rho_x(d) dP_\theta(x)$$

and an analogous definition will clearly hold for decision rules based on $\mathcal{F}$.

With a little abuse of notation, let $R(\mathcal{E}, L)$ denote the set of all functions that are actual risk functions, or larger than risk functions, obtainable by decision rules based on X when L is the assumed loss function.

Bohnenblust, Shapley and Sherman, and Blackwell after them, defined a preference relation on the set of possible experiments in the following way:

Definition 2.2.1 *$\mathcal{E}$ is said to be at least as informative [2] than $\mathcal{F}$ if, for every loss function L, $R(\mathcal{E}, L) \supset R(\mathcal{F}, L)$.*

Basically this definition requires, for $\mathcal{E}$ to be *more informative* than $\mathcal{F}$, that every risk function obtainable with a decision rule based on $\mathcal{F}$ can also be obtained with a decision rule based on $\mathcal{E}$: in this case, we could *reproduce* - or even improve on - the performance of every decision rule $\tilde{\rho}_y$ based on Y via a decision rule ρ_x based on X, for every decision problem (identified by the loss function, L). [3]

This clearly leads to the conclusion that $\mathcal{E}$ should be preferred to $\mathcal{F}$, since the former is at least as good as the latter (where good here is intended in the usual risk-comparison sense), for every decision problem.

As mentioned, this criterion is very strong since a short examination almost immediately qualifies it as an extended Admissibility criterion: while, in general studies on Admissibility, the loss function is assumed fixed (the decision problem is intrinsically made precise by this assumption), and we need to choose, if possible, an optimal procedure with respect to this loss function, in the present framework we are considering all possible loss functions (e.g. all possible decision problems!) and we are trying to establish if the outcomes of one experiment are more informative, in the sense of leading to a smaller expected loss (risk), than those of another one, for every decision problem that could be considered interesting.

[2] We will interchangeably use the words 'more informative' for the same concept.

[3] Partially different criteria of Informativeness of statistical experiments can be formulated: for a relatively recent review, see LeCam [39], for example.

These things considered, it should not be surprising that once has been possible to conclude that an experiment $\mathcal{E}$ is more informative than another one $\mathcal{F}$, it will also be possible to conclude that the former will be better than the latter according to any other usual optimality criterion: optimal Bayesian procedure (Bayes risk minimizers) based on $\mathcal{E}$ will outperform optimal Bayesian procedure based on $\mathcal{F}$, for every specified prior (and loss function) and analogoulsy for minimax procedure.
Indeed, it can be stated that the previous definition of 'more informative' is equivalent to the following:

Definition 2.2.2 *$\mathcal{E}$ is said to be at least as informative as $\mathcal{F}$ if, for every loss function L and every prior distribution on Θ, π, the expected Bayes risk from $\mathcal{F}$ is not less than that from $\mathcal{E}$.*

As usual the expected Bayes risk is defined as

$$r(\pi) = min_{\rho} E_{\pi}[R(\Theta, \rho)]$$

where π is the assumed prior distribution on Θ.
At the same time, everyone who has dealt with Admissibility problems for a while will need little effort to comprehend how complex this kind of problems can be and, most of all, how easily they can lead to no clear-cut solutions.
In particular, even after examination of just two experiments $\mathcal{E}$ and $\mathcal{F}$, we could end up finding that they are *not comparable*: typically, one will be better for some decision problems and the opposite will be true for some other problems. Note that it suffices to find just one decision problem for which $\mathcal{F}$ is better than $\mathcal{E}$ to encounter this 'non-comparability', even if paradoxically $\mathcal{E}$ is better than $\mathcal{F}$, in all other decision problems.
In other terms the induced preference relation on the set of considered experiments easily fails to be complete, and an optimal experiment frequently does not exist.

Both the presented definitions of 'more informative' are not very operational: in general it is difficult to directly exploit them to find out if one experiment is 'better' than another.
A sufficient condition is provided by Lehmann [40]:

Lemma 2.2.1 *$\mathcal{E}$ is at least as informative as $\mathcal{F}$ if there exist a function $\psi : \Re^{n+r} \to \Re^m$ and an r-dimensional random vector $\mathbf{Z}$, independent of $\mathbf{X}$ - the outcome of $\mathcal{E}$ - and whose distribution is free from θ, such that $\mathbf{Y} \overset{d}{\sim} \psi(\mathbf{X}, \mathbf{Z}) \quad \forall \theta \in \Theta$.* [4]

In the preceding lemma, the function ψ and the random vector $\mathbf{Z}$ provide one of the possible formal expressions of the concept of '**stochastic transformation**', that is a random transformation (basically a Markov kernel) of the outcome of $\mathcal{E}$: by means of a stochastic transformation, we reach the goal of 'reproducing the outcome of (and consequently the performances of procedures based on) $\mathcal{F}$' via outcomes of $\mathcal{E}$.
The case in which such a stochastic transformation exists is so relevant that it deserves its own

Definition 2.2.3 *If there exists a stochastic transformation as the one described in Lemma 2.2.1, then experiment $\mathcal{E}$ is said to be **sufficient** for experiment $\mathcal{F}$.*

So Lemma 2.2.1 just states that if $\mathcal{E}$ is sufficient for $\mathcal{F}$, then $\mathcal{E}$ is also at least as informative as $\mathcal{F}$.
If we also assume that the families of distributions defining the experiments are *dominated*, condition in Lemma 2.2.1 becomes necessary too, and

Theorem 2.2.1 *$\mathcal{E}$ is more informative than $\mathcal{F}$ if and only if $\mathcal{E}$ is sufficient for $\mathcal{F}$*

[4] In the lemma, the outcomes of experiments $\mathcal{E}$ and $\mathcal{F}$ previously varying generically in spaces $\mathcal{X}$ and $\mathcal{Y}$ have been assumed to be random vectors of dimensions n and m, respectively. With almost no loss of generality, this case characterizes almost all practical situations.

The notion of *Sufficiency between experiments* and its relation with the Information content characterizing an experiment is clearly more 'operational' than the original notion of Informative content based on sets of risk functions: from the abstract comparison of these latter for an infinite variety of loss functions we are lead to the search of just one stochastic transformation.

Nonetheless, even if the attention is now focussed on a narrower and well defined task, the search for such stochastic transformations is still a very hard task.

At the risk of being annoying, we recall again that nothing ensures the existence of such a stochastic transformation: via a particular transformation, we could get close, in some sense, to the distribution of $\mathbf{Y}$, but not be able to *exactly* 'reproduce' its behaviour in terms of $\mathbf{X}$. [5]

At the end of this brief presentation of classical 'Comparison of Experiments', we would like to focus on two observations:

- In the following, classical results, when available, will be very valuable. Our problem - comparing hierarchical models via Mutual Information - can simply be seen as a particular decision problem (a purely inferential problem), with a specific loss function (a proper logarithmic scoring utility function): if $\mathcal{E}$ is more informative than $\mathcal{F}$ in the classical sense, then it will also lead to a higher value of Mutual Information.

- On the contrary, when no result is available from the classical theory, our criterion would always provide a reasonable way to choose, on a Bayesian basis, among experiments: the ordering of experiments based on Mutual Information values is always complete, and sigles out an 'optimal one' in any comparison.

[5] Actually, the classical theory of Comparison of Experiments does not halt when faced with this 'black or white' situation: the notion of 'Deficiency' developed by LeCam [38] tries to cope with the problem of non-comparability in absolute sense, by measuring the degree of 'closeness' mentioned above.

2.2.2 Allocation of observations

We finally arrived at the particular class of hierarchical models which we wish to focus on in this work.

This subclass of hierarchical structures and of problems in the field of comparison of experiments is the one dealing with allocation of observations.

As described by DeGroot [16],

Definition 2.2.4 ***(Allocation Problem)*** *In an Allocation Problem, to investigate a phenomenon of interest, a researcher*

- *has a fixed number of observations, n, at her disposal;*
- *can allocate them to (at least n) different units or individuals*

To learn as much as possible, how many units should be used and how should the observations be allocated to the units?

The previous definition makes it clear that the Allocation Problem is a problem of **optimal experimental design**: in this framework, the researcher is planning to run an experiment and she wants to choose the 'best' one among a collection of available ones. This choice clearly implies **comparisons** between experiments and 'best' is defined on the basis of some **optimality criterion**.

In the specific situation considered here, experiments are hierarchical, as explained in the following.

Remark 2.2.1 *Since in design problems observations have not been taken yet, their values are unknown, and it makes sense to average over their distribution, even in a Bayesian approach.*

Example 2.2.1 *Suppose that a new drug has been devised to cure a particular disease. The drug will be introduced in a therapy followed in some hospitals. Before the drug is*

steadily adopted, it has been decided to run a trial on n individuals, chosen among patients in hospitals with the considered therapy, and to record its effectiveness.
How should the individuals be allocated to hospitals? Should all individuals be selected from a single hospital or should each individual be chosen from a different hospital? Or instead should groups of individuals be selected from different hospitals?
Note that in this example individuals represent observations and hospitals units.

We will indicate with $\mathcal{A}(n; k; n_1, \ldots, n_k)$ the allocation of n observations to k units, with n_i observations taken on the i-th unit ($i = 1, \ldots, k$): any choice of the number of units, k, between 1 and n, and observations per unit is possible, provided that $\sum_{i=1}^{k} n_i = n$. According to this notation, adopted following DeGroot [16], for example, $\mathcal{A}(n; 1; n)$ represents the case of all observations taken on a single unit while $\mathcal{A}(n; n; 1, \ldots, 1)$ that with n units and only one observation per unit.

To complete the description of the problem at hand, it remains to specify

- which are the statistical relationships between units, and between observations taken on the same unit.
- on what 'scale' we are going to measure the amount of learning, that is, which specific measure of Information we are going to adopt;

For an allocation $\mathcal{A}(n; k; n_1, \ldots, n_k)$, the former can be expressed as

- $\Theta^{(1)}, \ldots, \Theta^{(k)} | \mu \overset{i.i.d.}{\sim} f_{\Theta|\mu}(\cdot|\mu)$
- $(X_i)_1^{n_i} \equiv (X_{ij} \;\; j = 1, \ldots, n_i) | \Theta^{(i)} = \theta^{(i)} \overset{i.i.d.}{\sim} f_{X|\Theta}(\cdot|\theta^{(i)})$, $i = 1, \ldots, k$

According to the notation just introduced,

- $\Theta^{(i)}$ represents the unknown value of some characteristic for unit i, $i = 1, \ldots, k$;
- units are assumed to be a random sample from a population which can be modelled by a density $\pi_{\Theta|\mu}(\cdot|\mu)$;

- in general, the 'true' value of $\Theta^{(i)}$ is unobservable, since it is some abstractly defined quantity or since direct observation is too costly, for example;

- for each unit i, it is possible to obtain measurements, $X_{ij} \quad j = 1, \ldots, n_i$, somehow related to the unit's characteristic $\Theta^{(i)}$: this dependence is modelled by $f_{X|\Theta}(\cdot|\theta^{(i)})$.

Remark 2.2.2 *In the subsequent analysis, the* ***interest will focus on inferences about the parameter*** μ. *The presented model can be justified by the following situation: some characteristics of a population are under study, but it is not possible to obtain* ***direct observations*** *-* Θ*'s in our model - on the individuals belonging to it, but it is only possible to take* ***indirect observations*** *on each individual -* X*'s in our model - so that all inferences must be based on this set of indirect measurements only.*

General Results from Classical Theory

The Blackwell-DeGroot optimality criterion for the allocation problem states

Definition 2.2.5 *Consider the class* C_n *of all allocations* $\mathcal{A}(n; k; n_1, \ldots, n_k)$, $k = 1, \ldots, n$ *and* $\sum_{i=1}^{k} n_i = n$. *An allocation* $\mathcal{A}^*(n; k^*; n_1^*, \ldots, n_{k^*}^*)$ *is optimal in* C_n *if it is sufficient for every other allocation in* C_n.

By a strict application of this definition, to verify that an allocation is indeed optimal would require a long search for stochastic transformations between a supposed optimal allocation and every other allocation in C_n.

DeGroot [16] proves some results that are particularly useful in reducing the comparisons to be made. These theorems pertain the comparison of what can be called *'extremal' allocations*: allocations that maximally concentrate or distribute observations among units.

Theorem 2.2.2 *Let* n *be a fixed positive integer. Allocation* $\mathcal{A}(m; m; 1, \ldots, 1)$ *is optimal in* C_m *for every* $0 \leq m \leq n$ *if and only if it sufficient for* $\mathcal{A}(m; 1; m)$ *for every* $0 \leq m \leq n$.

Theorem 2.2.3 *Let n be a fixed positive integer. Allocation $\mathcal{A}(m;1;m)$ is optimal in C_m for every $0 \leq m \leq n$ if and only if it sufficient for $\mathcal{A}(m;2;m_1,m_2)$ for every $0 \leq m \leq n$.*

Using these theorems, we can then reduce the number of comparisons to just one, for a generic m: obviously nothing grants that the considered 'extremal' allocations will prove to be optimal; on the contrary, if these 'sufficiency check' have affirmative conclusions, with just one stochastic transformation we have been able to find the optimal allocation.

It is also possible to simplify the search for the optimal allocation by resorting to **Sufficient Statistics**: in particular, since consideration of Sufficient Statistics reduces the dimensions of the spaces considered in the problem, the task of constructing a stochastic transformation is recast between spaces of lower dimension.
Without stating explicitly the theorems, we shortly recall their contents:

- In allocation $\mathcal{A}(n;1;n)$, if $T_n = T_n(X_1,\ldots,X_n)$ is a sufficient statistic for Θ in the family $f_{X|\Theta}(\cdot|\theta)$, then it is also sufficient for μ.

- In allocations $\mathcal{A}(n;k;n_1,\ldots,n_k)$, if $T_m = T_m(X_1,\ldots,X_m)$ is sufficient for Θ for every $m = 1,\ldots,n$, then the vector

$$(T_{n_1}(X_{11},\ldots,X_{1n_1}),\ldots,T_{n_k}(X_{k1},\ldots,X_{kn_k}))$$

  is sufficient for μ.

The relevance of these results is that they allow to find sufficient statistics for the hierarchically specified dstribution of observables by focussing on particular components: the conditional distributions of the observables given the lowest level parameters.
Furthermore their importance remains valid even in a Bayesian analysis since the posterior distribution will be a function of the sample through sufficient statistics only.

Also the classical property of **Completeness** of a family of distributions can aid in the analysis.

Its role requires the notion of a *non-negatively estimable* function ϕ on Θ is by means of $\mathcal{E}$: if there exists a non-negative function f on $\mathcal{X}$ such that

$$\phi(\theta) = \int_{\mathcal{X}} f(x) dP_\theta(x) \qquad \forall \theta \in \Theta$$

then ϕ is said to be non-negatively estimable by means of $\mathcal{E}$.

Suppose now that the family of distributions $\{P_\theta, \theta \in \Theta\}$, defining experiment $\mathcal{E}$, is complete, and that $(\mathcal{Y}, \mathcal{B})$ in experiment $\mathcal{F}$ are an Euclidean space and the class of Borel sets respectively. Then Theorem 4.1 of DeGroot [16] states that $\mathcal{E}$ is sufficient for $\mathcal{F}$ if and only if, for each fixed $B \in \mathcal{B}$, $Q_\theta(B)$, as a function of θ, is *non-negatively estimable* by means of $\mathcal{E}$,that is if there exists a non-negative function $f(\cdot|\cdot)$ such that

$$Q_\theta(B) = \int_{\mathcal{X}} f(B|x) dP_\theta(x) \quad \forall \theta \in \Theta \tag{2.17}$$

It is evident how the function $f(\cdot|\cdot)$ captures the notion of stochastic transformation: it can indeed be thought as the conditional distribution of the random variable $Y|X = x$ or alternatively identified (almost surely) with the distribution of $\psi(Z, x)$, considered in Lemma 2.2.1.

Mutual Information Optimality

Choosing Mutual Information as an optimality criterion leads instead to

Definition 2.2.6 *For a fixed n, an allocation $\mathcal{A}^*(n; k^*; n_1^*, \ldots, n_{k^*}^*)$ is called an **optimal allocation** if it maximizes Mutual Information between the parameter μ and the observations $((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})$ among all allocations, that is, if*

$$I\Big(\mu, ((X_1)_1^{n_1^*}, \ldots, (X_{k^*})_1^{n_{k^*}^*})\Big) \geq I\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})\Big) \tag{2.18}$$

where $I(\mu, ((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k}))$ is the Mutual Information between μ and $((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})$.

Note that the allocation problem here dealt with, following a Bayesian approach, is an instance of Bayesian experimental design. See Chaloner and Verdinelli [8], for example, for a review on the topic.

In particular the criterion according to which is considered **optimal** the design maximizing Mutual Information between parameters and observations is known in the literature as **D-optimality** for the linear model.

This work could then be considered as an attempt to apply this criterion to hierarchical linear models, and to general hierarchical models.

Remark 2.2.3 *The restriction to inferences about μ only makes different allocation comparable: they all share the same parameter set .*

If also the unit-specific characteristics were under study, the comparison between different allocations would be somehow troublesome. For example, consider two allocations having k and $k-1$ units respectively: how the complete absence of information about the 'missing' unit in the parsimonious allocation would compare with the plausible increase of information about the remaining $k-1$, due to the reallocation of the same n observations to the fewer $k-1$ units?

Example 2.2.2 *(Example 2.2.1 continued) In this case μ represents the effectiveness of the drug on the overall population of patients with the considered disease (proportion of healed if the whole population of patients were treated with the drug); the hospital-specific parameters, Θ's, take into account potential heterogeneity in the field results (local proportion of healed patients), for example due to slight differences in the implementation of the therapy or to sensible diversity in the patients'characteristics (covariates) at different hospitals. Finally observations, X's, are the results of the therapy for each of the patient: healed or not.*

Note that both μ and the Θ's are unobserved and, in general, unobservables.

The intrinsic hierarchical nature of this sampling structure is now clearly becoming apparent, and a fully Bayesian approach follow just by specifying a prior $\pi_\mu(\cdot)$ on the hyperparameter μ.

Observation 2.2.1 *Each Allocation is a 2-level hierarchical model: it fits into the general notation by defining* $\mathbf{\Theta}_1 \equiv \mu$, $\mathbf{\Theta} \equiv (\Theta^{(1)}, \ldots, \Theta^{(k)})$, *and obviuosly* $\mathbf{X} \equiv (X_{ij} : \ i = 1, \ldots, k; \ j = 1, \ldots, n_i) = ((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})$.

A brief remark is necessary here to simplify notation in the sections that follow.

Remark 2.2.4 *Since we will be dealing with comparison of allocations and allocations are nothing but hierarchical models, clearly we will be comparing different statistical models. A strict formalism would then require to keep a distinct notation for each model: for example, we should write* $\Theta^{(i)}$ *in one model and* $\Theta'^{(i)}$ *in another with a different number of direct observations and/or different allocation of indirect observations. The same should be done for the* X*'s.*
Fortunately such a cumbersome notation is not necessary: in the present framework what is important are the distributions of these quantities. So $\Theta^{(1)}$ *and* $\Theta'^{(1)}$ *will have the same distribution in both models; moreover, in two allocations with* k *and* r, $r < k$, *direct observations, any subset of* r Θ*'s from the model with* k *of them will be distributed as those in the model with only* r.
An analogous reasoning applies to the X*'s.*

From the observation above and section 2.1.1, we can immediately state the following

Fact 2.2.1 ***(Indirect Observations are Less Informative than Direct Observations)***
For any allocation $A^{(n;k;n_1,\ldots,n_k)}$ *and for any convex function* $f(\cdot)$ *on* $(0, +\infty)$, *the amount of f-Information about* μ *provided by the indirect observations* $((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})$ *is*

smaller than or at most equal to the amount of f-Information about μ provided by the direct observations $(\Theta^{(1)}, \ldots, \Theta^{(k)})$, that is

$$I_f\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})\Big) \leq I_f\left(\mu, (\Theta^{(1)}, \ldots, \Theta^{(k)})\right) \tag{2.19}$$

and, in particular, for Mutual Information,

$$I\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_k)_1^{n_k})\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(k)})\Big) \tag{2.20}$$

Proof. Upon inversion of the direction of the Markov chain, the proof is obvious for the part on Mutual Information by use of the same techniques of equation 2.13, and for the general f-Information, by use of Theorem 2.1.2.

◇

We can now point out another fact that contributes partially to the general comparison of Allocations.

Fact 2.2.2 *For $r < k$, consider the two allocations $\mathcal{A}(n; k; n_1, \ldots, n_k)$ and $\mathcal{A}(n; r; n'_1, \ldots, n'_r)$, and ignore for the moment the indirect observations in both allocations.*
Then

$$I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(k)})\Big) \geq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \tag{2.21}$$

Proof. By the chain rule for Mutual Information,

$$\begin{aligned} I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(k)})\Big) &= I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \\ &\quad + I\Big(\mu, (\Theta^{(r+1)}, \ldots, \Theta^{(k)}) | (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \\ &\geq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \end{aligned} \tag{2.22}$$

since $I\Big(\mu, (\Theta^{(r+1)}, \ldots, \Theta^{(k)}) | (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \geq 0$.

◇

Considering an inequality analogous to 2.18 for allocation $\mathcal{A}(n; r; n'_1, \ldots, n'_r)$, we then have

$$I\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_r)_1^{n_r})\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(k)})\Big) \qquad (2.23)$$

so that we can observe that **allocations with fewer direct observations seem to possess** what could be called **a smaller 'informative potential'** than those with more direct observations.

Unfortunately this is not sufficient to conclude that they actually provide smaller information than those with more direct observations: this larger potential amount could not be accessed by indirect observations to the same extent as the smaller amount in the parsimonious allocation.

2.2.3 A simple model: Bernoulli Observables

We will analyse a very simple model that nonetheless illustrates the main issues arising in the kind of comparisons we are dealing with.

The component distributions at the different levels (prior and conditional distributions for direct and indirect observations) are assumed to be

- $\mu \sim \mathcal{U}(0, 1)$
- $\Theta \mid \mu \sim \mathcal{U}(0, \mu)$
- $X \mid \Theta = \theta \sim Bern(\theta)$

Before proceding with the analysis in terms of Mutual Information, we prove that $\mathcal{A}(n; n; 1, \ldots, 1)$ is not actually optimal in Blackwell-DeGroot sense.

This somehow contrasts with the findings in terms of Mutual Information that seem to emerge in the present work.

We follow strictly the line of reasoning in DeGroot [16], Example 5.2, that uses Completeness of the sufficient statistic in allocation $\mathcal{A}(n; n; 1, \ldots, 1)$ and the notion of non-negatively estimable parametric function, to compare $\mathcal{A}(n; n; 1, \ldots, 1)$ and $\mathcal{A}(n; 1; n)$.
We use some of the general results from classical theory to simplify the analysis.
First of all, it is clear that the univariate marginal distribution of each X_i is still a Bernoulli distribution with parameter $\mu/2$; indeed it clearly takes value in $\mathcal{X} = \{0, 1\}$ and

$$P_\mu(X_i = 1) = E_\mu[X_i] = E_\mu[E_\Theta[X_i]] = E_\mu[\Theta] = \frac{\mu}{2}$$

Consequently, under allocation $\mathcal{A}(n; n; 1, \ldots, 1)$, $X_1, \ldots, X_n$ represent an i.i.d. sample from a Bernoulli($\mu/2$) population with $0 < \mu < 1$, and we can immediately conclude that

- $T_n = \sum_{i=1}^n X_i$ is a sufficient statistic for this allocation;
- $T_n|\mu \overset{(n;n;1,\ldots,1)}{\sim} Binomial\left(n, \frac{\mu}{2}\right)$, which is indeed a complete family for $0 < \mu < 1$

Its distribution is given by

$$h(t; n, \mu) = \binom{n}{t} \left(\frac{\mu}{2}\right)^t \left(1 - \frac{\mu}{2}\right)^{n-t} = \binom{n}{t} \left(E_\mu[\Theta]\right)^t \left(1 - E_\mu[\Theta]\right)^{n-t}$$

for $t = 0, 1, \ldots, n$ and $0 < \mu < 1$.
We note two facts:

- $h(t; n, \mu)$ depends on the distribution of the direct observation, Θ, only through its expected value, $E_\mu[\Theta] = \mu/2$;
- The expected value $E_\mu[\Theta] = \mu/2$ uniquely identifies each member in the family $\{h(t; n, \mu), 0 < \mu < 1\}$: that is, if $E_{\mu_1}[\Theta] \neq E_{\mu_2}[\Theta]$ then clearly $\mu_1 \neq \mu_2$.

As stressed by DeGroot [16] Section 5, the second fact above is a relevant necessary (but not sufficient) condition for $\mathcal{A}(n; n; 1, \ldots, 1)$ to be a valid candidate for optimality: since the distribution of T_n depends on $E_\mu[\Theta]$ only, observation of T_n can 'teach' us about μ

only through $E_\mu[\Theta]$.

Thus if for some $\mu_1 \neq \mu_2$, $E_{\mu_1}[\Theta] = E_{\mu_2}[\Theta]$ but $E_{\mu_1}[\Theta^r] = E_{\mu_2}[\Theta^r]$, for some integer $r > 1$, T_n could not help in distinguishing between the two parameter values, and thus $\mathcal{A}(n; n; 1, \ldots, 1)$ could not be optimal.

The mentioned second fact above clearly rules out this possibility.

The relevance of this argument becomes even more evident when contrasted with the distribution of the sufficient statistic in $\mathcal{A}(n; 1; n)$, which is still T_n.

To arrive at this conclusion, we can rely on the general results on Sufficiency mentioned in the previous section: since $X_1, \ldots, X_n|\Theta$ represent an i.i.d. sample from a Bernoulli population with parameter Θ, $T_n = \sum_{i=1}^n X_i$ is a sufficient statistic for Θ, and it will be sufficient for μ also.

Obviously $T_n|\Theta = \theta \sim Binomial(n, \theta)$, and its distribution under $\mathcal{A}(n; 1; n)$ will be given by

$$g(t; n, \mu) = \binom{n}{t} \int_0^\mu \theta^t (1-\theta)^{n-t} \mu^{-1} d\theta \qquad t = 0, 1, \ldots, n \qquad 0 < \mu < 1$$

We can then observe that $g(t; n, \mu)$ depends on all the moments, up to the n-th order, of the distribution of Θ, $E_\mu[\Theta^r]$ $r = 1, \ldots, n$, not only on $E_\mu[\Theta]$.

So even if, for some $\mu_1 \neq \mu_2$, $E_{\mu_1}[\Theta] = E_{\mu_2}[\Theta]$, there could be an r such that $E_{\mu_1}[\Theta^r] \neq E_{\mu_2}[\Theta^r]$. This implies that T_n in the present allocation seems to be characterized by a larger informative potential than in the case of $\mathcal{A}(n; n; 1, \ldots, 1)$.

But this consideration does not lead to optimality of $\mathcal{A}(n; 1; n)$.

On the contrary, it would be sufficient to rule out $\mathcal{A}(n; n; 1, \ldots, 1)$ if the one-to-one correspondence between μ and $E_\mu[\Theta]$ failed to be verified: indeed, in this case, knowledge of higher order moments of the distribution of Θ would be necessary to distinguish different values of μ. An example of such an instance is considered in DeGroot [16] Section 5, and it will be mentioned in section 2.3.1 of the present work.

We arrived then at the conclusion that T_n is a complete sufficient statistic under $\mathcal{A}(n;n;1,\ldots,1)$; we now proceed using Theorem 2.2.2, that allows to conclude about optimality of $\mathcal{A}(n;n;1,\ldots,1)$ by comparing it with $\mathcal{A}(n;1;n)$.

For allocation $\mathcal{A}(n;n;1,\ldots,1)$ to be sufficient for $\mathcal{A}(n;1;n)$, and consequently optimal in Blackwell-DeGroot sense, for all $t=0,1,\ldots,n$, $g(t;n,\mu)$ should be non-negatively estimable in terms of an observation of the sufficient statistic T_n under $\mathcal{A}(n;n;1,\ldots,1)$. So, for $t=0,1,\ldots,n$, there must exist non-negative constants, $a_s \quad s=0,1,\ldots,n$, possibly depending on n and t, such that [6]

$$g(t;n,\mu)=\sum_{s=0}^{n} a_s h(s;n;\mu) \quad 0<\mu<1$$

Substituting the previously introduced distributions, we obtain the condition

$$\binom{n}{t}\int_0^{\mu}\theta^t(1-\theta)^{n-t}\mu^{-1}d\theta=\sum_{s=0}^{n}a_s\binom{n}{s}\left(\frac{\mu}{2}\right)^s\left(1-\frac{\mu}{2}\right)^{n-s} \quad 0<\mu<1$$

Now, the integral in the left-hand side term is an incomplete Beta function, $B(\mu;t+1,n+1-t)$ that can be expressed as

$$\begin{aligned}\int_0^{\mu}\theta^t(1-\theta)^{n-t}d\theta &= B(\mu;t+1,n+1-t)=B(t+1,n+1-t)I_{\mu}(t+1,n+1-t)\\ &= B(t+1,n+1-t)\sum_{j=t+1}^{n+1}\frac{(n+1)!}{j!(n+1-j)!}\mu^j(1-\mu)^{n+1-j}\\ &= \sum_{j=t+1}^{n+1}\frac{t!(n-t)!}{j!(n+1-j)!}\mu^j(1-\mu)^{n+1-j}\end{aligned}$$

where $I_{\mu}(t+1,n+1-t)$ denotes the regularized (incomplete) Beta function.

Pluggin this expression into the previous condition, after simple manipulations, we obtain

$$\sum_{s=t}^{n}\frac{1}{(s+1)!(n-s)!}\mu^s(1-\mu)^{n-s}=\sum_{s=0}^{n}b_s\mu^s(2-\mu)^{n-s}$$

[6] In the case of discrete spaces, the condition of Sufficiency based on the notion of a non-negatively estimable function modifies to $q_\theta(y)=\sum_{x\in\mathcal{X}}a_x^y p_\theta(x)\ \forall\theta\in\Theta$ and $\forall y\in\mathcal{Y}$ with $a_x^y\geq 0$ and $p_\theta(x)$ and $q_\theta(y)$ denote the pmfs of X and Y respectively.

Using the following facts

$$(1-\mu)^{n-s} = [(2-\mu)-1]^{n-s} = \sum_{h=0}^{n-s} \binom{n-s}{h} (2-\mu)^h (-1)^{n-s-h}$$

$$\mu^s (1-\mu)^h = \sum_{j=s}^{n-h} \binom{n-s-h}{j-s} \frac{1}{2^{n-s-h}} \mu^j (2-\mu)^{n-j}$$

we can write

$$\begin{aligned}
\sum_{s=t}^{n} \frac{1}{(s+1)!(n-s)!} \mu^s (1-\mu)^{n-s} &= \sum_{s=t}^{n} \sum_{h=t}^{n-s} \frac{(-1)^{n-s-h}}{(s+1)!h!(n-s-h)!} \mu^s (2-\mu)^h \\
&= \sum_{s=t}^{n} \sum_{h=t}^{n-s} \frac{(-1)^{n-s-h}}{(s+1)!j!(n-s-h)!} \sum_{j=s}^{n-h} \binom{n-s-h}{j-s} \frac{1}{2^{n-s-h}} \mu^j (2-\mu)^{n-j} \\
&= \sum_{j=t}^{n} \left\{ \sum_{s=t}^{j} \sum_{h=0}^{n-j} \frac{(-1)^{n-s-h} 2^{-(n-s-h)}}{(s+1)!h!(j-s)!(n-j-h)!} \right\} \mu^j (2-\mu)^{n-j} \\
&= \sum_{j=t}^{n} \left\{ \frac{2^{-n}}{j!(n-j)!} \sum_{s=t}^{j} \binom{j}{s} (-1)^{j-s} \frac{2^s}{s+1} \right\} \mu^j (2-\mu)^{n-j} \\
&= \sum_{j=t}^{n} b_j \mu^j (2-\mu)^{n-j}
\end{aligned}$$

which is the required expression.

Being the family of distributions of T_n complete, the found constants are the only ones satisfying the condition: unfortunately they are not all non-negative, as can be easily checked chosing $j = t+1$, for example.

In turn this implies that allocation $\mathcal{A}(n; n, 1, \ldots, 1)$ is not sufficient for $\mathcal{A}(n; 1, n)$, and so it cannot be optimal.

Note that the Classical Theory leads us to rule out allocation $\mathcal{A}(n; n, 1, \ldots, 1)$ as sub-optimal but it does not states that there surely exists an alternative one which will prove to be optimal nor it provides us with an alternative allocation that could perform better than this in most decision problems. We could actually be in a case in which $\mathcal{A}(n; n, 1, \ldots, 1)$ does very good - better than all other allocations - in most problems but fails to excell in some minor and peculiarly structured problems.

We consider now the same hierarchical structure, and analyse it in terms of Mutual Information.

For this model we fix $n = 2$, so the possible allocations are only

- $\mathcal{A}(2; 1; 2)$
- $\mathcal{A}(2; 2; 1, 1)$

We will treat them in turn and then compare them.

To do this we should compute $I(\mu, (X_1)_1^2)$ and $I(\mu, (X_{11}, X_{21}))$ respectively, and these quantities require to obtain posteriors distribution of μ given $(X_1)_1^2$ and (X_{11}, X_{21}).

Then we could compute Mutual Information as the difference between prior and expected posterior entropies for each model, , for example, following the lines of equation 2.6 of Definition 2.1.4 with uncertainty function U_3 of Example 2.1.3.

Unfortunately posterior distributions are not explicitly obtainable in this case, and before using computational tools, we can somehow simply the problem.

First of all, note that for both allocations we can write

$$I(\mu, (X_1)_1^2) = I(\mu, X_{11}) + I(\mu, X_{12}|X_{11}) \tag{2.24}$$

$$I(\mu, (X_{11}, X_{21})) = I(\mu, X_{11}) + I(\mu, X_{21}|X_{11}) \tag{2.25}$$

by the chain rule for Mutual Information.

Since the first terms on the right-hand sides of both equations are equal [7], it seems possible to compare just the second terms: this can be directly interpreted as the **additional amount of information provided by a second observation after a first one has been taken**.

In general it is not possible to compute these expressions directly since they involve

[7] Recall Remark 2.2.4.

distributions not easily (or not at all) obtainable.

However we can use the following identities, again a consequence of the chain rule,

$$\begin{aligned} I((\mu, X_{12}), X_{11}) &= I(\mu, X_{11}) + I(X_{11}, X_{12}|\mu) \\ &= I(X_{11}, X_{12}) + I(\mu, X_{11}|X_{12}) \\ I((\mu, X_{21}), X_{11}) &= I(\mu, X_{11}) + I(X_{11}, X_{21}|\mu) \\ &= I(X_{11}, X_{21}) + I(\mu, X_{11}|X_{21}) \end{aligned} \tag{2.26}$$

Finally, equating the first two lines and the second two lines on the right-hand side of 2.26, noting that by simmetry $I(\mu, X_{11}|X_{12}) = I(\mu, X_{12}|X_{11})$ and $I(\mu, X_{11}|X_{21}) = I(\mu, X_{21}|X_{11})$ and rearranging terms we arrive at

$$\begin{aligned} I(\mu, X_{12}|X_{11}) &= I(\mu, X_{11}) - [I(X_{11}, X_{12}) - I(X_{11}, X_{12}|\mu)] \\ I(\mu, X_{21}|X_{11}) &= I(\mu, X_{11}) - [I(X_{11}, X_{21}) - I(X_{11}, X_{21}|\mu)] \\ &= I(\mu, X_{11}) - I(X_{11}, X_{21}) \end{aligned} \tag{2.27}$$

The last equality follows from observing that, in allocation $A^{(2;2;1,1)}$, $X_{11} \perp X_{21}|\mu$.

Since the first terms in the two equations have the same value, comparison will clearly be based on the terms in square brackets.

We proceed now to evaluate the expressions in square brackets by finding first the joint distributions of the X's conditional on μ and then their marginal joint distributions for both allocations.

For $\mathcal{A}(2;1;2)$, we have

$$P(X_{11} = x_1, X_{12} = x_2) = \begin{cases} 1 - \mu + \frac{\mu^2}{3} & x_1 = x_2 = 0 \\ \frac{\mu}{2} - \frac{\mu^2}{3} & x_1 = 0, x_2 = 1 \quad \text{or} \quad x_1 = 1, x_2 = 0 \\ \frac{\mu^2}{3} & x_1 = x_2 = 1 \end{cases}$$

with equal univariate marginals, $j = 1, 2$

$$P(X_{1j} = x) = \begin{cases} 1 - \frac{\mu}{2} & x = 0 \\ \frac{\mu}{2} & x = 1 \end{cases}$$

so that [8]

$$I(X_{11}, X_{12}|\mu) = 0.0125$$

The marginal joint distribution is instead

$$P(X_{11} = x_1, X_{12} = x_2) = \begin{cases} \frac{22}{36} & x_1 = x_2 = 0 \\ \frac{5}{36} & x_1 = 0, x_2 = 1 \text{or} x_1 = 1, x_2 = 0 \\ \frac{4}{36} & x_1 = x_2 = 1 \end{cases}$$

with univariate marginals

$$P(X_{1j} = x) = \begin{cases} \frac{27}{36} & x = 0 \\ \frac{9}{36} & x = 1 \end{cases}$$

and they lead to the exact evaluation

$$I(X_{11}, X_{12}) = 0.0312$$

For $\mathcal{A}(2; 2; 1, 1)$, we need to compute the overall marginal distribution only, since as mentioned $I(X_{11}, X_{21}|\mu)$ vanishes, and obviuosly the univariate marginals are equal to those already obtained for $\mathcal{A}(2; 1; 2)$.

We have

$$P(X_{11} = x_1, X_{12} = x_2) = \begin{cases} \frac{21}{36} & x_1 = x_2 = 0 \\ \frac{6}{36} & x_1 = 0, x_2 = 1 \text{or} x_1 = 1, x_2 = 0 \\ \frac{3}{36} & x_1 = x_2 = 1 \end{cases}$$

which leads to

$$I(X_{11}, X_{21}) = 0.0059$$

[8] This evaluation is the only one that required a simple Monte Carlo step: we generated i.i.d. variates from the distribution of μ ($\mathcal{U}(0, 1)$) and computed the Mutual Information between the observations, conditional on each simulated value: then we averaged the obtained outcomes. The simulation step is necessary since direct integration of the expression for Mutual Information conditional on a particular value of μ with respect to the (any) prior on μ cannot be carried out explicitly.

The final conclusion is then that allocation $\mathcal{A}(2;2;1,1)$ is more informative than allocation $\mathcal{A}(2;1;2)$ since

$$I(X_{11}, X_{12}) - I(X_{11}, X_{12}|\mu) = 0.0312 - 0.0125 = 0.0187 \quad (2.28)$$

$$I(X_{11}, X_{21}) - I(X_{11}, X_{21}|\mu) = 0.0059 \quad (2.29)$$

or, by identities 2.27, equivalently

$$I(\mu, X_{21}|X_{11}) > I(\mu, X_{12}|X_{11}) \quad (2.30)$$

Some observations seem to be suggested by this example:

- The terms in square brackets in equation 2.27 we just compared are Mutual Information amounts between observations in each allocation, and not between parameter and observations: the focus seems to have shifted from information provided by observations about parameters to **information provided by one observation about the other**, or in other words to **dependence between observations** (as measured by Mutual Information).

- They have a straightforward interpretation: they measure the **dependence** between observations **induced by μ only**, that is when dependence due to the 'middle level' (Θ's) is accounted for. Indeed

  - $I(X_{11}, X_{12})$ and $I(X_{11}, X_{21})$ measure the global strength of the dependence in each allocation, due to both μ and the Θ's;

  - $I(X_{11}, X_{12}|\mu)$ and $I(X_{11}, X_{21}|\mu)$ measure the average dependence due only to the Θ's, since μ is held fixed.

  Equality between the last two lines of 2.27 agrees with this reasoning: in allocation $\mathcal{A}(2;2;1,1)$, all dependence between X_{11} and X_{21} is indeed induced by μ.

- The larger this dependence between observations the smaller the additional amount of information provided by a second one: most of its informative content is already carried by the first observation.

Note that in the classical analysis of the present problem, we drew the conclusion that $\mathcal{A}(n; n; 1, \ldots, 1)$ is not optimal, for any chosen n - thus also for $n = 2$ as assumed in the Bayesian approach: nonetheless direct computation of Mutual Information has shown that it is indeed optimal according to the Mutual Information criterion. There must be some other problem for which it is not the best experiment to choose.
This finding is clearly only a suggestion, since optimality has been defined with respect to a small class of allocations (one with just two elements).
On the opposite side, the larger value of $I(\mu, (X_{11}, X_{21}))$ with respect to $I(\mu, (X_{11}, X_{12}))$ basically provides an example of a decision problem in which $\mathcal{A}(2; 2; 1, 1)$ outperforms $\mathcal{A}(2; 1; 2)$: thus $\mathcal{A}(2; 1; 2)$ cannot be optimal in Blackwell-DeGroot sense.
We are then lead to the conclusion that for the present problem with $n = 2$ no optimal allocation exists! And considering that we were comparing experiments with just two observations the finding is moderately annoying: since the complexity of the problem is quite small, one would hope more intensely for a clear-cut solution than in more complex problems where 'more things can go wrong'.
This reasoning seems also to suggest that the classical criterion to measure the 'Informative content' may be too strict to be applied in normal routine.

Hypothetically removing the constraint on the sample size, we could suppose that the basic remarks about the inverse relation between dependence among observations and the consequent 'Informative content' still remain valid: the more dependent are the observations in a hierarchical structure the less informative (about the parameter of interest) they prove to be in terms in Mutual Information.

A natural conclusion would then be that the optimal allocation is the one with observations less dependent as possible, pointing towards $\mathcal{A}(n; n; 1, \ldots, 1)$.

Clearly, this line of reasoning needs formal arguments to be justified or, at least, significant evidence in its support.

2.2.4 Normal hierarchical model

The normal model is one of the most widely studied and applied in the class of hierarchical ones, and actually the one for which clear results can be obtained.

We consider hierarchical structures composed by 3 levels, the lowest being defined by the normal likelihood for the observations and the second and third ones consisting of normal priors on the mean parameters of the likelihood and of the second-level normal prior respectively. Variances are assumed known.

The class of models we consider can be identified with the class

$$C_n = \left\{ (n; k; n_1, \ldots, n_k) : k \in \{1, \ldots, n\}, n_j \in \{1, \ldots, n-k+1\}, j = 1, \ldots, k, \sum_{j=1}^{k} n_j = n \right\} \tag{2.31}$$

The reason to set the range of the n_j's as $\{1, \ldots, n-k+1\}$ is that we identify models $(n; k, n_1, \ldots, n_k)$ and $(n; k'; n'_1, \ldots, n'_k)$ if $\{n_j : j \in \{1, \ldots, k\}, n_j > 0\} = \{n'_j : j \in \{1, \ldots, k'\}, n'_j > 0\}$.

To be explicit, we identify allocations with the same number of units with at least one observation, and the same number of observations on each of these units. So for a fixed k, the free observations to be distributed among the k units are only $n-k$, and on each unit at most $n-k+1$ observation can be taken.

In the light of the Information measures introduced in Chapter 1, we propose to judge the relative merits of these two models according to the quantity of Information they are able to provide about the parameter μ.

In particular, we will base our evaluations on the Mutual Information $I(\mu, \underline{X}^n)$, that is on

the Information a sample of size n is able to provide about the parameter of higher level: clearly in this notation something is missing, and it is the reference to the hierarchical structure which gave rise to the observations.

To be consistent but slightly simplify the notation, we will then write $I_{(n;1;n)}(\mu, \underline{X}^n)$ and $I_{(n;n;1,\ldots,1)}(\mu, \underline{X}^n)$, for example.

Independently of the hierarchical structure considered, by expression 1.28 of Chapter 1 for Mutual Information, we can compute the exact values of these two quantities as

$$I(\mu, \underline{X}^n) = h(\underline{X}^n) - h(\underline{X}^n|\mu) = h(\mu) - h(\mu|\underline{X}^n) \tag{2.32}$$

Actually the second expression above is the one that proves to be more useful, and it will be applied after some considerations that help simplify the computations.

Note that, in both cases, to apply the preceding expression we need to find out the posterior distribution of μ given the observations $\underline{X}^n$: in general the second level of the hierarchy leads one out of the usual model characterized by i.i.d. observations since it induces dependence between the observables or it leads to marginal distributions for the X_i given μ which are not analytically tractable.

Nevertheless, for the models considered below Sufficiency and the properties of the Normal distribution help simplify the problem.

To begin with, we consider the two simple models, $(n; 1; n)$ and $(n; n; 1, \ldots, 1)$.

We assume known variances, that is σ^2, τ_Θ^2 and τ_μ^2 are known positive constants in the following.

Actually these two models are the only ones that need to be considered: DeGroot [16] indeed shows how in this problem it is possible to construct a stochastic transformation between the two experiments, and show the sufficiency of $(n; n; 1, \ldots, 1)$ for $(n; 1; n)$, which in turn implies optimality of the former.

We postpone the discussion until more details are available and DeGroot's argument can

be made evident.

In the first case, we have

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $\Theta|\mu \sim \mathcal{N}(\mu, \tau_\Theta^2)$
3. $X_1, \ldots, X_n|\Theta \overset{i.i.d.}{\sim} \mathcal{N}(\Theta, \sigma^2)$

In model $(n; 1; n)$, reasoning conditionally on Θ, we can note that $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the sample average is sufficient for Θ and, according to the results mentioned in the general section about classical comparison of allocations, also for μ.
Clearly $\overline{X}_n|\Theta \sim \mathcal{N}(\Theta, \frac{\sigma^2}{n})$. Furthermore marginalizing out Θ we arrive at the model

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $\overline{X}_n|\mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \tau_\Theta^2 + \frac{\sigma^2}{n})$

which is the standard Normal model with conjugate prior and posterior

$$\mu|\underline{X}^n \overset{(n;1;n)}{\sim} \mu|\overline{X}_n \overset{(n;1;n)}{\sim} \mathcal{N}\left(\frac{\tau_\mu^2}{\tau_\mu^2 + \tau_\Theta^2 + \sigma^2/n}\overline{X}_n, \tau_\mu^2\frac{\tau_\Theta^2 + \sigma^2/n}{\tau_\mu^2 + \tau_\Theta^2 + \sigma^2/n}\right) \tag{2.33}$$

where the notation $\overset{(n;1;n)}{\sim}$ indicates under which allocation is computed the posterior distribution.
Now, from the expression of the differential entropy for Normal r.v., the value of Mutual Information for model $(n; 1; n)$ can straightforwardly be found. Indeed

$$\begin{aligned} I_{(n;1;n)}(\mu, \underline{X}^n) &= h(\mu) - h_{(n;1;n)}(\mu|\underline{X}^n) \\ &= \frac{1}{2}\log\left(2\pi e\tau_\mu^2\right) - \frac{1}{2}\log\left(2\pi e\tau_\mu^2\frac{\tau_\Theta^2 + \sigma^2/n}{\tau_\mu^2 + \tau_\Theta^2 + \sigma^2/n}\right) \\ &= \frac{1}{2}\log\left(\frac{\tau_\mu^2 + \tau_\Theta^2 + \sigma^2/n}{\tau_\Theta^2 + \sigma^2/n}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{n\tau_\mu^2}{n\tau_\Theta^2 + \sigma^2}\right) = -\frac{1}{2}\log\left(1 - \rho_{(n;1;n)}^2\right) \end{aligned} \tag{2.34}$$

where $\rho^2_{(n;1;n)}$ is the squared correlation coefficient between μ and $\overline{X}_n$ in the present allocation.

If now we focus on model $(n; n; 1, \ldots, 1)$, defined by

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $\Theta_1, \ldots, \Theta_n | \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \tau_\Theta^2)$
3. $X_i | \Theta_i \overset{ind.}{\sim} \mathcal{N}(\Theta_i, \sigma^2)$ for $i = 1, \ldots, n$

we can, first of all, marginalize out $\Theta_1, \ldots, \Theta_n$, as we are interested in inferences about μ only. By known results, we then arrive at the model

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $X_i | \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2 + \tau_\Theta^2)$ for $i = 1, \ldots, n$

which reconducts us to the well known case of a random sample from a normal distribution with known variance and conjugate prior on the unknown mean parameter.
A classical analysis directly recognizes again $\overline{X}_n$ as a sufficient statistic for μ, and distributed as $\overline{X}_n | \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \frac{\sigma^2 + \tau_\Theta^2}{n})$.

The difference in the marginal distribution of the X_i's is the inflated variance, which exemplifies quite clearly the effect of the availability of indirect observation only.

Standard computations lead to the conclusion that the posterior distribution in this model is

$$\begin{aligned} \mu | \underline{X}^n &\overset{(n;n;1,\ldots,1)}{\sim} \mu | \overline{X}_n \\ &\overset{(n;n;1,\ldots,1)}{\sim} \mathcal{N}\left(\frac{\tau_\mu^2}{\tau_\mu^2 + (\tau_\Theta^2 + \sigma^2)/n} \overline{X}_n, \tau_\mu^2 \frac{(\tau_\Theta^2 + \sigma^2)/n}{\tau_\mu^2 + (\tau_\Theta^2 + \sigma^2)/n} \right) \end{aligned} \tag{2.35}$$

so that we can compute the Mutual Information, according to the same expression used for the preceding model.

$$\begin{aligned} I_{(n;n;1,\ldots,1)}(\mu, \underline{X}^n) &= h(\mu) - h_{(n;n;1,\ldots,1)}(\mu|\underline{X}^n) \\ &= \frac{1}{2}\log\left(2\pi e\tau_\mu^2\right) - \frac{1}{2}\log\left(2\pi e\tau_\mu^2 \frac{(\tau_\Theta^2+\sigma^2)/n}{\tau_\mu^2 + (\tau_\Theta^2+\sigma^2)/n}\right) \\ &= \frac{1}{2}\log\left(\frac{\tau_\mu^2 + (\tau_\Theta^2+\sigma^2)/n}{(\tau_\Theta^2+\sigma^2)/n}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{n\tau_\mu^2}{\tau_\Theta^2+\sigma^2}\right) = -\frac{1}{2}\log\left(1-\rho^2_{(n;n;1,\ldots,1)}\right) \end{aligned} \tag{2.36}$$

and $\rho^2_{(n;n;1,\ldots,1)}$ has the same meaning as above but with respect to $(n;n;1,\ldots,1)$.
By monotonicity of the logarithm function, a direct comparison of the last lines of equations 2.34 and 2.36 clearly shows that the different information contents of the two models are captured by the ratios defining the arguments of the logarithms

$$r_{(n;1;n)} = \frac{n\tau_\mu^2}{n\tau_\Theta^2+\sigma^2}; \qquad r_{(n;n;1,\ldots,1)} = \frac{n\tau_\mu^2}{\tau_\Theta^2+\sigma^2} \tag{2.37}$$

respectively.

Obviously the two ratios are equal for $n = 1$ while $r_{(n;n;1,\ldots,1)} > r_{(n;1;n)}$ for $n \geq 2$, and consequently

$$I_{(n;n;1,\ldots,1)}(\mu, \underline{X}^n) \geq I_{(n;1;n)}(\mu, \underline{X}^n) \quad n \geq 1 \tag{2.38}$$

with strict inequality for $n \geq 2$ for every strictly positive values of the variances τ_Θ^2 and τ_μ^2.
A partial conclusion can thus be drawn about this hierarchical normal model: the experiment with each observation taken on a different unit (or individual) provides more information than the "opposite" one taking all the observations on a single unit.

We can now formalize the argument in DeGroot [16] that actually prove sufficiency of $\mathcal{A}(n;n;1,\ldots,1)$ for $\mathcal{A}(n;1;n)$, directly constructing a stochastic transformation between

the reduced experiments represented by the sufficient statistics.

More precisely, the sample mean $\overline{X}_n$ is the sufficient statistic in both allocations, with Normal distribution centered on μ but with different variances:

$$Var_\mu[\overline{X}_n] \overset{(n;1;n)}{=} Var_\mu[\overline{X}'_n] = \tau_\Theta + \frac{\sigma^2}{n} \quad \text{and} \quad Var_\mu[\overline{X}_n] \overset{(n;n;1,\ldots,1)}{=} \frac{\sigma^2 + \tau_\Theta}{n}$$

[9] respectively.

To apply Lemma 2.2.1, we should find a random variable Z and a function ψ such that $\psi(\overline{X}_n, Z) =_d \overline{X}'_n$. But it suffices to choose $Z \sim \mathcal{N}(0, \tau_\Theta^2 \frac{n-1}{n})$ and consider the function $\psi(x, z) = x + z$, to immediately find $\overline{X}_n + Z =_d \overline{X}'_n$.

By this simple conclusion and Theorem 2.2.2, it also follows that $\mathcal{A}(n; n; 1, \ldots, 1)$ is actually optimal in Blackwell-DeGroot sense.

The classical result, mentioned above, clearly implies that $\mathcal{A}(n; n; 1, \ldots, 1)$ is optimal in terms of Mutual Information too. Furthermore for normal random vectors stronger Sufficiency results have been proven for dependent r.v.'s: see Shaked and Tong [57], [58], and Stepniak [63].

Nonetheless we proceed to show how the result on Mutual Information can also be obtained in a different way.

After having proved to be effective when dealing with the current problem (for which a definite result already exists), this different approach will then be exploited in section 2.2.5, facing a different model for which classical results seem to be absent.

$$I_{(n;n;1,\ldots,1)}(\mu, \underline{X}^n) \geq I_{(n;k;n_1,\ldots,n_k)}(\mu, \underline{X}^n) \tag{2.39}$$

for every $n \geq 1$ and for all $k = 1, \ldots, n-1$, $n_i \in \aleph$ with $\sum_{i=1}^k n_i = n$.

[9] We slightly changed the notation to distinguish the same sufficient statistic in the two different allocations and avoid confusion.

For a fixed k, consider the model $(n; k; n_1, \ldots, n_k)$, which by Sufficiency considerations, can be reduced to

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $\Theta_1, \ldots, \Theta_k | \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \tau_\Theta^2)$
3. $\overline{X}_i | \Theta_i \overset{ind.}{\sim} \mathcal{N}(\Theta_i, \sigma^2/n_i)$ for $i = 1, \ldots, k$

with $\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ and X_{ij} for $j = 1, \ldots, n_i$ are the observations belonging to the i-th group. Marginalizing out $\Theta_1, \ldots, \Theta_k$ we are again lead to the model

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $\overline{X}_i | \mu \overset{ind.}{\sim} \mathcal{N}(\mu, \tau_\Theta^2 + \sigma^2/n_i)$ for $i = 1, \ldots, k$

The posterior distribution of μ is

$$\mu | \underline{X}^n \overset{(n;k;n_1,\ldots,n_k)}{\sim} \mathcal{N}\left(\sum_{i=1}^k w_i \overline{X}_i, \left(\frac{1}{\tau_\mu^2} + \sum_{i=1}^k \frac{1}{\tau_\Theta^2 + \sigma^2/n_i} \right)^{-1} \right) \tag{2.40}$$

To arrive at this conclusion, for ease of notation define $\sigma_i^2 = \tau_\Theta^2 + \sigma^2/n_i$ and $Y_i = \overline{X}_i$ so that we are reconduced to the model where we have indipendent normal observations $Y_1, \ldots, Y_k$ with same mean μ but different variances σ_i^2. Clearly we have

$$\begin{aligned} \pi(\mu | y_1, \ldots, y_k) &\propto \pi(\mu) \prod_{i=1}^k f(y_i | \mu) \\ &\propto \mathrm{e}^{-\frac{1}{2\tau_\mu^2}\mu^2} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^k \frac{(y_i-\mu)^2}{\sigma_i^2}} \\ &\propto \mathrm{e}^{-\frac{1}{2}\left(\frac{1}{\tau_\mu^2} + \sum_{i=1}^k \frac{1}{\sigma_i^2}\right)\mu^2 + \mu \sum_{i=1}^k \frac{1}{\sigma_i^2} y_i} \\ &\propto \mathrm{e}^{-\frac{1}{2}\left(\frac{1}{\tau_\mu^2} + \sum_{i=1}^k \frac{1}{\sigma_i^2}\right)\left(\mu - \sum_{i=1}^k \frac{1/\sigma_i^2}{1/\tau_\mu^2 + \sum_{j=1}^k 1/\sigma_j^2} y_i\right)^2} \end{aligned} \tag{2.41}$$

Using this result, we can recognize the weights

$$w_i = \frac{1/\sigma_i^2}{1/\tau_\mu^2 + \sum_{j=1}^k 1/\sigma_j^2} = \frac{1/(\tau_\Theta^2 + \sigma^2/n_i)}{1/\tau_\mu^2 + \sum_{j=1}^k 1/(\tau_\Theta^2 + \sigma^2/n_j)} \tag{2.42}$$

and the posterior variance

$$\begin{aligned} V(\mu|\underline{X}^n) &= \left(\frac{1}{\tau_\mu^2} + \sum_{i=1}^{k} \frac{1}{\sigma_i^2}\right)^{-1} = \left(\frac{1}{\tau_\mu^2} + \sum_{i=1}^{k} \frac{1}{\tau_\Theta^2 + \sigma^2/n_i}\right)^{-1} \\ &= \tau_\mu^2 \left(1 + \tau_\mu^2 \sum_{i=1}^{k} \frac{1}{\tau_\Theta^2 + \sigma^2/n_i}\right)^{-1} \end{aligned} \tag{2.43}$$

which is the relevant quantity in the present problem.

Note that, obviously, choosing $k = 1$ ($n_1 = n$) or $k = n$ ($n_1 = \cdots = n_i = \cdots = n_n = 1$) we arrive at the expressions for the posterior variances in 2.33 and 2.35 respectively, e.g. to the posterior variances for the "extreme" models.

A more or less explicit expression for the Mutual Information of each model in the class $\mathcal{C}_n$ can thus be given:

$$\begin{aligned} & I_{(n;k;n_1,\ldots,n_k)}(\mu, \underline{X}^n) \\ &\quad = h(\mu) - h(\mu|\underline{X}^n) \\ &\quad = \frac{1}{2}\log\left(2\pi e \tau_\mu^2\right) - \frac{1}{2}\log\left(2\pi e \tau_\mu^2 \left(1 + \tau_\mu^2 \sum_{i=1}^{k} \frac{1}{\tau_\Theta^2 + \sigma^2/n_i}\right)^{-1}\right) \\ &\quad = \frac{1}{2}\log\left(1 + \tau_\mu^2 \sum_{i=1}^{k} \frac{1}{\tau_\Theta^2 + \sigma^2/n_i}\right) \end{aligned} \tag{2.44}$$

The value of Mutual Information for a model is clearly directly related to the value of the summation in Equation 2.44.

The proof of 2.39 will be structured in two steps: first we will show that, for a fixed k, the summation is maximixed at the allocation(s) which distribute observations as evenly as possible - considered the discrete nature of the variables - among units. Secondly we will show that the maximum for each k is increasing in k itself.

As already noted, the Mutual Information 2.44 is increasing in the summation in the argument of the logarithmic function, so that maximizing the latter is equivalent to maximizing the former.

We can express the summation as

$$\begin{aligned}\sum_{i=1}^{k}\frac{1}{\tau_\Theta^2+\sigma^2/n_i} &= \sum_{i=1}^{k}\frac{n_i}{n_i\tau_\Theta^2+\sigma^2} = \frac{1}{\tau_\Theta^2}\sum_{i=1}^{k}\frac{n_i}{n_i+\sigma^2/\tau_\Theta^2} \\ &= \frac{1}{\tau_\Theta^2}\sum_{i=1}^{k}\left\{1-\frac{c}{n_i+c}\right\} = \frac{1}{\tau_\Theta^2}\left\{k-c\sum_{i=1}^{k}\frac{1}{n_i+c}\right\} \\ &= \frac{1}{\tau_\Theta^2}\left\{k-c\sum_{i=1}^{k}h_i(\underline{n})\right\} = \frac{h(\underline{n})}{\tau_\Theta^2}\end{aligned} \tag{2.45}$$

where we set $c=\sigma^2/\tau_\Theta^2$, $\underline{n}=(n_1,\ldots,n_k)$ and $h_i(\underline{n})=\frac{1}{n_i+c}$.

We are lead to the problem of finding extrema of

$$\mathcal{L}(\underline{n},\lambda)=h(\underline{n})+\lambda\left\{\sum_{i=1}^{k}n_i-n\right\} \tag{2.46}$$

where, for the moment, it is useful to let each n_i vary in $\Re^+$.

Note that $h(\underline{n})$ is a concave function of $\underline{n}$; indeed it is easy to check that each $h_i(\underline{n})$ is a convex function of $\underline{n}$ so that their sum is still a convex function, and the sign change leads to concavity. This property grants that the extrema found are really maxima.

The first order conditions are given by

$$\begin{aligned}\mathcal{L}'_i(\underline{n},\lambda) &= \frac{\partial}{\partial n_i}\mathcal{L}(\underline{n},\lambda) = \frac{c}{(n_i+c)^2}+\lambda=0 \quad i=1,\ldots,k \\ \mathcal{L}'_\lambda(\underline{n},\lambda) &= \frac{\partial}{\partial\lambda}\mathcal{L}(\underline{n},\lambda) = \sum_{i=1}^{k}n_i-n=0\end{aligned} \tag{2.47}$$

In particular each of the first k conditions can be reexpressed as

$$n_i=\sqrt{-c\lambda^{-1}}-c=n^*=constant \tag{2.48}$$

so that combined with the last condition lead to $n^*=\frac{n}{k}$ (and $\lambda=-c[(n/k)+c]^{-2}$).

Obviously only in fortunate cases n will be a multiple of k, so that, by invariance of the function $h(\underline{n})$ with respect to permutations of the elements of $\underline{n}$, with integer-valued n_i the (typically multiple) maxima will be attained with those allocations which place $\lfloor n/k\rfloor$ observations on each unit and put the remaining $n-k\lfloor n/k\rfloor$ each on a different unit.

There are $\binom{k}{n - k\lfloor n/k \rfloor}$ such allocations.

Consider now the case $k = n$. The previous analysis shows that $n_i = 1$ is the optimal (and, with integer-valued n_i's, indeed unique) allocation.

Formally, each maximization problem with $k < n$, is equivalent to the one with $k = n$ when additional constraints are imposed: in particular we impose $n_i = 0$ for $i = k + 1, \ldots, n$.

Since a constrained maximum is always inferior to or at most equal to an uncontrained one, we are lead to conclude that the maximum value of Mutual Information is attained by the allocation $(n; n; 1, \ldots, 1)$.

2.2.5 Normal hierarchical model with different unit variances

We will now remove the assumption of equal variance for all units.

At the beginning it seems reasonable to modify slightly the problem: instead of assuming a variable number of units, we fix their number at ν and look for the best allocation of the usual n observations to these units.

In a real setting, when sample size is larger than the number of units the above reformulation of the problem seems more reasonable. Instead, in the case in which n is relatively small it seems possible to assume availability of as many units as observations: then we can just set $\nu = n$.

Without loss of generality we can assume that

$$\sigma_1^2 < \cdots < \sigma_i^2 < \ldots \sigma_\nu^2 \tag{2.49}$$

Analogously to the the preceding section, by sufficiency, allocation $\mathcal{A}(n; \nu; n_1, \ldots, n_{\nu)}$ is represented as

1. $\mu \sim \mathcal{N}(0, \tau_\mu^2)$
2. $\Theta_1, \ldots, \Theta_\nu | \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \tau_\Theta^2)$

3. $\overline{X}_i|\Theta_i \overset{ind.}{\sim} \mathcal{N}(\Theta_i, \sigma_i^2/n_i)$ for $i = 1, \ldots, \nu$

By steps analogous to the preceding section, the posterior distribution of μ given $\underline{X}^n$ is shown to be

$$\mu|\underline{X}^n \overset{(n;\nu;n_1,\ldots,n_\nu)}{\sim} \mathcal{N}\left(\sum_{i=1}^{\nu} w_i \overline{X}_i, \left(\frac{1}{\tau_\mu^2} + \sum_{i=1}^{\nu} \frac{1}{\tau_\Theta^2 + \sigma_i^2/n_i}\right)^{-1}\right) \tag{2.50}$$

to which corresponds a value of Mutual Information of

$$I_{(n;\nu;n_1,\ldots,n_\nu)}(\mu, \underline{X}^n) = \frac{1}{2}\log\left(1 + \tau_\mu^2 \sum_{i=1}^{\nu} \frac{1}{\tau_\Theta^2 + \sigma_i^2/n_i}\right) \tag{2.51}$$

The function comparing as summation in the argument of the logarithm can be expressed as

$$\sum_{i=1}^{\nu} \frac{1}{\tau_\Theta^2 + \sigma_i^2/n_i} = \tau_\Theta^{-2}\left\{\nu - \sum_{i=1}^{\nu} \frac{c_i}{c_i + n_i}\right\} \tag{2.52}$$

with $c_i = \frac{\sigma_i^2}{\tau_\Theta^2}$ and clearly is still concave, but no more symmetric in the n_i's.

Ignoring for the moment that solutions must be non-negative integers, and looking for real solutions, the first order conditions analogous to 2.47 are given by

$$\begin{aligned} \mathcal{L}_i'(\underline{n}, \lambda) &= \frac{\partial}{\partial n_i}\mathcal{L}(\underline{n}, \lambda) = \frac{c_i}{(n_i + c_i)^2} + \lambda = 0 \quad i = 1, \ldots, k \\ \mathcal{L}_\lambda'(\underline{n}, \lambda) &= \frac{\partial}{\partial \lambda}\mathcal{L}(\underline{n}, \lambda) = \sum_{i=1}^{k} n_i - n = 0 \end{aligned} \tag{2.53}$$

leading to

$$n_i^* = \sqrt{\frac{c_i}{\lambda^*}} - c_i = \sqrt{\frac{\sigma_i^2}{\lambda^* \tau_\Theta^2}} - \frac{\sigma_i^2}{\tau_\Theta^2} \quad i = 1, \ldots, \nu \tag{2.54}$$

and λ^* is determined in such a way that $\sum_{i=1}^{\nu} n_i^* = n$.

Note that $g(\lambda) = \sum_{i=1}^{\nu} n_i(\lambda)$ with $n_i(\lambda) \equiv \sqrt{\frac{c_i}{\lambda}} - c_i$ is monotonically decreasing as a function of λ so the optimal value of λ could in principle be determined by progressively raising λ starting from 0 until $g(\lambda) = n$.

However, since we have not explicitly taken into account non-negativity of the n_i's, the optimal value found could lead to some negative allocations and some others larger than

n.

We redefine the solutions as

$$n_i^* = max\left\{0, \sqrt{\frac{c_i}{\lambda^*}} - c_i\right\} \quad i = 1, \ldots, \nu \tag{2.55}$$

and λ^* such that $\sum_{i=1}^{\nu} n_i^* = n$.

$$\tilde{g}(\lambda) = \sum_{i=1}^{\nu} max\left\{0, \sqrt{\frac{c_i}{\lambda}} - c_i\right\} \tag{2.56}$$

is still a monotone decreasing function of λ so setting $\tilde{g}(\lambda) = n$ leads to the optimal value, λ^* and via 2.55 to the n_i^* - now all non negative (real) and summing to n. An approximation is then called for to arrive at integer-valued allocations.

Some considerations about the nature of the **allocation rule** just obtained can be done:

- For each $i = 1, \ldots, \nu$, we can define a value λ_i such that $n_i > 0$ if and only if $\lambda < \lambda_i$:

$$n_i = \sqrt{\frac{c_i}{\lambda}} - c_i > 0 \Leftrightarrow \lambda < \frac{\tau_\Theta^2}{\sigma_i^2} = \lambda_i \tag{2.57}$$

- By the ordering of the variances we can conclude the reversed ordering

$$\lambda_\nu < \cdots < \lambda_i < \cdots < \lambda_1 \tag{2.58}$$

- If λ^* is larger than some of the λ_i's, the corresponding n_i's will be zero: **units with larger variances are the first to be excluded from sampling**.

- We can write $\lambda_i = \frac{\tau_\Theta^2}{\sigma_i^2} = \frac{1/\sigma_i^2}{1/\tau_\Theta^2}$, so that λ_i can be interpreted as a **relative precision** of indirect observations from group i with respect to direct observation: if this relative precision is too low, sampling on that unit is not activated.

- Consequently λ^* can be viewed as a **required minimum precision**

- For those units that are sampled - $\lambda_i > \lambda^*$, the allocations are fixed by

$$\begin{aligned} n_i^* &= \frac{\sigma_i}{\tau_\Theta}\left(\frac{1}{\lambda^{*1/2}} - \frac{\sigma_i}{\tau_\Theta}\right) = \frac{\sigma_i}{\tau_\Theta}\left(\frac{\sigma^*}{\tau_\Theta} - \frac{\sigma_i}{\tau_\Theta}\right) \\ &= \tau_\Theta^{-2}\sigma_i(\sigma^* - \sigma_i) = \tau_\Theta^{-2}K(\sigma_i) \end{aligned} \tag{2.59}$$

  for $\sigma_i \in (0, \sigma^*)$ with $\sigma^* \equiv \frac{\tau_\Theta}{\lambda^{*1/2}}$, 'maximum admitted standard deviation'.

  By a simple analysis of the function $K(\sigma) = \sigma(\sigma^* - \sigma)$ it can be concluded that units for which σ_i **is closer to** $\sigma^*/2$ receive more observations.

2.3 Simulations

The attempt to extend the analysis of the previous sections to models with different distributions at the various levels of the hierarchy seems to lead almost inevitably to analytical difficulties:

- mostly due to the iterated mixing, posterior **distributions for** μ and **overall marginals** ('overall' meaning with respect to both Θ's and μ) are hardly ever obtainable in explicit form, making impossible the computation of the log-ratio's

$$\log\frac{\pi(\mu|\mathbf{x})}{\pi(\mu)} \qquad \log\frac{f(\mathbf{x}|\mu)}{f(\mathbf{x})} \tag{2.60}$$

  in the double integral defining Mutual Information.

- In the Normal case, posterior entropy is not a function of $\mathbf{x}$ values, making unnecessary integration with respect to the distribution of the observables.

  In general, however, posterior entropy is indeed a function of the specific sample values, and, even if the posterior distribution is available explicitly, **integration** with respect to $\mathbf{X}$ and μ **of the log posterior-to-prior ratio** (or of the log μ-conditional-to-marginal likelihood ratio) becomes easily untractable.

We have then tried to proceed via computational methods.

In particular the following analysis represents an attempt to draw some conclusions about

a more realistic version of the simple model with Bernoulli observables in section 2.2.3. The fact that observations are still unobserved makes the use of relatively simple Monte Carlo methods directly available: a short description of the algorithm used for simulations is presented in the appendix, together with some general considerations.

2.3.1 A Gamma-Beta prior for Bernoulli Observables

The following analysis was inspired by Parmigiani and Berry [47] who studied utility surfaces and posterior distributions as functions of the chosen design, with a fixed prior.
As it will be made explicit in the following, we use a slightly different approach by examining few designs but letting the hyperparameters of the prior distribution change: in this way it was thought that different priors could lead to different optimal designs, and a 'prior effect' could emerge.
It seemed indeed reasonable to suppose that different prior opinions could be most sensible to different designs.
Results do not seem to confirm this supposition.

The class of models considered here, for fixed k and $\mathbf{n} = (n_1, \ldots, n_k)$, is described by

1. - $\alpha \sim Gamma(\gamma_1, \eta)$
   - $\beta \sim Gamma(\gamma_2, \eta)$, $\alpha \perp \beta$
2. $\Theta_1, \ldots, \Theta_k | \ \alpha, \beta \overset{i.i.d.}{\sim} Beta(\alpha, \beta)$
3. $X^{(i)} = (X_{i1}, \ldots, X_{in_i}) | \Theta_i \overset{i.i.d.}{\sim} Bernoulli(\Theta_i)$

with γ_1, γ_2 and η known positive constants.
A reduction by Sufficiency is clearly reasonable also in this case but it does not help the analytical work: instead it has been used in the design of the algorithm.

Before proceding with the exposition of the computational results, we briefly recall the arguments of DeGroot [16], already mentioned in section 2.2.3, that exclude the possibility for $\mathcal{A}(n; n; 1, \ldots, 1)$ to be optimal (in the classical sense).

The classical approach obviously takes into consideration only the component distributions at points 2. and 3. of the previous list, ignoring the prior distribution on μ.

Exactly the same argument used in section 2.2.3, leads to the conclusion that the statistic $T_n = \sum_{i=1}^{n} X_i$ is sufficient for μ in both extremal allocations, $\mathcal{A}(n; n; 1, \ldots, 1)$ and $\mathcal{A}(n; 1; n)$.

For a generic allocation $\mathcal{A}(n; k; n_1, \ldots, n_k)$, the vector-valued sufficient statistic will be given by

$$(T_{n_1}, \ldots, T_{n_k}) = \left(\sum_{i=1}^{n_1} X_{1i}, \ldots, \sum_{i=1}^{n_k} X_{ki} \right)$$

by the results about sufficient statistics in allocation problems introduced in section 2.2.2.

The distribution of each component of the vector will be given by

$$g(t_k; n_k, \alpha, \beta) = \binom{n_k}{t_k} \int_{[0,1]} \theta^{t_k} (1-\theta)^{n_k - t_k} \pi_\Theta(\theta|\alpha, \beta) d\theta \quad t_k = 0, 1, \ldots, n_k$$

and $\pi_\Theta(\theta|\alpha, \beta)$ is the Beta density assumed at point 2, which is clearly a Beta-Binomial distribution.

The joint distribution of the components of the sufficient statistic will be the product of these component distribution, considering their independence conditioning on μ.

The relevant point here is that this joint distribution will depend on the first $\tilde{n} = \max\{n_1, \ldots, n_k\}$ moments of the Beta distribution, $E_{\alpha,\beta}[\Theta^r]$ with $r = 1, \ldots, \tilde{n}$.

So, as particular cases, we encounter the general considerations of section 2.2.3: T_n has a distribution depending on the first n moments under $\mathcal{A}(n; 1; n)$, while its distribution will depend on the expected value $E_{\alpha,\beta}[\Theta]$ only, under $\mathcal{A}(n; n; 1, \ldots, 1)$, being a $Binomial(n, E_{\alpha,\beta}[\Theta])$.

If we now focus attention on this last distribution, we can immediately see that for two

different sets of hyperparameters (α, β) and $(\alpha', \beta') = (c\alpha, c\beta)$, for $c > 0$, we have

$$E_{\alpha,\beta}[\Theta] = \frac{\alpha}{\alpha+\beta} = \frac{c\alpha}{c\alpha+c\beta} = \frac{\alpha'}{\alpha'+\beta'} = E_{\alpha',\beta'}[\Theta]$$

so that T_n possesses the same distribution under different values of the hyperparameters: the one-to-one correspondence between (hyper)parameter value and expected value of Θ (and of X, as well) is lost here. The hyperparameters are no more identifiable through the knowledge of this quantity only.

On the contrary, other allocations will possess sufficient statistics with distributions depending on higher moments of Θ, for example

$$\begin{aligned} Var_{\alpha,\beta}[\Theta] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \neq \frac{\alpha\beta}{(\alpha+\beta)^2(c\alpha+c\beta+1)} \\ &= \frac{(c\alpha)(c\beta)}{(c\alpha+c\beta)^2(c\alpha+c\beta+1)} = Var_{\alpha',\beta'}[\Theta] \end{aligned}$$

which obviously implies different values of the second moment, even if the expected value is equal in the two cases, as seen before.

They will consequently provide additional information to discriminate between different values of the hyperparameters.

In this conditions, $\mathcal{A}(n; n; 1, \ldots, 1)$ cannot be sufficient for other allocations, and thus optimal.

These considerations will play a significant role when combined with the suggestions emerging from the simulations that follow.

Given the large number of possible allocations for even moderate n and the possible effect of different prior parameter specifications, to study the effect of different allocations, we devised the following framework:

- in all simulations, we fixed the values for some parameters:
  - sample size: $n = 10$

  - common scale parameter: $\eta = 2$

- we focussed attention on 4 different allocations, meant to be representative of some subclasses of the whole model class. Specifically

  - $\mathcal{A}(10; 1; 10)$
  - $\mathcal{A}(10; 10; 1, \ldots, 1)$
  - $\mathcal{A}(10; 3; 4, 3, 3)$
  - $\mathcal{A}(10; 3; 1, 3, 6)$

  The first two are the so called '**extremal**' ones already considered for the normal model: they proved to be relevant in the preceding sections - $\mathcal{A}(10; 10; 1, \ldots, 1)$ indeed optimal - and including them in the analysis makes comparisons possible.
  The other two allocations are versions of a '**middle type**' allocation: they both allocate observations to only $k = 3$ units. $\mathcal{A}(10; 3; 4, 3, 3)$ embodies a form of **balanced** allocation for fixed k, while $\mathcal{A}(10; 3; 1, 3, 6)$ uses a more **dispersed** choice. Recall that in the Normal model, for a fixed k, more balanced allocations were optimal.

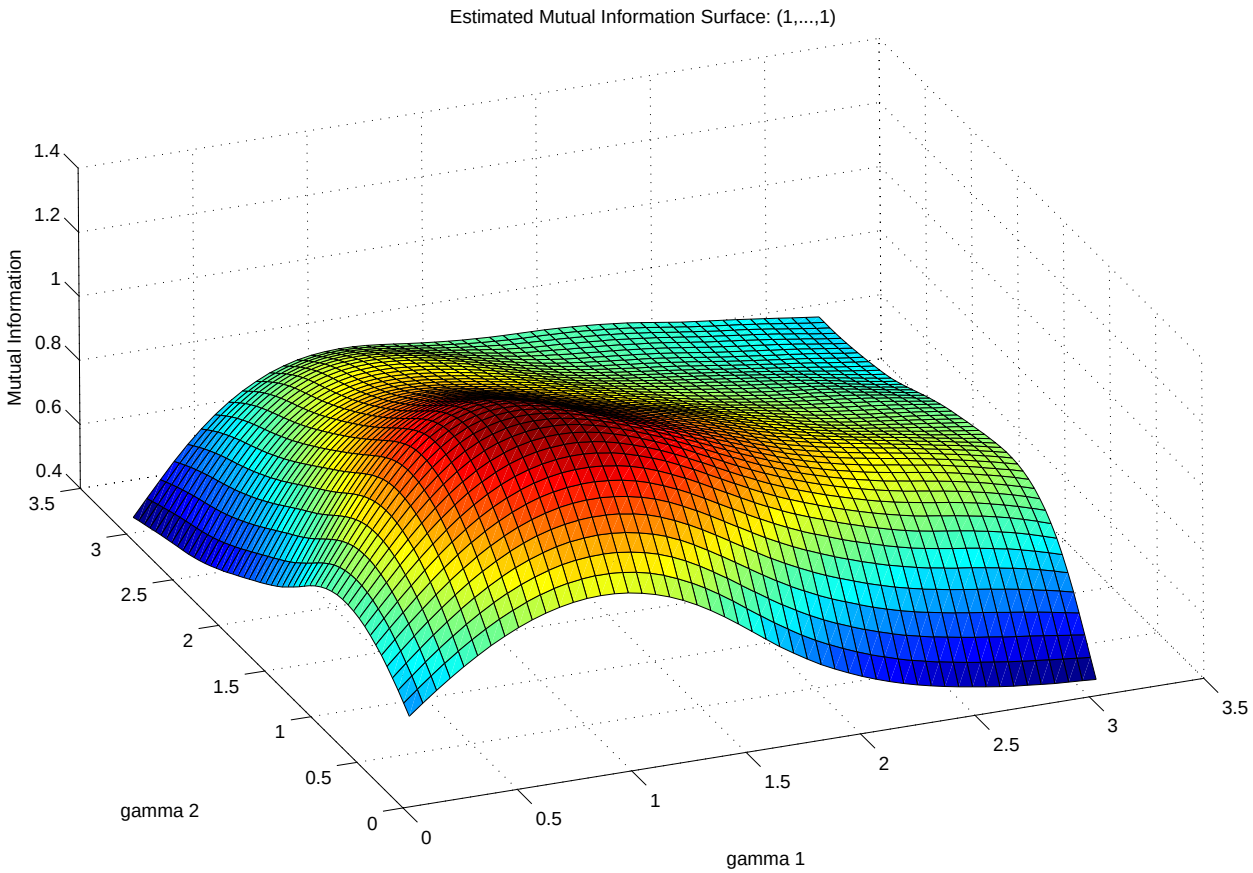
- we considered various possible specifications for prior parameters, γ_1 and γ_2. More specifically we let γ_1 and γ_2 vary on

  - a grid of equally spaced points in $(0, 3)^2$, [10] with
  - 0.2 as the chosen distance between points.

For each of the above mentioned allocations, we estimated the Mutual Information between parameters, (α, β), and observations, $\mathbf{X}$, at each grid point (γ_1, γ_2)

$$I_{(k,(\mathbf{n}))}\left((\alpha, \beta), \mathbf{X} | \gamma_1, \gamma_2, \eta\right)$$

[10] Actually the chosen interval is $(0.05, 3.05)$ to keep parameters into the proper parameter space.

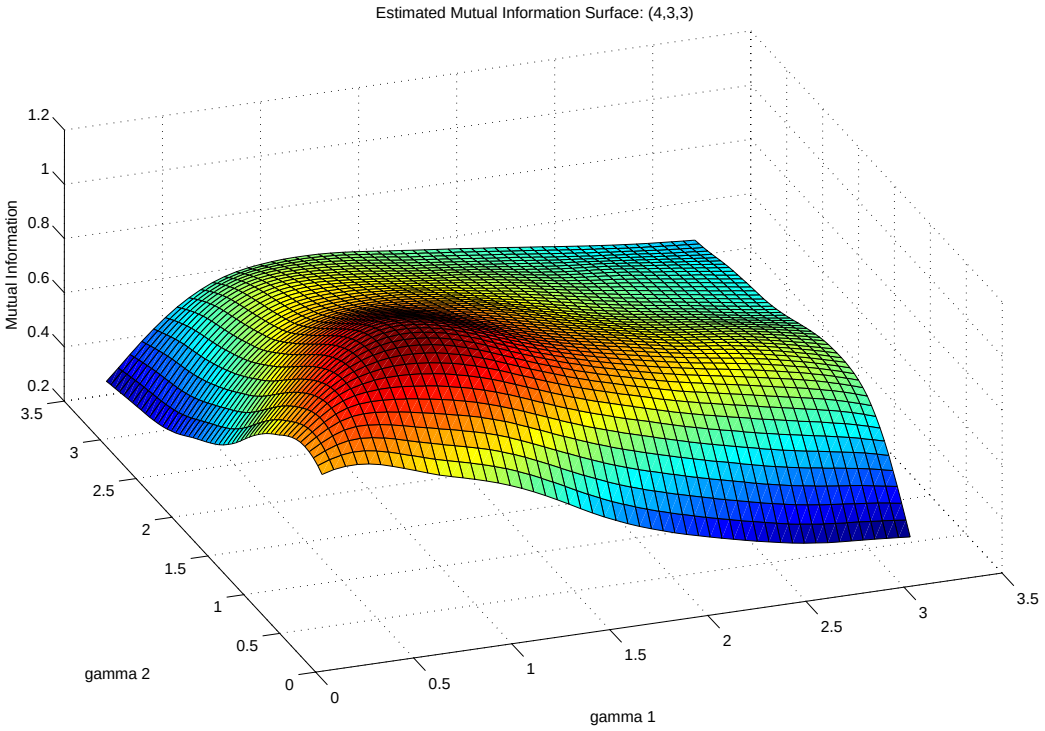


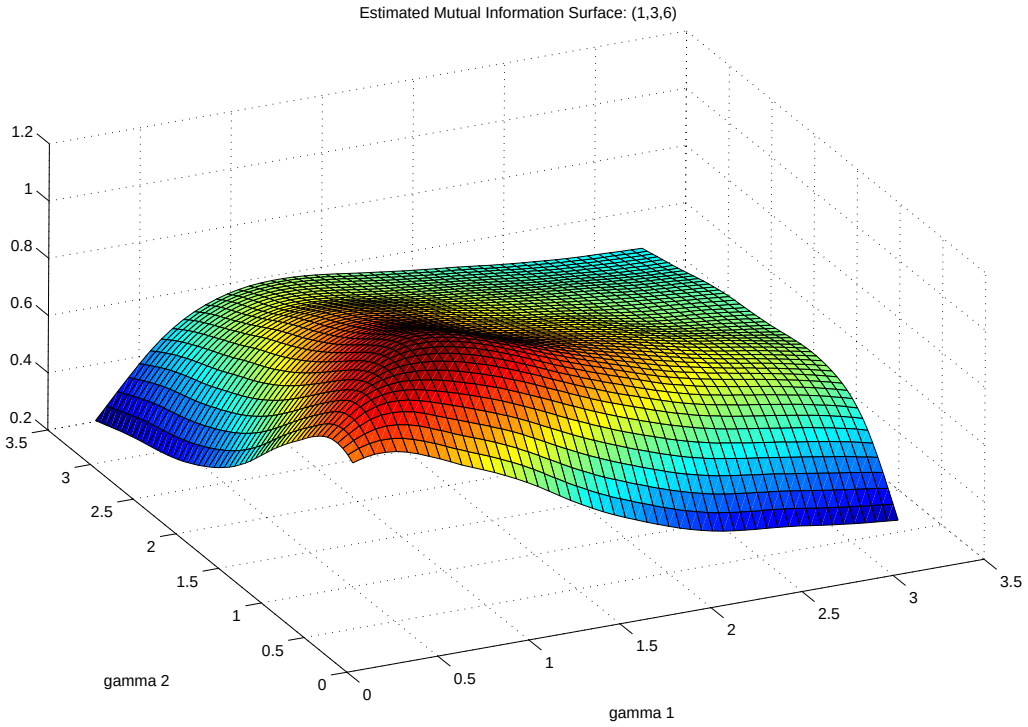
- that is, for each prior corresponding to the parameter values equal to the grid point coordinates.

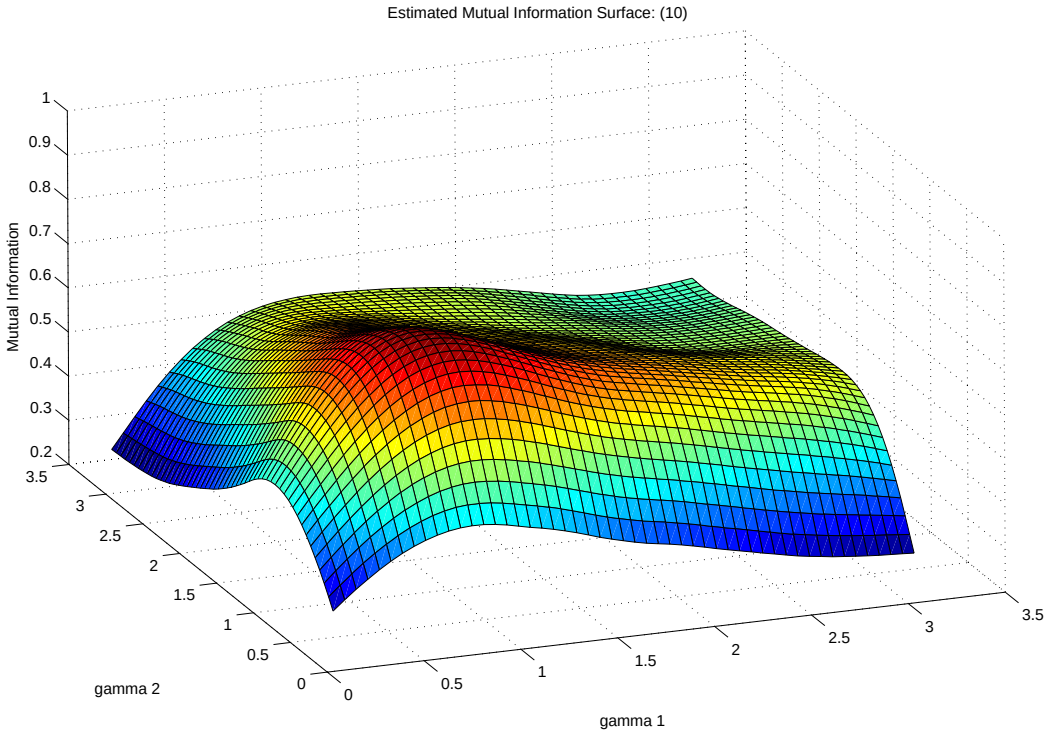
Consequently we obtained what we called a '**Mutual Information Surface**' for each allocation.

Here we show smoothed versions of the originally estimated surfaces, that seem to better capture salient features of the allocations, without requiring an extrapolation effort from the inevitable variability due to the estimation process. The original surfaces are reported in the appendix, with a brief discussion about estimation issues.

Some comments are straightforward.


Estimated Mutual Information Surface: (4,3,3)
Mutual Information
1.2
1
0.8
0.6
0.4
0.2
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5


Estimated Mutual Information Surface: (1,3,6)
1.2
1
0.8
0.6
0.4
0.2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 2
0
0.5
1
1.5
2
2.5
3
3.5
gamma 1


Estimated Mutual Information Surface: (10)
Mutual Information
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5

- Surfaces for different allocations are very similar in shape:

  - *they tend to 0 when at least one of the parameters tends to 0*: this indeed corresponds to prior variances tending to 0, and consequently to a very strong opinion which is hard to modify for observations:

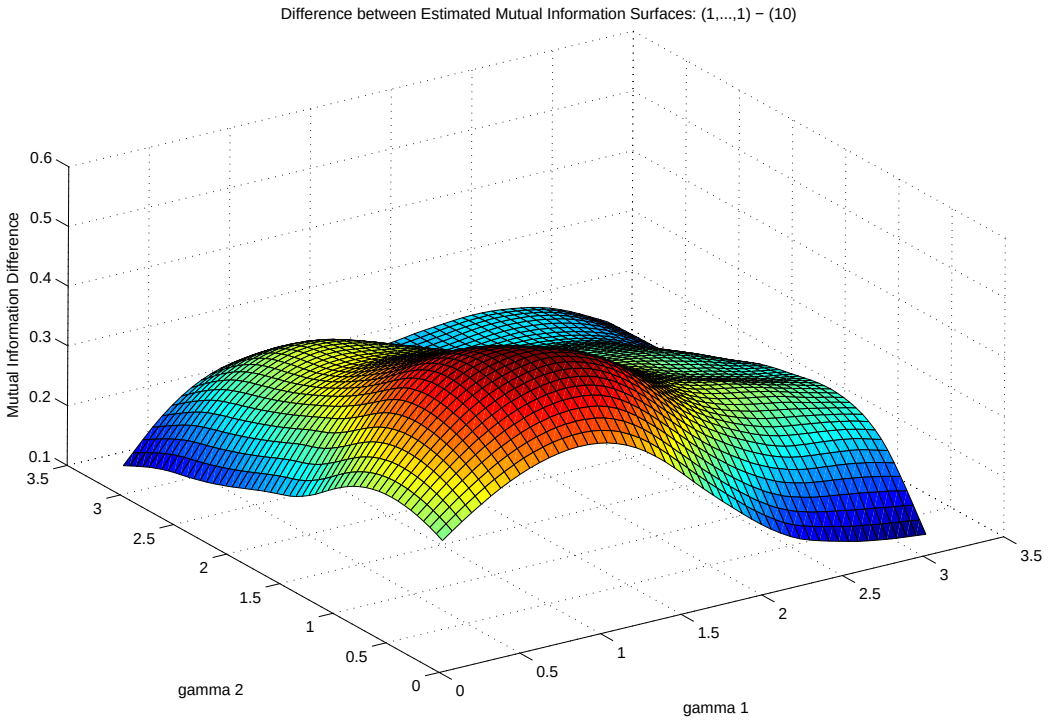
  - *they tend to 0 also when parameters assume relatively large values.*

    A possible explanation is that, in this case, α and β will in turn assume relatively large values, with a 'variance reduction effect' for the distribution of Θ's: these latter will on average be less different, and this reduced variability could clearly prevent us from learning about α and β through their variance. In some sense, we learn about α and β through a more precise, though disturbed, observation of a function of them - $E[\Theta_i|\alpha, \beta] = \alpha/(\alpha + \beta)$ - but we loose part of the information: some variability in the Θ's is necessary to learn more than just the relative magnitude of α with respect to β.

    Clearly it is also necessary to consider that we have at out disposal only indirect observations so that the variability reduction could operate as a 'bottleneck' for information coming from $\mathbf{X}$.

  - There seems to be an interval of values - approximately $(0.5, 1)$ -, common to all considered allocations, over which Mutual Information is maximized: prior distributions characterized by parameter values in this square in the plane could be thought as '**weak**', or less informative in the sense of being more influenced by data.

- Apparently different slopes and heights - though clearly present - are in general mainly a consequence of the different rotations of the axis: we adopted different angles to have the possibility of showing them all and at the same time of giving a global picture of essentially the same surface.



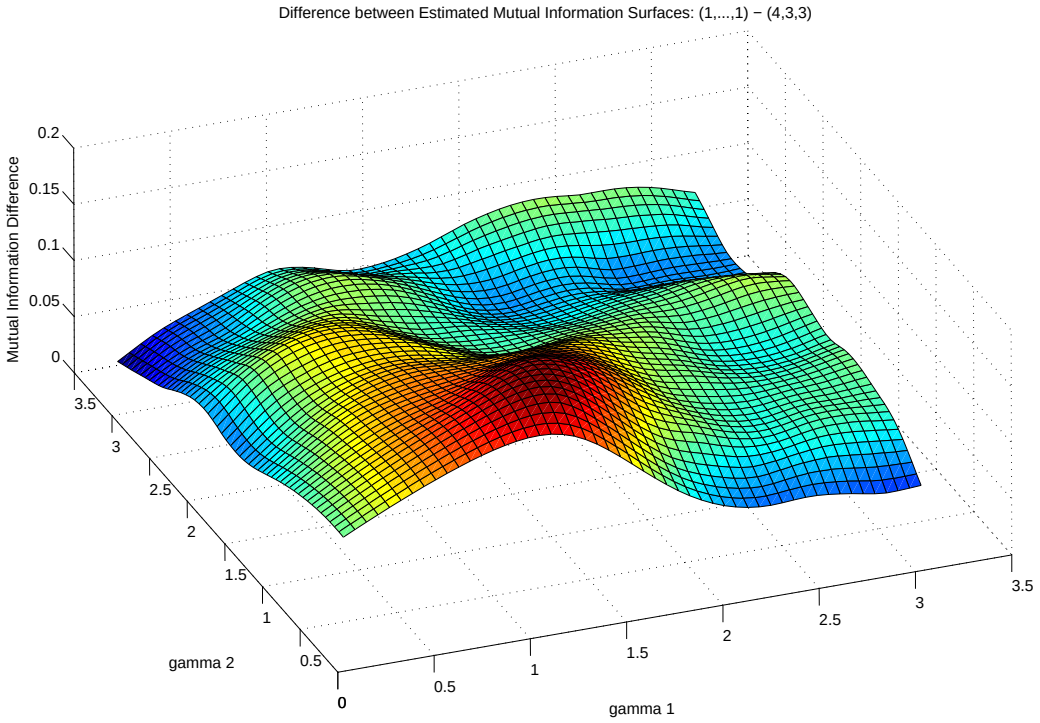
- A difference can be found on the vertical axis: it can be noticed that moving from $(1, ..., 1)$ [11] to (10), the maximum value progressively decreases.

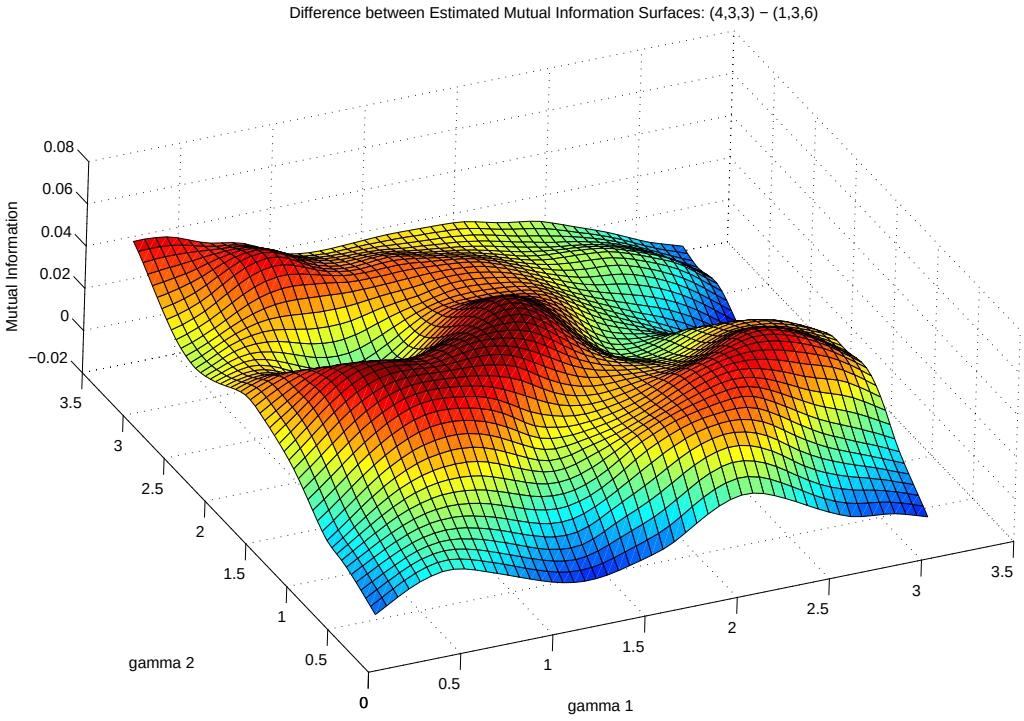
To compare different allocations, we now consider differences between surfaces.
In each of the following pictures, $(n_1, \ldots, n_k)$ - $(n_1^*, \ldots, n_{k^*}^*)$ means that the pictured surface is given by the difference between the surface corresponding to $\mathcal{A}(n; k; n_1, \ldots, n_k)$ and that to $\mathcal{A}(n; k^*; n_1^*, \ldots, n_{k^*}^*)$.

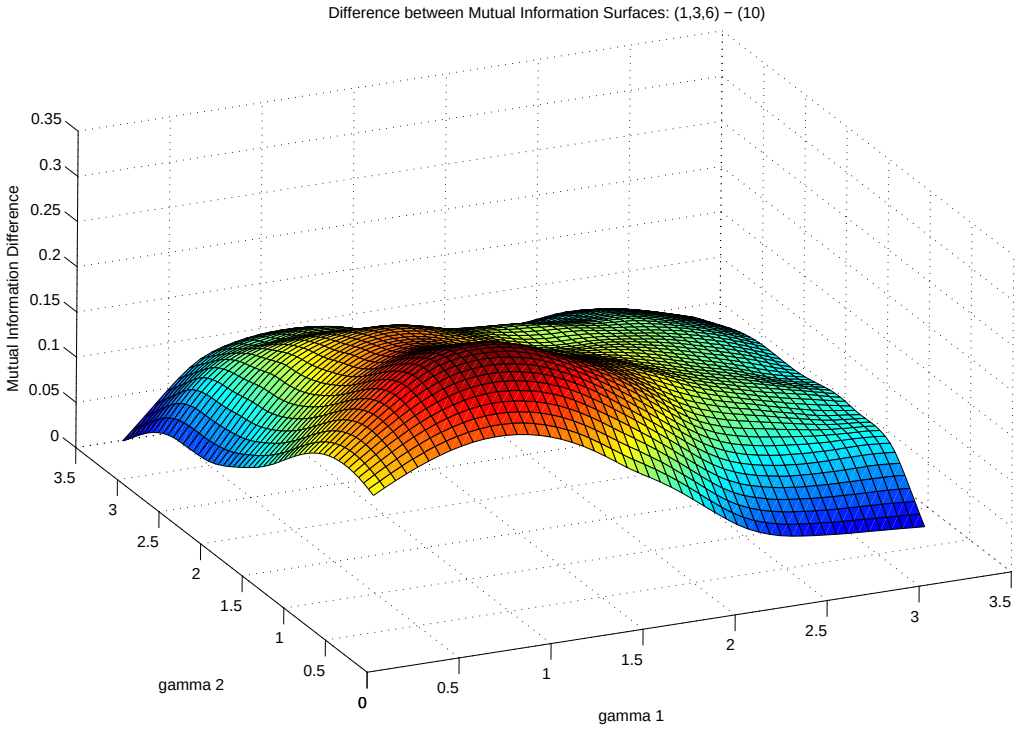
It appears that

- $\mathcal{A}(10; 10; 1, \ldots, 1)$ is clearly more informative than $\mathcal{A}(10; 1; 10)$: surface (10) - (1,. . . ,1) is negative over the whole parameter space considered, sometimes even considerably. In particular, the largest gains occur coherently for the 'weak' priors.

[11] for ease of notation, we use $(n_1, \ldots, n_k)$ in place of $(n; k; n_1, \ldots, n_k)$.


Difference between Estimated Mutual Information Surfaces: (1,...,1) − (4,3,3)
Mutual Information Difference
0.2
0.15
0.1
0.05
0
3.5
3
2.5
2
1.5
1
0.5
0
gamma 2
0
0.5
1
1.5
2
2.5
3
3.5
gamma 1


Difference between Estimated Mutual Information Surfaces: (4,3,3) − (1,3,6)
Mutual Information
0.08
0.06
0.04
0.02
0
−0.02
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5


Difference between Mutual Information Surfaces: (1,3,6) − (10)
Mutual Information Difference
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0.5
1
1.5
2
2.5
3
3.5

- $\mathcal{A}(10; 10; 1, \ldots, 1)$ is more informative than $\mathcal{A}(10; 3; 4, 3, 3)$, the 'middle type' balanced allocation: information gains are somehow smaller but still relevant.

- Between 'middle type' allocations, the balanced $\mathcal{A}(10; 3; 4, 3, 3)$ appears in general more informative than $\mathcal{A}(10; 3; 1, 3, 6)$, even if differences are small in size.

- The less informative 'middle type' allocation, $\mathcal{A}(10; 3; 1, 3, 6)$ is still more informative that the 'concentrated extremal' allocation $\mathcal{A}(10; 1; 10)$.

- Comparisons between surfaces are almost always 'mono-directional': one surface is above the other at each point grid: this seems to suggest that choice of the prior does not influence the ordering of the allocations.

- A part from some macroscopic behaviour, actual shapes of these different surfaces seem to be more the effect of the estimation variance of the algorithm than of structural differences between allocations.

The previous comparisons are clearly of a qualitative kind but the overall picture that emerges, though partial, can synthetically be expressed as

$$\mathcal{A}(10; 10; 1, \ldots, 1) \succ \mathcal{A}(10; 3; 4, 3, 3) \succ \mathcal{A}(10; 3; 1, 3, 6) \succ \mathcal{A}(10; 1; 10) \tag{2.61}$$

where '$\succ$' is defined as 'more informative' in terms of (estimated Mutual Information).

With all the caution due to the limited number of allocations considered and the necessarily restricted simulation settings, the similarity with the results on the Normal Model is clear: the optimal allocation seems to be the same in both cases. Analogously, it also appears that among allocations with the same number of units, the more balanced ones are preferable.
In synthesis, the same ordering of the allocations seems to prevail in both cases.

Finally we can somehow bring together the findings coming from the classical and the Bayesian analysis of this allocation model.

We found that $\mathcal{A}(n; n; 1, \ldots, 1)$ cannot be optimal in the Blackwell-DeGroot sense: this implies that there exist decision problems (at least one) for which decision rules based on observations derived from different allocations lead to lower risks.

But we also found, via simulation, that $\mathcal{A}(n; n; 1, \ldots, 1)$ in general performs better than some other allocations for a specific decision problem, that of maximizing Mutual Information [12]: the comparisons have clearly been based on estimated values of Mutual Information and have taken into account only a relatively small subset of the set of all allocations (for a fixed n) so its findings cannot be considered definitive. Nonetheless the suggestions emerging from these results appear to be substantially in agreement with the hypothesis that $\mathcal{A}(n; n; 1, \ldots, 1)$ qualifies as the optimal allocation in Mutual Information sense: this in turn leads us to the conclusion that *no optimal allocation exists* for the present comparison of hierarchical experiments.

This is essentially the same situation encountered in section 2.2.3: $\mathcal{A}(n; n; 1, \ldots, 1)$ cannot be optimal for the reasons mentioned above but no other allocation can perform better than it in the maximization of Mutual Information.

Thus no other allocation can be optimal, or in different terms, an optimal allocation does not exists.

2.4 Asymptotic Optimality of $\mathcal{A}(n; n; 1, \ldots, 1)$

In the previous sections it has emerged how it is generally hard to find an optimal allocation.

[12] Recall the interpretation of Mutual Information as expected Utility in a purely inferential decision problem.

One way of tackling with these difficulties has been considered in section 2.4: the use of simulation tecniques.

In the present section, we follow a different approach and basically rely on asymptotic arguments. Our aim will be to show that when the sample size, n, increases it seems more and more reasonable to adopt allocation $\mathcal{A}(n; n; 1, \ldots, 1)$.

For large sample sizes, this latter indeed will progressively stand out and - always in terms of Mutual Information - dominate the other allocations, characterized by a fixed number, r, of direct observations, with r. The problem then appears to be quite simplified in this framework.

We mention in passing that this result is not so obvious: one could reasonably expect that, as n increases, the problem of finding this optimal allocation becomes still harder: the class of possible allocations is indeed fast growing in n and in the enriched availability of chosing the number of direct observations, r.

The main intuition leading to the result is derived by Fact 2.2.2: allocations with a smaller number of direct observations (or units or individuals) seem to possess a limited 'informative potential'.

We state the result and then illustrate the line of reasoning leading to it.

In the statement of the result, the conditional density of an indirect observation, X, given the value of the highest order parameter, μ, $f_{X|\mu}(x|\mu)$, defined by

$$f(x|\mu) = \int_{\Theta} f(x|\theta)\pi(\theta|\mu)d\theta$$

plays a relevant role.

For this density, we use the notation $\mathcal{L}_n(\mu) = \sum_{i=1}^{n} log f(x_i|\mu)$, the log-likelihood for the preceding model under i.i.d. sampling, with a prime "$\prime$" denoting differentiantion w.r.t. μ.

In particular we will have

$$\mathcal{L}_n''(m) = \frac{\partial^2}{\partial \mu^2}\mathcal{L}_n(\mu)\big|_{\mu=m}$$

and, with $\hat{\mu}_n$ denoting the MLE of μ, we will have $\sigma_n = -\mathcal{L}_n''(\hat{\mu})^{-1}$.

Note that analogous definitions as the ones just introduced will be needed for the density $\pi(\theta|\mu)$. We will not go into these redundant notation: it just suffices to consider $\pi(\theta|\mu)$ in place of $f(x|\mu)$.

Furthermore we list some standard Regularity Conditions that will help define the framework:

1. The parameter space $\mathcal{M} \subseteq \Re$.

2. μ_0 (the true parameter value) is an interior point of $\mathcal{M}$.

3. The prior distribution of μ, $\pi(\mu)$, has a density with respect to Lebesgue measure that is positive and continuous at μ_0.

4. There exists a neighborhood $N_0 \subseteq \mathcal{M}$ of μ_0 on which $\mathcal{L}_n(\mu)$ is twice continuosly differentiable with respect to μ, a.s. $[P_{\mu_0}]$.

5. σ_n goes to 0 in probability.

6. For $\delta > 0$, define $N_0(\delta)$ to be the open ball of radius δ around μ_0. If $N_0(\delta) \subseteq \mathcal{M}$, then there exists $K(\delta) > 0$ such that

$$lim_{n\to\infty} P_{\mu_0}\left(sup_{\mu\in\mathcal{M}-N_0(\delta)}\sigma_n[\mathcal{L}_n(\mu) - \mathcal{L}_n(\mu_0)] < -K(\delta)\right) = 1 \tag{2.62}$$

7. For each $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that

$$lim_{n\to\infty} P_{\mu_0}\left(sup_{\mu\in N_0(\delta(\epsilon))} \|1 + \mathcal{L}_n''(\mu)\sigma_n\| < \epsilon\right) = 1 \tag{2.63}$$

Lemma 2.4.1 *If the families of densities* $\left\{f_{\Theta|\mu}(\theta|\mu); \mu \in \mathcal{M}\right\}$ *and* $\left\{f_{X|\mu}(x|\mu); \mu \in \mathcal{M}\right\}$ *both satisfy the Regularity Conditions defined above, allocation*

$\mathcal{A}(n; n; 1, \ldots, 1)$ *is* ***asymptotically optimal****, in the sense that, for any other allocation* $\mathcal{A}(n; r; n_1, \ldots, n_r)$*, with fixed* r*, there exists a* n_0 *such that,* $\forall n \geq n_0$*,*

$$I\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_r)_1^{n_r})\Big) \leq I\Big(\mu, (X_1, \ldots, X_n)\Big) \tag{2.64}$$

In other terms, Lemma 2.4.1 simply states that, for a sufficiently large sample size n, $\mathcal{A}(n; n; 1, \ldots, 1)$ will dominate any other allocation that concentrates indirect observations on same units: in a pragmatic sense, the results suggests that when n is large we can rely on $\mathcal{A}(n; n; 1, \ldots, 1)$ without losing too much information, being instead confident that this experimental design will provide us with nearly highest amount of information.

To arrive at the result note that, however large is n,

$$I\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_r)_1^{n_r})\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(n)})\Big)$$

The middle term, independent of n, in the preceding chain of inequalities is actually an upper bound for the amount of information provided by any allocation based on a fixed number r of direct observations. Furthermore, writing, by the chain rule,

$$I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(n)})\Big) = \sum_{i=1}^{n} I\Big(\mu, \Theta^{(i)} | \Theta_1^{i-1})\Big)$$

we find that $I(\mu, (\Theta^{(1)}, \ldots, \Theta^{(n)}))$ is a monotone non-decreasing function of n, by non-negativity of Mutual Information.

This clearly holds also for $I(\mu, X_1^n)$ (as for any other set of r.v.'s).

Even in allocation $\mathcal{A}(n; n; 1, \ldots, 1)$ we have the inequality

$$I\Big(\mu, (X_1, \ldots, X_n)\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(n)})\Big)$$

but if the right-hand side of the inequality increases without bound, also the left-hand side would be free to tend to ∞, and there would exist an n_0 such that

$$I\Big(\mu, ((X_1)_1^{n_1}, \ldots, (X_r)_1^{n_r})\Big) \leq I\Big(\mu, (\Theta^{(1)}, \ldots, \Theta^{(r)})\Big) \leq I\Big(\mu, (X_1, \ldots, X_{n_0})\Big)$$

But, under the assumed Regularity Conditions, the posterior distribution of μ given $(X_1, \ldots, X_n)$ (or given $(\Theta_{(1)}, \ldots, \Theta_{(n)})$) is asymptotically normal with variance σ_n tending to 0, as $n \to \infty$.

The entropy of this posterior distribution can then be suitably approximated by the entropy of a normal random variable with vanishing variance, that is

$$h(\mu|\mathbf{x}) \approx \frac{1}{2}\log(2\pi e\sigma_n) \to -\infty \quad n \to +\infty \quad P_X - a.s.$$

It is then immediate to conclude by expression 2.36 that, $I_{(n,(1,\ldots,1))}(\mu, X_1^n)$ will tend to $+\infty$.[13]

The assumed Regularity Conditions seem to pose mild restrictions on the possible models to be considered, widening the potential use of the result.

[13] See also Ibragimov and Hasminskii [33] for an asymptotic approximation of $I(\mu, (X_1, \ldots, X_n)) \sim \frac{d}{2}\log n$, where d is basically the dimension of the parameter space.

Chapter 3

Model Selection via Information Measures

The use of Information quantities in model selection procedures is widespread in statistical literature, as witnessed by the well known criteria of AIC, BIC and DIC.

In the preceding chapter we actually used a specific measure of Information -Mutual Information - to perform Bayesian Experimental Design, which can be interpreted as a form of "prior model selection" procedure.

It seems natural to ask if the concepts and tools used and developed in the preceding chapters can be modified and adapted to perform a different task: model selection for hierarchical models.

In this work, Mutual Information has evidently played a crucial role but, for some reasons soon made evident, its direct application in this framework is not feasible, and it requires some thinking and potential modifications.

Synthetically this chapter is still work in progress.

3.1 The Choice between Hierarchical Models

First of all, since we are still dealing with hierarchical models, it appears necessary to point out the fundamental difference between the problem treated in Chapter 2 and the one we are addressing here.

Chapter 2 dealt with the issue of Optimal Experimental Design: the statistician planning to collect evidence about a phenomenon of interest seeks the best way to 'question Nature' so as to gain as much Information as possible.

Observations have still to be taken: this gives the researcher an additional tool in her effort to gain Information, the experimental design, that is the possibility to influence, within a certain degree, the stochastic process that will generate the data.

The present chapter deals instead with the situation when the data have already been collected: these latter could be the product of a well designed experiment but might also be the result of an observational study, over which the control of the researcher has been moderate or completely absent.

In broad generality, we could still state that the goal of the statistician is to gain as much Information as possible, but now Information appears to be of a different nature: instead of looking for the best 'questions' to ask Nature, she is more likely trying to infer what questions best suit Nature's already available answers, the data.

The statistician is then looking for the best interpretation of the data: a statistical model.

Coherently with the choice of previous chapters, the possible models considered here will be hierarchical, basically with 3 levels as before.

If the assumption of a hierarchical structure was clearly justified by the chosen sampling design in the experimental setting, the adoption of a hierarchical model in data analysis can be made for several reasons: in general, these class of models are characterized by a huge amount of flexibility, both with respect to the prior opinion and to the data generating mechanism that can be represented by one of its member: thus they can be adapted

to many situations in a rich variety of contexts.
Literature on hierarchical analysis is huge: for a fundational paper in the Bayesian approach recall the already mentioned Lindley and Smith [43], or for more recent references see Congdon [11] or, for the possibility of integrating experimental and observational data in a variety of situations, Clarke and Gelfand [9].

To be more precise, the structures we will consider are almost identical, at least formally, to the ones analysed in chapter 2:

- $\mu \sim \pi_\mu$
- $\Theta_1, \ldots, \Theta_k | \mu \sim \pi_{\Theta|\mu}(\cdot|\mu)$
- $X_i | \Theta_j = \theta_j \overset{i.i.d.}{\sim} f_X | \Theta(\cdot|\theta_j)$ for $i \in S_j$

where $g = (S_1, \ldots, S_k)$ represent a partition of the set of integers indexing the observations, $I = \{1, \ldots, n\}$.
The notational difference in the above model specification, with respect to the one adopted in chapter 2, is indeed substantial: allocations were identified by vectors like $(n; k; n_1, \ldots, n_k)$ stating only the total number of observations, the number of units (groups) and the *number* of observations on each unit.
The parameter g, indexing the possible hierarchical models, provides additional information: not only the number of groups in the partition and the number of observations in each group are preserved but also the exact labelling of the observations in each group is specified.
This is necessary since data have already been observed, and the values of the observations are fixed quantities in the problem.
Furthermore a crucial fact is that the *grouping of observations* into subsets characterized by the same value of the middle-level parameter Θ is *uncertain*.

While in the experimental setting, the allocation of observations to units was under the control of the statistician, in the present setting a *partition*, g, can be interpreted just as *an hypothesis about the potential relations between observations*: those belonging to the same group are expected to be more dependent or more similar to one another than to those of other groups.

Confronted with the uncertainty about g, there are basically two available choices, essentially depending on the focus and goals of the analysis:

- If g represents on its own a 'quantity' of interest in the problem at hand, we can try to determine the 'best' grouping - that is, estimate g - where the qualifier 'best' is defined in terms of some inferential criterion: this goal is clearly the one characterizing the research field of (model based) Cluster Analysis.

  Actually the clustering is here almost identical to model choice.

  Many criteria can be chosen: wishing to proceed within a Bayesian approach, it clearly makes sense to state a prior distribution on the uncertain parameter g, $p(g)$, and subsequently obtain the posterior distribution given the data, $\mathbf{x}$, $p(g|\mathbf{x})$. It is then possible to estimate g by the posterior mode, for example, or since in general computational difficulties arise, the choice will be typically guided by Bayes factors. Note however that in this case the specification of a prior distribution on g is not necessary: another optimality criterion could be specified.

- If instead g is not of primary interest, it essentially represents or a nuisance parameter or alternatively a useful tool to optimize inference [29]: for example, the analysis could focus on the estimation of parameter, θ_j, characterizing a particular X_i. See for example, Malec and Sedransk [46] or Consonni and Veronese [12]. Since g determines the observations in the group with X_i, it crucially establishes which observations are to be used to estimate θ_j.

  Thinking about the $X_i's$ as the outcomes of experiments, the role of g is then crucial

in determining how influent the result of experiment X_h is on conclusions about experiment X_i: g then establishes if and to what extent it is possible to combine the results of different experiments. The role of the prior distribution, $p(g)$, will be very important.

These situations are typical of Meta-Analysis, where one tries to 'borrow strenght' analysing jointly a set of potentially interdependent experiments, or Model Averaging, where the uncertainty relative to the 'true' data generating mechanism is not faced by chosing a single 'optimal model' but by accurately weighting the inferences provided from all the competing models.

In the present setting, the competing models are clearly all the hierarchical models generated by partitioning the observations in groups.

These approaches fall into the field of Bayesian Clustering or into the class of the so called Partition Models, for obvious reasons.

In dealing with partitions, in general we are faced with some peculiar problems.
A critical one is the number of partitions of n observations in any number of groups, k, $k = 1, \ldots, n$. These numbers, B_n for $n = 0, 1, 2, \ldots$, also known as Bell numbers, become very huge even for moderate n: for example, for just $n = 10$ observations, there are more than 100 thousands partitions.
The class of models that need to be considered becomes quickly too large to be fully explored, if the proposed decision procedure has to be applied to realistic problems.
To obviate to such inconvenience - that seriously menaces to make heavy the computational burden and to harden the elicitation of a prior distribution on this huge class of objects - some strategies are available:

- As suggested by Hartigan [29] and Consonni and Veronese [12], one can reasonably assume that the observations are already grouped in a relatively moderate number

of subsets: the partitioning takes place at the level of the parameter characterizing these groups of observations. It is then possible to reduce considerably the dimension of the model space $\mathcal{G}$, sometimes even to the extent that makes possible explicit consideration of each partition.

Furthermore this dimensionality reduction aids in the elicitation of a coherent prior distribution for g: in turn, when $\mathcal{G}$ is not too large, it is also possible to consciously set to 0 the probabilities attached to some partitions, further reducing 'probabilistically' the dimension of the model space.

- A second possibility consists of exploring the model space $\mathcal{G}$ according to some algorithm with good properties, giving up from the beginning the requirement of an exhaustive exploration. See Quintana and Iglesias [48], for example.

- Finally, a minor strategy could be to impose some restriction on the cardinality of the partitions to be considered, that is, $|g| \leq \tilde{k}$, that is to limit a priori the number of groups in each partition, on the basis of pragmatic considerations.

3.2 A Criterion based on the Kullback-Leibler Divergence

In this section an attempt is made to apply the kind of reasoning, exploited in the experimental design setting of the previous chapter, to selection of hierarchical models.
We mention in passing that nothing prevents us from adopting the Mutual Information criterion to evaluate the performances of experiments that are not hierarchical: the criterion has originally been formulated by Lindley [41] with respect to generic experiments.
Obviously even the potential model selection criterion that will be here formulated need not be applied to hierarchical models exclusively.

The effort to choose hierarchical models directly on the basis of the amount of Mutual Information they provide immediately runs into the inevitable observation that

$$I(\mu, \mathbf{x}) = 0 \tag{3.1}$$

Since data, $\mathbf{x}$, are known and fixed constants, by definition, they provide no (Mutual) Information about the parameters.

A partial remedy to this fact may be found in the original reasoning that lead to the adoption of Mutual Information: $I(\mu, \mathbf{X})$ has initially been derived as an **expected reduction in Uncertainty** (as measured by Entropy) provided by an experiment, $\mathbf{X}$; recalling the definition

$$I(\mu, \mathbf{X}) = \int_{\mathcal{X}} f_\pi(\mathbf{x}) \left\{ \int_{\mathcal{M}} \pi(\mu|\mathbf{x}) log \frac{\pi(\mu|\mathbf{x})}{\pi(\mu)} d\mu \right\} d\mathbf{x} \tag{3.2}$$

When data are observed, we indeed face an **actual reduction in Uncertainty**,whose value is provided by the inner integral in the above expression,

$$KL(\pi(\cdot|\mathbf{x})|\pi(\cdot)) = \int_{\mathcal{M}} \pi(\mu|\mathbf{x}) log \frac{\pi(\mu|\mathbf{x})}{\pi(\mu)} d\mu \geq 0 \tag{3.3}$$

Then if we agreed in choosing experiments on the basis of an expected reduction in Uncertainty, it would seem reasonable to choose models on the basis of the actual reduction in Uncertainty they are able to provide: essentially we choose the model that appears to be more convincing and that seems to cast away our doubts, consolidating our opinion on firmer ground.

It seems that a different interpretation can also be given to such a criterion.
We can write

$$KL(\pi(\cdot|\mathbf{x}) \mid \pi(\cdot)) = \int_{\mathcal{M}} \pi(\mu|\mathbf{x}) log \frac{f(\mathbf{x}|\mu)}{f_\pi(\mathbf{x})} \, d\mu \tag{3.4}$$

where $f_\pi(\mathbf{x})$ is the marginal density of the observables with respect to the prior π.
Note that a more appropriate notation for the model would be $f(\mathbf{x}|\mu, g)$ since g indexes

the class and each member is given by

$$f(\mathbf{x}|\mu, g) = \prod_{i=1}^{|g|} \int_{\Omega} \prod_{i \in S_j} f_{X|\Theta}(x_i|\theta_j) \pi_{\Theta|\mu}(\theta_j|\mu) d\theta_j \tag{3.5}$$

and analogously we should write $\pi(\mu|\mathbf{x}, g)$ in place of $\pi(\mu|\mathbf{x})$.

Now, if we act with the goal of maximizing 3.4, we are trying to choose the model $f(\mathbf{x}|\mu)$ whose members appear a posteriori maximally dispersed ('far' in terms of logarithmic distance) from their 'prior mean', $f_{\pi}(\mathbf{x})$, in terms of density value for the observed data. Using a basic interpretation of dispersion, this would imply that, a posteriori, the 'prior mean' is not very much representative of the family of distributions, or alternatively that we tend to accept models that provide widely variable probability evaluations.

Such models embody some form of prudential approach: they include members that model observables' behaviour in a significantly different way, from a probabilistical point of view. They could be thus defined rich or flexible.

We could name this potential criterion **Maximum Kullback-Leibler Divergence** rule.

At the present state, it is difficult to conclude about its properties: the analogy with Mutual Information that lead to it partially justifies its adoption but a deeper study is needed.

Some insight into the implications of its use can be drawn from Example 3.2.1, where its explicit form is found for the normal hierarchical model.

The previous considerations don't take into account in any way the computational difficulties of the problem.

These follow from

- the already commented huge number of models (partitions), that will require the design of an algorithm for the search for the maximum in the model space $\mathcal{G}$;

- the usual MCMC step, to sample values from the posterior, in general is not sufficient to evaluate information quantites as the ones presented above, since they require knowledge of the analytical expression of the posterior to evaluate the posterior-to-prior density ratio, or knowledge of the overall marginal at $\mathbf{x}$ when the ratio is expressed in term of densities for the observables. Otherwise, approximations to this ratios will be required.

A comparison with other widely used methods mentioned above - Partition Models and Clustering methods - is hard at the moment: the application of the criterion is clearly more similar to a clustering method than to the fully Bayesian inference possible in Partition Models.

A potential advantage is the fact that it does not require the elicitation of a prior for the partition parameter, g: this task can be quite difficult in situations where it has not been possible to significantly reduce the cardinality of $\mathcal{G}$.

On the other side, this aspect precludes the possibility of a complete Bayesian analysis, and of the consequent fully probabilistic reasoning.

Example 3.2.1 *Consider the 2-level normal hierarchical models of the previous chapter, where now* $\mathbf{X} = \mathbf{x}$*: that is the* X*'s are observed.* n *is still the sample size.*

k *will now represent the number of groups in which the observations are divided. Note that now to describe a model is no more sufficient to fix a* k*-vector of integers summing up to* n*: clearly two models with the same* k *and the same number of observations for each group are different if the groups are formed by using different subsets of observations.*

In general, for each k*, it will be necessary to specify the group membership for each observation, for example through a* k*-vector of '0's and '1's, summing up to '1'. Notation becomes quite heavy, and for the present purposes it is not needed.*

We will be content with the consideration of just one single model, to illustrate what the maximum Kullback-Leibler criterion states for this particular model class.

$\mathbf{x}^{(i)}$ *will then represent the observations in group* i, n_i *and* $\overline{x}_i$ *will represent their number and their mean respectively.*

We recall briefly that the Kullback-Leibler divergence between a $\mathcal{N}(\mu_1, \sigma_1^2)$ *and a* $\mathcal{N}(\mu_2, \sigma_2^2)$ *is given by*

$$KL\left((\mu_1, \sigma_1^2)|(\mu_2, \sigma_2^2)\right) = \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} \tag{3.6}$$

Since prior and posterior for μ *are normal in this case, the criterion can then be expressed as*

$$KL\left((\mu_p, \tau_p^2)|(\mu_0, \tau_\mu^2)\right) = \frac{1}{2}\log\frac{\tau_\mu^2}{\tau_p^2} - \frac{1}{2} + \frac{\tau_p^2}{2\tau_\mu^2} + \frac{(\mu_0 - \mu_p)^2}{2\tau_\mu^2} \tag{3.7}$$

with μ_p *and* τ_p^2 *posterior mean and variance respectively, given by*

$$\mu_p = \sum_{i=0}^{k} w_i \overline{x}_i \qquad \tau_p^2 = \tau_\mu^2 \left(1 + \tau_\mu^2 \sum_{i=1}^{k} \frac{1}{\tau_\Theta + \sigma^2/n_i}\right)^{-1} \tag{3.8}$$

with $\overline{x}_0 \equiv \mu_0$, $w_i = \frac{1/(\tau_\Theta + \sigma^2/n_i)}{1/\tau_\mu^2 + \sum_{j=1}^{k} 1/(\tau_\Theta + \sigma^2/n_j)}$ *for* $i = 1, \ldots, k$, *and* $w_0 = \frac{1/\tau_\mu^2}{1/\tau_\mu^2 + \sum_{j=1}^{k} 1/(\tau_\Theta + \sigma^2/n_j)}$.
This divergence can be written as

$$KL\left((\mu_p, \tau_p^2)|(\mu_0, \tau_\mu^2)\right) = \frac{1}{2}\left\{\log\frac{\tau_\mu^2}{\tau_p^2} - \frac{\tau_\mu^2 - \tau_p^2}{\tau_\mu^2}\right\} + \frac{(\mu_0 - \mu_p)^2}{2\tau_\mu^2} \tag{3.9}$$

In the present framework, for every $\mathbf{x}$, $\tau_p^2 \leq \tau_\mu^2$, *and the term in curly braces is a positive, monotone decreasing function of* τ_p^2 *on* $(0, \tau_\mu^2)$*: since the criterion requires the maximization of 3.7, models that reduce posterior variance more than others will be preferred. So this term embodies a "posterior precision" issue.*

Recalling from chapter 2 that the optimal allocation was the one minimizing the posterior variance, this first component will guide our choice towards the model with n different groups each with only one observation.

The final choice will however be influenced by the second term which seems to require a model that shifts the posterior mean as far as possible from the prior mean. In this regard, we are pursuing the above mentioned 'partially non-informative' analysis, letting data change as much as possible our opinion on prior mean: implicitly we are trying to

specify a weak prior.

Note that the same weak prior could be obtained by raising the prior variance but that would raise posterior variance too: for a fixed τ_μ^2, the criterion implicitly tries to introduce prior weakness by uncertainty about model structure (that is about the dependence structure among observations, and between observations and parameters).

The two component considered together define a mixed criterion which tries to balance a maximum gain in posterior precision with a weak opinion on its actual value.

It appears difficult to develop further the analysis since specific values obtained for the observations and the prior mean clearly define the specific solution reached at the end.

Simulation studies are in program. Note that by equation 3.8 the criterion can then be expressed more concisely as

$$\begin{aligned} KL\left((\mu_p, \tau_p^2)|(\mu_0, \tau_\mu^2)\right) = \frac{1}{2}\log\left(1 + \tau_\mu^2 \sum_{i=1}^{k} \frac{1}{\tau_\Theta + \sigma^2/n_i}\right) - \frac{1}{2} \\ + \frac{1}{2}\left(1 + \tau_\mu^2 \sum_{i=1}^{k} \frac{1}{\tau_\Theta + \sigma^2/n_i}\right)^{-1} + \frac{(\mu_0 - \sum_{i=0}^{k} w_i \overline{x}_i)^2}{2\tau_\mu^2} \end{aligned} \tag{3.10}$$

This expression can then by used directly for simulation.

Chapter 4

Conclusions and Outline

The use of Mutual Information, and of measures of Information in general, appears to be natural in Statistics: they seem to be 'native' concepts in Statistical Theory, suggest some kind of intuitive manipulation of information relations where Uncertainty is exchanged with Information, and this in turn with Dependence.

Unfortunately this easiness of manipulation does not go along with a general analytical tractability of the same quantities, that almost always requires computational tools.

We obtained explicit expressions of Mutual Information for 2-level normal hierarchical models, under the restrictive hypothesis of known and equal variances, and under the slightly less restrictive one of known but unequal variances.

Surely, at least two extensions seem desirable: first of all, we can remove the unrealistic hypothesis of known variances. Almost surely this will require the use of computational tools even in the normal case, and the effect of these new assumptions appear to be uncertain. If the inferential focus still remains fixed on the highest level parameters, it seems possible to conjecture that minor changes are to be expected in the final conclusions: unknown but still equal variance should inflate the variance of the observations and induce some dependence among them, but the best way to cope with this still seems to be to choose the usual extremal allocation, which even in this case minimizes the dependence

induced by the $\Theta's$.

While if either unequal unknown variances are considered or the variance parameter(s) are assumed to be of interest, conclusions appear difficult to guess. Further work is needed, and it could probably contribute to make this approach more applicable.

A second extension would be to raise the number of levels in the considered hierarchical model. Considering varying numbers of parameters at different levels, the number of potential designs increases considerably: this seems to make harder the maximization problem but not to change the nature of the problem.

There exists a third possibility: to assume some forms of dependence among observations taken on the same unit. For example, always considering the Normal model with just 2 observations and negatively correlated observations taken on the same unit, it is possible to show that the allocation with both observations on the same unit is optimal, for opportune values of the correlation coefficient.

Outside the Normal model, conclusions can be drawn almost only via computations: the insights provided by our simulations, even if they agree with Normal results, are difficult to extend to other cases. Furthermore no general guideline can be put forward if we introduce some form of heterogeneity in the groups, analogoulsy to the unequal variances case in the normal model: we can reasonably guess that the optimal allocation will change but an explicit rule can hardly be worked out. Faster and more precise algorithms are thus called for.

Chapter 5

Appendix: Mutual Information Surfaces and Algorithm

The Algorithm

To compute the Estimated Mutual Information at each grid point γ_1, γ_2, we used the following Monte Carlo algorithm.
It is relatively straightforward and it can easily be adapted to different specifications of the distributions at different levels of the hierarchical model.
In part this generality can also be a drawback since it does not exploit specific properties of the densities involved.

We fix the number of iterations, I, and of what we could call 'subiterations', J.
For each $h = 1, \ldots, I$

1. Generate

   α_h from $Gamma(\gamma_1, \eta)$

   β_h from $Gamma(\gamma_2, \eta)$

2. Generate
   $(\theta_1^{(h)}, \ldots, \theta_k^{(h)})$ independently from $Beta(\alpha_h, \beta_h)$

3. For $l = 1, \ldots, J$, generate
   $(\theta_1^{(h,l)}, \ldots, \theta_k^{(h,l)})$ independently from $Beta(\alpha_h, \beta_h)$

4. Generate
   $s_i^{(h)}$ independently from $Binomial(n_i, \theta_i^{(h)})$ for $i = 1, \ldots, k$

5. Estimate the conditional distribution of the vector of sufficient statistics $(s_1^{(h)}, \ldots, s_k^{(h)})$, given (α_h, β_h) as

$$f(s_1^{(h)}, \ldots, s_k^{(h)} | \alpha_h, \beta_h) = \frac{1}{J} \sum_{l=1}^{J} \prod_{i=1}^{k} Binomial(s_i^{(h)} | n_i, \theta_i^{(h,l)})$$

6. Estimate the marginal distribution of the vector of sufficient statistics as

$$f(s_1^{(h)}, \ldots, s_k^{(h)} | \gamma_1, \gamma_2) = \frac{1}{I} \sum_{h=1}^{I} f(s_1^{(h)}, \ldots, s_k^{(h)} | \alpha_h, \beta_h)$$

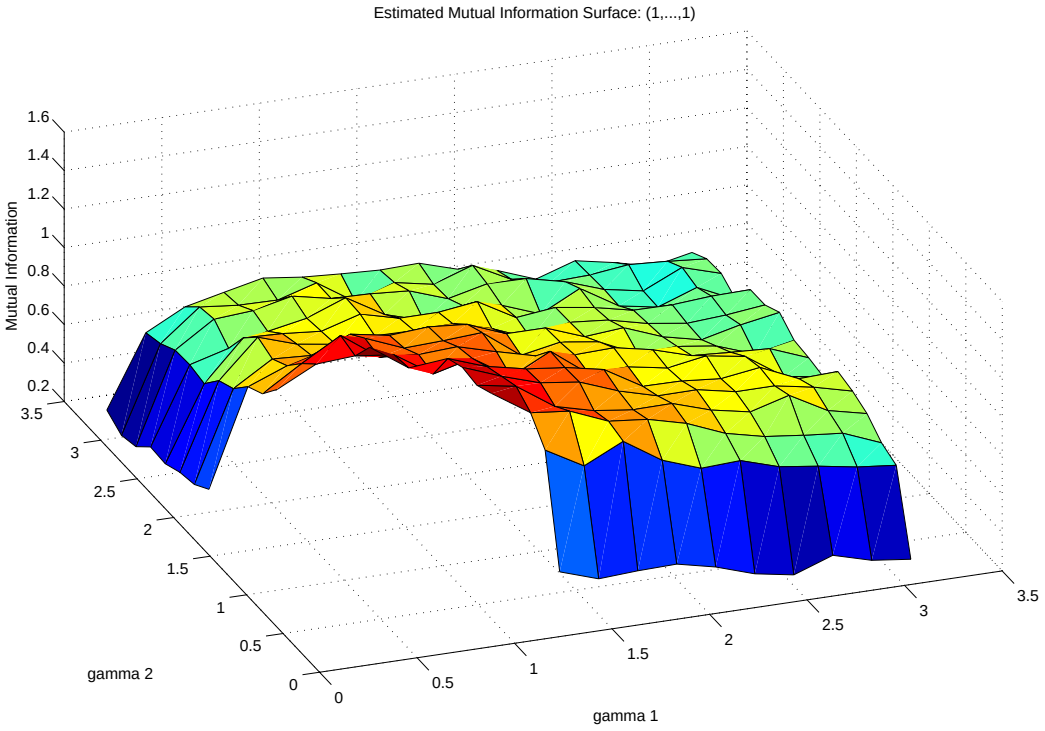
7. Estimate the Mutual Information as

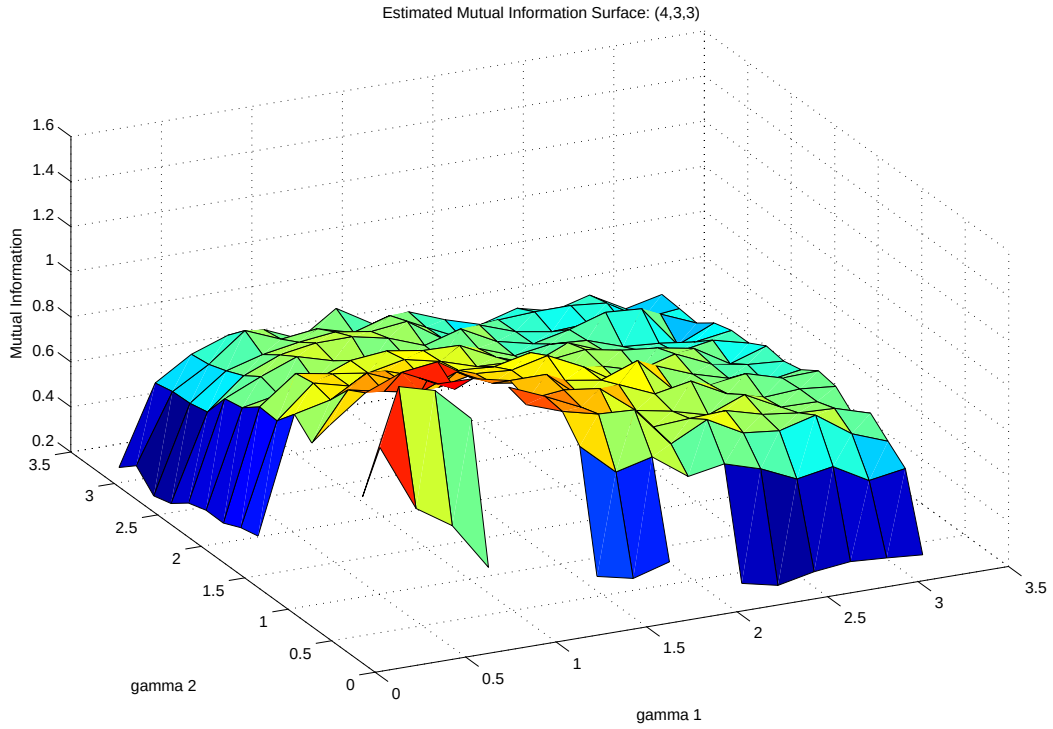
$$\hat{I}((\alpha, \beta)), \mathbf{X} | \gamma_1, \gamma_2) = \frac{1}{I} \sum_{h=1}^{I} \log \frac{f(s_1^{(h)}, \ldots, s_k^{(h)} | \alpha_h, \beta_h)}{f(s_1^{(h)}, \ldots, s_k^{(h)} | \gamma_1, \gamma_2)}$$

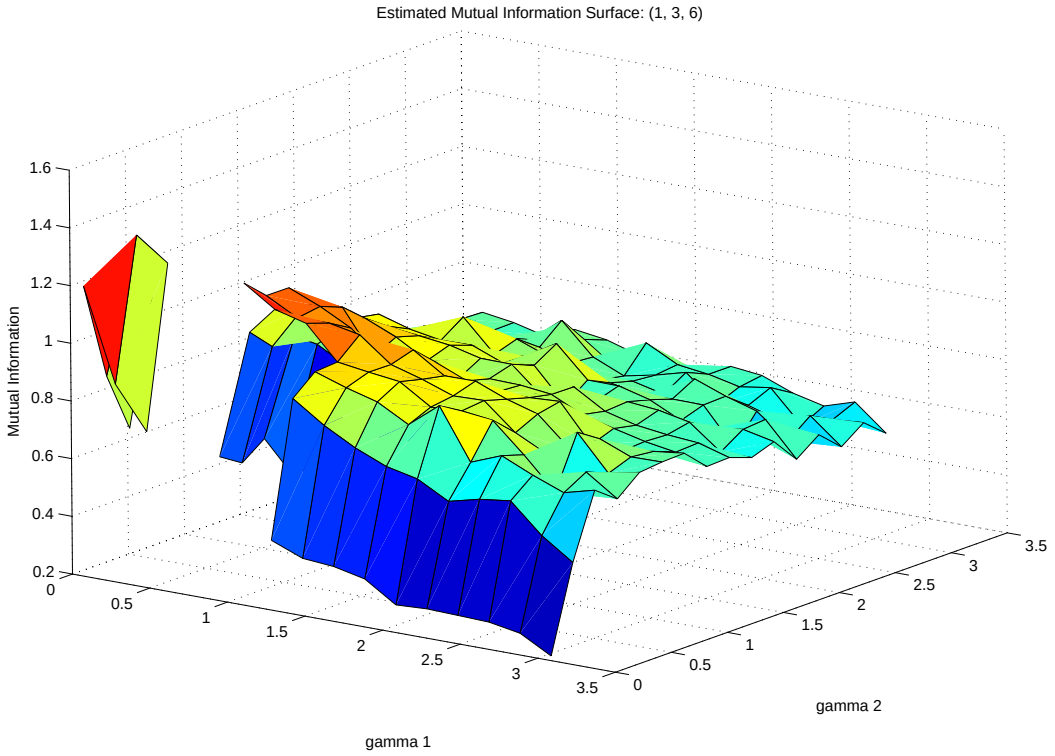
The Surfaces

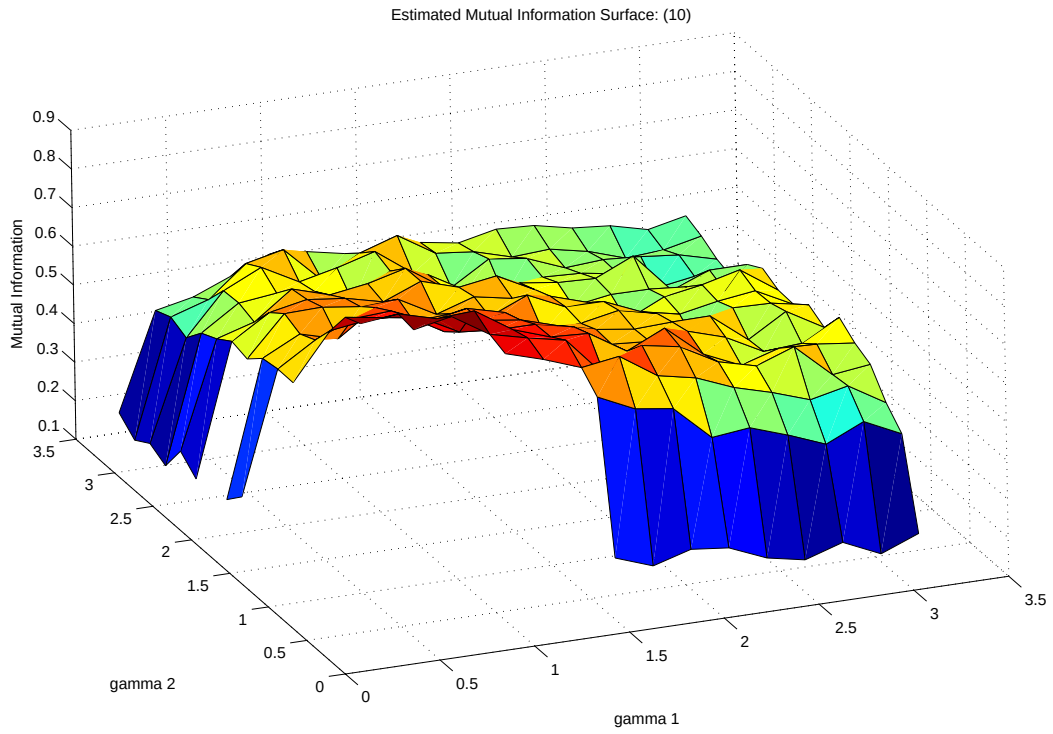
Some final considerations:

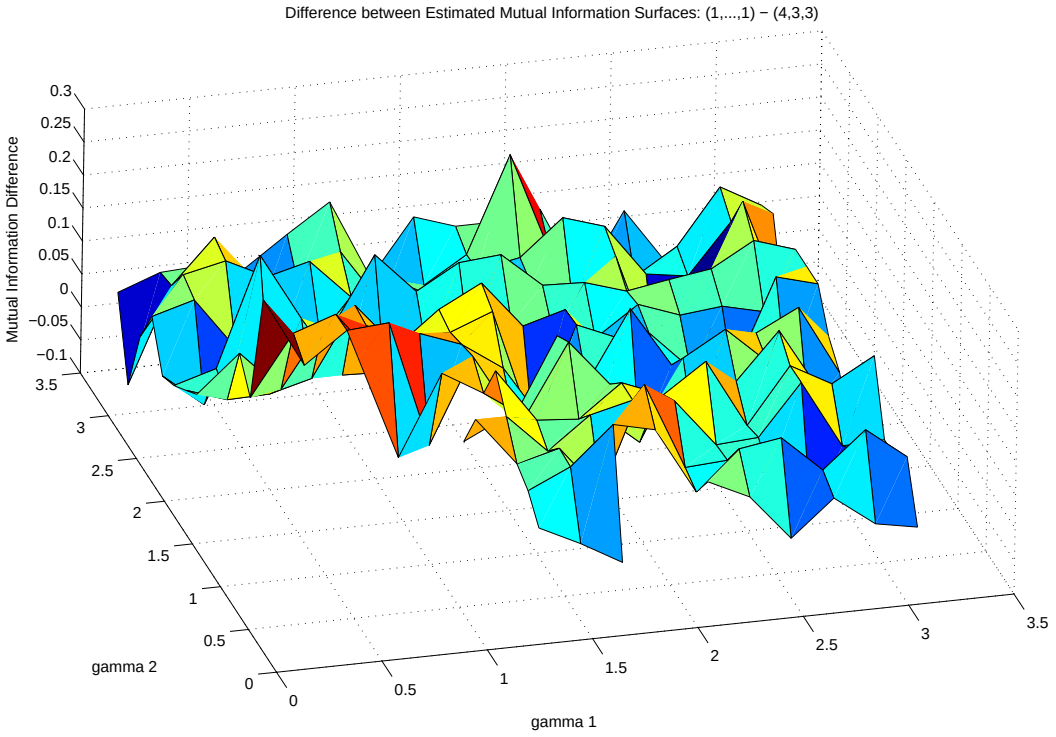
- The algorithm encounters some problems in proximity of the origin, where it probably needs to evaluate very high ratio's: this leads to missing estimates for some grid points which can be partially reconstructed via smoothing.

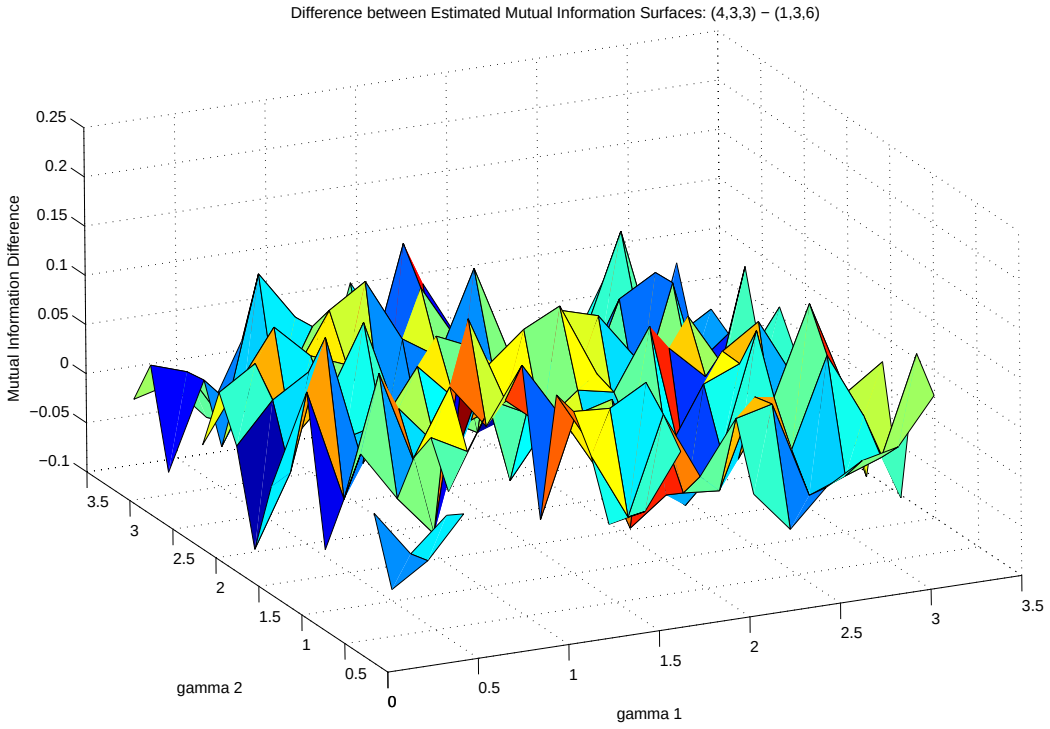

Estimated Mutual Information Surface: (1,...,1)
Mutual Information
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5

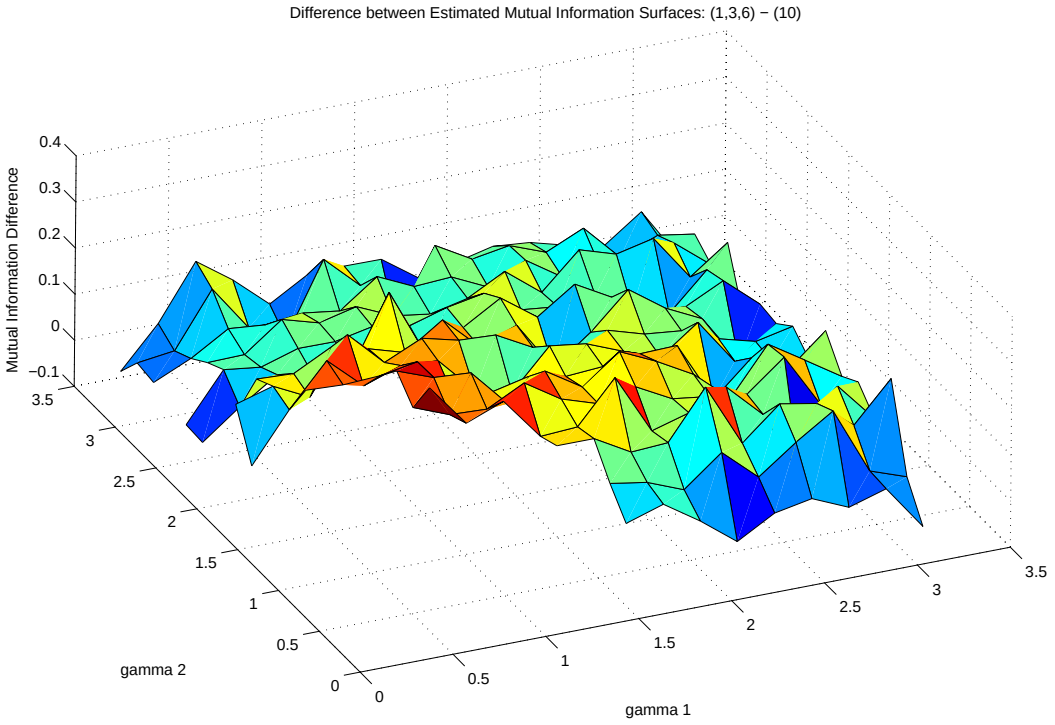

Estimated Mutual Information Surface: (4,3,3)
Mutual Information
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5


Estimated Mutual Information Surface: (1, 3, 6)
Mutual Information
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0.5
1
1.5
2
2.5
3
3.5
gamma 1
gamma 2


Estimated Mutual Information Surface: (10)
Mutual Information
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5


Difference between Estimated Mutual Information Surfaces: (1,...,1) − (4,3,3)
Mutual Information Difference
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
gamma 2
0
0.5
1
1.5
2
2.5
3
3.5
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5


Difference between Estimated Mutual Information Surfaces: (4,3,3) − (1,3,6)
Mutual Information Difference
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
gamma 2
3.5
3
2.5
2
1.5
1
0.5
0
gamma 1
0
0.5
1
1.5
2
2.5
3
3.5


Difference between Estimated Mutual Information Surfaces: (1,3,6) − (10)
Mutual Information Difference
0.4
0.3
0.2
0.1
0
−0.1
3.5
3
2.5
2
1.5
1
0.5
0
gamma 2
0
0.5
1
1.5
2
2.5
3
3.5
gamma 1

- Theoretical surfaces are in general continuous functions of the parameters: with this respect, the estimated surfaces do not seem to exhibit an eccessive irregularity.

- In comparisons between surfaces, it would be pleasant to have confidence bounds on the surface values to evaluate how much significantly positive or negative is a difference: clearly the standard deviation of the Monte Carlo replications can be used but a better estimate would be significant.

Bibliography

[1] Ali S.M. and Silvey S.D. (1965). Association Between Random Variables and the Dispersion of a Radon-Nikodym Derivative. *J.R.Statist.Soc.* **27** 100-107.

[2] Ali S.M. and Silvey S.D. (1966). A General Class of Coefficients of Divergence of One Distribution from Another. *J.R.Statist.Soc.* **28** 131-142.

[3] Balatoni J. and Renyi A. (1956). On the Notion of Entropy. (English translation) *MTA Matematikai Kutato Intezetenek Kozlemenyei.* **1** 9-40.

[4] Bernardo, J.M. (1979). Expected Information as Expected Utility. *Ann.Statist.* **7** 686-690.

[5] Blackwell, D. (1951). Comparison of Experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics*. 93-102.

[6] Blackwell, D. (1953). Equivalent Comparison of Experiments. *Ann.Math.Statist.* **24** 265-272.

[7] Bohnenblust, H.F., Shapley, L.S. and Sherman, S. (1949). Reconnaissance in Game Theory, Rand Research Memorandum **RM-208**.

[8] Chaloner, K. and Verdinelli, I. (1995). Bayesian Experimental Design: A Review, *Statis.Sci.* **10** 273-304.

[9] Clark, J.S. and Gelfand, A.E. (2006). *Hierarchical Modelling for the Environmental Sciences.* Oxford University Press, New York.

[10] Colin, B. (1989). Information Mutuelle Generalisee et Modeles Baysiens Hierarchiques. Pub.Inst.Stat.Univ.Paris **XXXIV** 67-96.

[11] Congdon, P. (2006). *Bayesian Statistical Modelling*. Wiley, New York.

[12] A Bayesian Method for Combining Results from Several Binomial Experiments. *J.Amer.Statist.Assoc.* **90** 935-944.

[13] Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. Wiley, New York.

[14] Csiszar, I. (1967). Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Studia Scientiarum Mathematicarum Hungarica* **2** 299-318.

[15] DeGroot, M.H. (1962). Uncertainty, Information, and Sequential Experiments. *Ann.Math.Statist.* **33** 404-419.

[16] DeGroot, M.H. (1966). Optimal Allocation of Observation. *Ann.Inst.Math.Statist.*. 13-28.

[17] DeGroot, M.H. (1984). Changes in Utility as Information. *Theory and Decision* **17** 287-303.

[18] Gelfand I.M. and Yaglom A.M. (1959) Calculation of the Amount of Information About a Random Variable Contained in Another Such Fucntion. *American Mathematical Society Translations* **12** 199-245.

[19] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. Chapman & Hall, New York.

[20] Gavurin, M.K. (1963). On the Value of Information. *Vestnik Leningrad University Series* **4** 27-34; translation (1968). *Selected Translations in Mathematical Statistics and Probabilty* **7** 193-202.

[21] Ginebra, J. (2007). On the Measure of the Information in a Statistical Experiment. *Bayesian Analysis* **2** 167-212.

[22] Goel, P.K. (1983). Information Measures and Bayesian Hierarchical Models. *J.Amer.Statist.Assoc.* **78** 408-410.

[23] Goel, P.K. and DeGroot, M.H. (1979). Comparison of Experiments and Information Measures. *Ann.Statist.* **7** 1066-1077.

[24] Goel, P.K. and DeGroot, M.H. (1981). Information About Hyperparameters in Hierarchical Models. *J.Amer.Statist.Assoc.* **73** 140-147.

[25] Goel, P.K. and Ginebra, J. (2003). When Is One Experiment 'Always Better than' Another? *The Statist.* **4** 515-537.

[26] Good, I.J. (1952). Rational Decisions. *J.R.Statist.Soc. B* **14** 107-114.

[27] Good, I.J. (1953). The population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* **40** 237-264.

[28] Good, I.J. (1980). Some History of the Hierarchical Bayesian Methodology. In *Proceedings of the International Meeting on Bayesian Statistics*. Uni. Valencia, 489-519.

[29] Hartigan J.A. (1990) Partition Models. *Comm. Statist. - Theory Meth.* **19** 2745-2756.

[30] Hartley, R.V. (1928). Transmission of Information. *Bell.System Tech.J.* **7** 535-563.

[31] Haussler, D. and Opper, M. (1997). Mutual Information, Metric Entropy and Cumulative Relative Entropy Risk. *Ann.Statist.* **25** 2451-2482.

[32] Jaynes E.T. (1957) Information Theory and Statistical Mechanics I. *Physical Review.* **106** 620-630.

[33] Ibragimov I. and Hasminskii R. (1972) On the Information in a Sample about a Parameter. In *Second International Symposium on Information Theory* IEEE, New York. 295-309.

[34] Jaynes E.T. (1957) Information Theory and Statistical Mechanics II. *Physical Review.* **108** 171-190.

[35] Jaynes E.T. (2003) *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

[36] Khinchin, A.I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications, New York.

[37] Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications, New York.

[38] LeCam, L. (1964). Sufficiency and Approximate Sufficiency. *Ann.Math.Statist.* **35** 1419-1455.

[39] LeCam, L. (1996). Comparison of Experiments - A Short Review. *Statistics, Probability and Game Theory* IMS Lecture Notes - Monograph Series. Vol. **30** 127-138.

[40] Lehmann, E.L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

[41] Lindley, D.V. (1956). On a Measure of the Information Provided by an Experiment. *Ann.Math.Stat.* **27** 986-1005.

[42] Lindley, D.V. (1957). Binomial Sampling Schemes and the Concept of Information. *Biometrika* **44** 179-186.

[43] Lindley, D.V. and Smith, A.F.M. (1972). Bayes Estimates for the Linear Model (with Discussion). *J.R.Statist.Soc. B* **34** 1-41.

[44] Lindqvist, B. (1977). How Fast Does a Markov Chain Forget the Initial State? A Decision Theoretical Approach. *Scand.J.Statist.* **4** 145-152.

[45] Lindqvist, B. (1978). On the Loss of Information Incurred by Lumping States of a MArkov Chain. *Scand.J.Statist.* **5** 92-98.

[46] Maled D. and Sedransk J. (1992) Bayesian Methodology for Combining the Results from Different Experiments when the Specifications for Pooling are Uncertain. *Biometrika* **79** 593-601.

[47] Parmigiani, G. and Berry D.A. (1994). Applications of Lindley's Information Measure to the Design of Clinical Experiments. In *Aspects of Uncertainty* edited by Smith, A.F.M. and Freeman, P.R., Wiley, New York.

[48] Quintana F.A. and Iglesias P.L. (2003). Bayesian Clustering and Product Partition Models. *J.R.Statist.Soc. B* **65** 557-574.

[49] Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory.* M.I.T. Press, Cambridge.

[50] Renyi, A. (1959). On the Dimension and Entropy of Probability Distributions. *Acta Mathematica Academiae Scientiarum Hungaricae.* **10** 193-215.

[51] Renyi, A. (1960). Dimension, Entropy and Information. *Transactions of the Prague Conference on Information Theory, Statistical Decision Theory functions, Random Processes.* 545-556.

[52] Renyi, A. (1965). On the Foundations of Information Theory. *Rev.Inst.Internat.Stat.* **33** 1-14.

[53] Renyi, A. (1984). *A Diary on Information Theory*. Akademiai Kiado, Budapest.

[54] Reza, F.M. (1961). *An Introduction to Information Theory*. Dover Publications, New York.

[55] Robert, C.P. (2001). *The Bayesian Choice* 2nd Ed. Springer-Verlag, New York.

[56] Rossi, P.E., Allenby G. and McCulloch R. (2005). *Bayesian Statistics and Marketing* Wiley, New York.

[57] Comparison of Experiments via Dependence of Normal Variables with a Common Marginal Distribution. (1992). *Ann.Statist.* **20** 614-618.

[58] Comparison of Experiments of Some Multivariate Distributions with a Common Marginal. (1993). In *Stochastic Inequalities*. IMS Lecture Notes - Monograph Series **22** 388-398.

[59] Shannon, C.E. (1948). A mathematical Theory of Communication. Bell.Syst.Tech.Journ. **27** 379-423; 623-656.

[60] Silvey, S.D. (1964). On a Measure of Association. *Ann.Math.Statist.* **35** 1157-1166.

[61] Smith, A.F.M. (1973). A General Bayesian Linear Model. *J.R.Statist.Soc. B* **35** 67-75.

[62] Soofi, E.S. (1994). Capturing the Intangible Concept of Information. *J.Amer.Statist.Assoc.* **89** 1243-1254.

[63] Stepniak, C. (1994). A Note on Comparison of Genetic Experiments. *Ann.Statist.* **22** 1630-1632.

[64] Torgersen, E. (1991). *Comparison of Statistical Experiments* in *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge.

[65] Yeung, R.W. (1991). A New Outlook on Shannon's Information Measures. *IEEE Trans.Inf.Th.* **37** 466-474.

[66] Wiener, N. (1948). *Cybernetics*. MIT Press, Cambridge MA and Wiley, New York.