Some Incoherencies Resulting from Minimal Information Methods

Haim Gaifman[1] Anubav Vasudevan[1]

[1]Department of Philosophy
Columbia University

FUPIR Conference, 2009

Outline

Overview

The technique of minimizing information (infomin) has been widely considered as a general method for both choosing and updating one's prior probabilities.

Advocates of the infomin methodology have made strong claims on its behalf, arguing that it can be justified on the basis of *a priori* principles of rationality (see, Shore & Johnson (1980), Csiszár (1991) and Paris & Vencovská (1997))

Overview

The technique of minimizing information (infomin) has been widely considered as a general method for both choosing and updating one's prior probabilities.

Advocates of the infomin methodology have made strong claims on its behalf, arguing that it can be justified on the basis of *a priori* principles of rationality (see, Shore & Johnson (1980), Csiszár (1991) and Paris & Vencovská (1997))

Overview

On the other hand, the method has been criticized by Bayesians, who have observed that the updating behavior recommended by infomin cannot be coherently represented as the result of conditionalization of a probability measure defined on an extended algebra in which the constraints themselves are events (see, Shimony (1985) and Seidenfeld (1986)).

We argue that, in a wide class of cases, infomin can be faulted on more fundamental grounds without presupposing a higher order Bayesian framework.

Overview

On the other hand, the method has been criticized by Bayesians, who have observed that the updating behavior recommended by infomin cannot be coherently represented as the result of conditionalization of a probability measure defined on an extended algebra in which the constraints themselves are events (see, Shimony (1985) and Seidenfeld (1986)).

We argue that, in a wide class of cases, infomin can be faulted on more fundamental grounds without presupposing a higher order Bayesian framework.

Preliminaries

A rational agent whose subjective degrees of belief are given by a probability function Pr defined over the Boolean algebra, $\mathcal{B}$, of all subsets of a (finite) set Ω.

New information in the form of a constraint C that should be satisfied by the agent's probability function.

An updating rule U, which maps Pr and C to an 'updated' probability, $U(Pr; C)$, which satisfies C.

Preliminaries

A rational agent whose subjective degrees of belief are given by a probability function Pr defined over the Boolean algebra, $\mathcal{B}$, of all subsets of a (finite) set Ω.

New information in the form of a constraint C that should be satisfied by the agent's probability function.

An updating rule U, which maps Pr and C to an 'updated' probability, $U(Pr; C)$, which satisfies C.

Preliminaries

A rational agent whose subjective degrees of belief are given by a probability function Pr defined over the Boolean algebra, $\mathcal{B}$, of all subsets of a (finite) set Ω.

New information in the form of a constraint C that should be satisfied by the agent's probability function.

An updating rule U, which maps Pr and C to an 'updated' probability, $U(Pr; C)$, which satisfies C.

Bayesian Conditionalization

The agent is informed of the truth of an event $A \in \mathcal{B}$, where $Pr(A) > 0$.

The constraint is $P(A) = 1$.

$U(Pr; P(A) = 1) = Pr(\ \ |A)$, the conditional probability function, given A.

Bayesian Conditionalization

The agent is informed of the truth of an event $A \in \mathcal{B}$, where $Pr(A) > 0$.

The constraint is $P(A) = 1$.

$U(Pr; P(A) = 1) = Pr(\ \ |A)$, the conditional probability function, given A.

Bayesian Conditionalization

The agent is informed of the truth of an event $A \in \mathcal{B}$, where $Pr(A) > 0$.

The constraint is $P(A) = 1$.

$U(Pr; P(A) = 1) = Pr(\ \ |A)$, the conditional probability function, given A.

Jeffrey Conditionalization

The agent is informed of the probabilities of a collection of pairwise disjoint events $A_1, A_2, \ldots, A_m \in \mathcal{B}$, where $Pr(A_i) > 0$ for $i = 1, 2, \ldots, m$.

The constraint is of the form $P(A_i) = \lambda_i$, for $i = 1, \ldots, m$, where $\{A_i\}_i$ is a partition of Ω and the λ_i's are non-negative reals which sum to 1.

The updated probability Pr^* is defined by:

$$Pr^*(A) = \sum_{i=1}^{m} Pr(A|A_i)\lambda_i \qquad (A \in \mathcal{B})$$

Jeffrey Conditionalization

The agent is informed of the probabilities of a collection of pairwise disjoint events $A_1, A_2, \ldots, A_m \in \mathcal{B}$, where $Pr(A_i) > 0$ for $i = 1, 2, \ldots, m$.

The constraint is of the form $P(A_i) = \lambda_i$, for $i = 1, \ldots, m$, where $\{A_i\}_i$ is a partition of Ω and the λ_i's are non-negative reals which sum to 1.

The updated probability Pr^* is defined by:

$$Pr^*(A) = \sum_{i=1}^{m} Pr(A|A_i)\lambda_i \qquad (A \in \mathcal{B})$$

Jeffrey Conditionalization

The agent is informed of the probabilities of a collection of pairwise disjoint events $A_1, A_2, \ldots, A_m \in \mathcal{B}$, where $Pr(A_i) > 0$ for $i = 1, 2, \ldots, m$.

The constraint is of the form $P(A_i) = \lambda_i$, for $i = 1, \ldots, m$, where $\{A_i\}_i$ is a partition of Ω and the λ_i's are non-negative reals which sum to 1.

The updated probability Pr^* is defined by:

$$Pr^*(A) = \sum_{i=1}^{m} Pr(A|A_i)\lambda_i \qquad (A \in \mathcal{B})$$

.

Linear Constraints

A linear constraint is any constraint of the form:

$$\sum_{i=1}^{m} a_i P(A_i) = b \quad (A_i \in \mathcal{B}),$$

where $a_1, a_2, \ldots, a_m, b$ are reals.

Question: Is there a general updating rule which applies to all linear constraints?

Linear Constraints

A linear constraint is any constraint of the form:

$$\sum_{i=1}^{m} a_i P(A_i) = b \quad (A_i \in \mathcal{B}),$$

where $a_1, a_2, \ldots, a_m, b$ are reals.

Question: Is there a general updating rule which applies to all linear constraints?

Infomin Updating

The agent's probabilities should be updated so as to minimize the amount of new 'information' gained in the process.

Given a linear constraint, C, update to the probability Pr^* satisfying C, which minimizes the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(Pr, Pr^*) = \sum_{\omega \in \Omega} Pr^*(\omega) \log \left(\frac{Pr^*(\omega)}{Pr(\omega)} \right)$$

Infomin updating generalizes both Jeffrey and Bayesian conditionalization.

Infomin Updating

The agent's probabilities should be updated so as to minimize the amount of new 'information' gained in the process.

Given a linear constraint, C, update to the probability Pr^* satisfying C, which minimizes the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(Pr, Pr^*) = \sum_{\omega \in \Omega} Pr^*(\omega) \log \left(\frac{Pr^*(\omega)}{Pr(\omega)} \right)$$

Infomin updating generalizes both Jeffrey and Bayesian conditionalization.

Infomin Updating

The agent's probabilities should be updated so as to minimize the amount of new 'information' gained in the process.

Given a linear constraint, C, update to the probability Pr^* satisfying C, which minimizes the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(Pr, Pr^*) = \sum_{\omega \in \Omega} Pr^*(\omega) \log \left(\frac{Pr^*(\omega)}{Pr(\omega)} \right)$$

Infomin updating generalizes both Jeffrey and Bayesian conditionalization.

Infomin Prior Selection

The agent's prior probability should be minimally informative.

Given a linear constraint, C, choose as a prior, the probability Pr^* satisfying C, possessing minimal Shannon information:

$$S(Pr^*) = \sum_{\omega \in \Omega} Pr^*(\omega) \log Pr^*(\omega)$$

Since $S(Pr^*) = D_{\mathrm{KL}}(Pr^0, Pr^*) + c$, where Pr^0 is the uniform prior, as far as the mathematics is concerned, one can treat infomin prior selection as the special case of infomin updating of Pr^0.

Infomin Prior Selection

The agent's prior probability should be minimally informative.

Given a linear constraint, C, choose as a prior, the probability Pr^* satisfying C, possessing minimal Shannon information:

$$S(Pr^*) = \sum_{\omega \in \Omega} Pr^*(\omega) \log Pr^*(\omega)$$

Since $S(Pr^*) = D_{\mathrm{KL}}(Pr^0, Pr^*) + c$, where Pr^0 is the uniform prior, as far as the mathematics is concerned, one can treat infomin prior selection as the special case of infomin updating of Pr^0.

Infomin Prior Selection

The agent's prior probability should be minimally informative.

Given a linear constraint, C, choose as a prior, the probability Pr^* satisfying C, possessing minimal Shannon information:

$$S(Pr^*) = \sum_{\omega \in \Omega} Pr^*(\omega) \log Pr^*(\omega)$$

Since $S(Pr^*) = D_{\mathrm{KL}}(Pr^0, Pr^*) + c$, where Pr^0 is the uniform prior, as far as the mathematics is concerned, one can treat infomin prior selection as the special case of infomin updating of Pr^0.

Outline

Families of Constraints

We consider families of linear constraints $\{C_\lambda\}_{\lambda\in\Lambda}$. We assume that U is non-vacuous with respect to $\{C_\lambda\}_{\lambda\in\Lambda}$, i.e., for some $\lambda \in \Lambda$, $Pr \neq U(Pr; C_\lambda)$.

Expected-Value constraints:

(E) $$E(X) = \sum_{\omega\in\Omega} X(\omega)\cdot P(\omega) = \lambda$$

Conditional-Probability Constraints:

(C) $$P(B|A) = \lambda$$

Families of Constraints

We consider families of linear constraints $\{C_\lambda\}_{\lambda\in\Lambda}$. We assume that U is non-vacuous with respect to $\{C_\lambda\}_{\lambda\in\Lambda}$, i.e., for some $\lambda \in \Lambda$, $Pr \neq U(Pr; C_\lambda)$.

Expected-Value constraints:

(E) $$E(X) = \sum_{\omega\in\Omega} X(\omega)\cdot P(\omega) = \lambda$$

Conditional-Probability Constraints:

(C) $$P(B|A) = \lambda$$

Families of Constraints

We consider families of linear constraints $\{C_\lambda\}_{\lambda\in\Lambda}$. We assume that U is non-vacuous with respect to $\{C_\lambda\}_{\lambda\in\Lambda}$, i.e., for some $\lambda \in \Lambda$, $Pr \neq U(Pr; C_\lambda)$.

Expected-Value constraints:

(E) $$E(X) = \sum_{\omega\in\Omega} X(\omega)\cdot P(\omega) = \lambda$$

Conditional-Probability Constraints:

(C) $$P(B|A) = \lambda$$

Always-Decreasing(Increasing) Events

Definition

An event $A \in \mathcal{B}$ is *always decreasing under U* with respect to the family $\{C_\lambda\}_{\lambda \in \Lambda}$ and the probability Pr, if for all $\lambda \in \Lambda$ either $U(Pr; C_\lambda) = Pr$ or $U(Pr; C_\lambda)(A) < Pr(A)$. *Always increasing* events are defined similarly.

Deceptiveness

Definition

For a given probability Pr, U is *deceptive over* A with respect to the family of constraints $\{C_\lambda\}_{\lambda\in\Lambda}$, if A is either always decreasing or always increasing under U. We call U *deceptive* if it is deceptive over some A.

What's wrong with Deceptiveness?

Consider a rational agent, Ann, whose current degrees of belief are given by the probability function Pr, and who is committed to updating her probabilities by means of U for all $\lambda \in \Lambda$. Suppose that A is always decreasing under U.

Ann is given a sealed envelope, which contains a report describing the true value of λ.

What's wrong with Deceptiveness?

Consider a rational agent, Ann, whose current degrees of belief are given by the probability function Pr, and who is committed to updating her probabilities by means of U for all $\lambda \in \Lambda$. Suppose that A is always decreasing under U.

Ann is given a sealed envelope, which contains a report describing the true value of λ.

What's wrong with Deceptiveness?

Does Ann admit the possibility that $\lambda \in \{\lambda' \in \Lambda : U(Pr; C_{\lambda'}) \neq Pr\}$?

- If so, then her current degree of belief in A must be $< Pr(A)$.
- If not, then there is no point in her opening the envelope, since she is already certain that the information it contains will only confirm what she already believes.

In this case 'updating' is a deceptive name.

What's wrong with Deceptiveness?

Does Ann admit the possibility that
$\lambda \in \{\lambda' \in \Lambda : U(Pr; C_{\lambda'}) \neq Pr\}$?

- If so, then her current degree of belief in A must be $< Pr(A)$.
- If not, then there is no point in her opening the envelope, since she is already certain that the information it contains will only confirm what she already believes.

In this case 'updating' is a deceptive name.

What's wrong with Deceptiveness?

Does Ann admit the possibility that $\lambda \in \{\lambda' \in \Lambda : U(Pr; C_{\lambda'}) \neq Pr\}$?

- If so, then her current degree of belief in A must be $< Pr(A)$.
- If not, then there is no point in her opening the envelope, since she is already certain that the information it contains will only confirm what she already believes.

In this case 'updating' is a deceptive name.

What's wrong with Deceptiveness?

Does Ann admit the possibility that $\lambda \in \{\lambda' \in \Lambda : U(Pr; C_{\lambda'}) \neq Pr\}$?

- If so, then her current degree of belief in A must be $< Pr(A)$.
- If not, then there is no point in her opening the envelope, since she is already certain that the information it contains will only confirm what she already believes.

In this case 'updating' is a deceptive name.

What's wrong with Deceptiveness?

This argument can also be made out in terms of betting odds. Ann is aware that her current commitments oblige her to accept the following two bets:

1. A bet staking \$1 on A at odds $(1 - Pr(A))/Pr(A) : 1$.
2. A future bet, made after receiving the information, staking $\$(1 - Pr(A))/Pr(A)$ against A at odds $Pr_\lambda(A)/(1 - Pr_\lambda(A)) : 1$.

Ann knows that in accepting these two bets, she cannot, under any circumstances, earn a positive return, and moreover, she will lose money if $\lambda \in \Lambda^-$ and A does not occur.

What's wrong with Deceptiveness?

This argument can also be made out in terms of betting odds. Ann is aware that her current commitments oblige her to accept the following two bets:

1. A bet staking \$1 on A at odds $(1 - Pr(A))/Pr(A) : 1$.
2. A future bet, made after receiving the information, staking $\$(1 - Pr(A))/Pr(A)$ against A at odds $Pr_\lambda(A)/(1 - Pr_\lambda(A)) : 1$.

Ann knows that in accepting these two bets, she cannot, under any circumstances, earn a positive return, and moreover, she will lose money if $\lambda \in \Lambda^-$ and A does not occur.

What's wrong with Deceptiveness?

This argument can also be made out in terms of betting odds. Ann is aware that her current commitments oblige her to accept the following two bets:

1. A bet staking \$1 on A at odds $(1 - Pr(A))/Pr(A) : 1$.
2. A future bet, made after receiving the information, staking $\$(1 - Pr(A))/Pr(A)$ against A at odds $Pr_\lambda(A)/(1 - Pr_\lambda(A)) : 1$.

Ann knows that in accepting these two bets, she cannot, under any circumstances, earn a positive return, and moreover, she will lose money if $\lambda \in \Lambda^-$ and A does not occur.

What does this Argument Assume?

- Ann is capable of recognizing that U is deceptive
- Ann is capable of acknowledging the possibility that $\lambda \in \Lambda^{<}$.

In particular, there is no need to assume that this latter acknowledgement is an expression of any more detailed estimate, on Ann's part, of how *likely* it is that the updating will be non-trivial.

What does this Argument Assume?

- Ann is capable of recognizing that U is deceptive
- Ann is capable of acknowledging the possibility that $\lambda \in \Lambda^{<}$.

In particular, there is no need to assume that this latter acknowledgement is an expression of any more detailed estimate, on Ann's part, of how *likely* it is that the updating will be non-trivial.

What does this Argument Assume?

- Ann is capable of recognizing that U is deceptive
- Ann is capable of acknowledging the possibility that $\lambda \in \Lambda^{<}$.

In particular, there is no need to assume that this latter acknowledgement is an expression of any more detailed estimate, on Ann's part, of how *likely* it is that the updating will be non-trivial.

Deceptiveness and Infomin: (E) Constraints

Theorem

Assume that Ω has n points ($n > 3$). Then, for every event $A \subseteq \Omega$, such that $2 \leq |A| \leq n-2$, there is a random variable, X, such that A is always decreasing under infomin updating of the uniform prior with respect to the family of constraints $\{E(X) = \lambda\}_\lambda$.

On the other hand, one can always find a random variable, X, such that infomin updating of the uniform prior on $E(X)$ is not deceptive.

Deceptiveness and Infomin: (E) Constraints

Theorem

Assume that Ω has n points ($n > 3$). Then, for every event $A \subseteq \Omega$, such that $2 \leq |A| \leq n-2$, there is a random variable, X, such that A is always decreasing under infomin updating of the uniform prior with respect to the family of constraints $\{E(X) = \lambda\}_\lambda$.

On the other hand, one can always find a random variable, X, such that infomin updating of the uniform prior on $E(X)$ is not deceptive.

Deceptiveness and Infomin: (C) Constraints

Theorem

Let $A, B \in \mathcal{B}$ be such that $A \neq \Omega$ and $A \cap B$ is a non-empty, proper subset of A. Then for any strictly positive prior Pr, A is always decreasing under infomin updating of Pr with respect to the family $\{P(B|A) = \lambda\}_\lambda$.

Outline

Choosing a Prior

Consider an agent Abe, who, like Ann, is about to learn the true value of λ. Abe does not yet have probabilities over $\mathcal{B}$, but is committed to choosing as his prior $U(Pr_0; C_\lambda)$, for all $\lambda \in \Lambda$.

Suppose that A is always decreasing under U with respect to $\{C_\lambda\}_{\lambda \in \Lambda}$ and Pr_0.

Abe is in a position to infer that, regardless of the value of λ, the prior he will choose will satisfy the inequality $U(Pr_0; C_\lambda)(A) \leq |A|/n$, where $n = |\Omega|$.

Choosing a Prior

Consider an agent Abe, who, like Ann, is about to learn the true value of λ. Abe does not yet have probabilities over $\mathcal{B}$, but is committed to choosing as his prior $U(Pr_0; C_\lambda)$, for all $\lambda \in \Lambda$.

Suppose that A is always decreasing under U with respect to $\{C_\lambda\}_{\lambda \in \Lambda}$ and Pr_0.

Abe is in a position to infer that, regardless of the value of λ, the prior he will choose will satisfy the inequality $U(Pr_0; C_\lambda)(A) \leq |A|/n$, where $n = |\Omega|$.

Choosing a Prior

Consider an agent Abe, who, like Ann, is about to learn the true value of λ. Abe does not yet have probabilities over $\mathcal{B}$, but is committed to choosing as his prior $U(Pr_0; C_\lambda)$, for all $\lambda \in \Lambda$.

Suppose that A is always decreasing under U with respect to $\{C_\lambda\}_{\lambda \in \Lambda}$ and Pr_0.

Abe is in a position to infer that, regardless of the value of λ, the prior he will choose will satisfy the inequality $U(Pr_0; C_\lambda)(A) \leq |A|/n$, where $n = |\Omega|$.

Choosing a Prior

Imagine a bag containing a large number of apples and pears of two kinds: expensive and inexpensive. Let A be the event that the next drawn object will be an apple and let B be the event that it will be an expensive apple.

Abe is about to learn the value of $\lambda = P(B|A)$. This means that the relative frequency of expensive apples among apples is λ.

If A is always decreasing under U with respect to Pr_0 and the family of constraints $\{P(B|A) = \lambda\}_\lambda$, then Abe is already in a position to know that the prior he chooses will assign to the event A a probability $\leq 1/2$!

Choosing a Prior

Imagine a bag containing a large number of apples and pears of two kinds: expensive and inexpensive. Let A be the event that the next drawn object will be an apple and let B be the event that it will be an expensive apple.

Abe is about to learn the value of $\lambda = P(B|A)$. This means that the relative frequency of expensive apples among apples is λ.

If A is always decreasing under U with respect to Pr_0 and the family of constraints $\{P(B|A) = \lambda\}_\lambda$, then Abe is already in a position to know that the prior he chooses will assign to the event A a probability $\leq 1/2$!

Choosing a Prior

Imagine a bag containing a large number of apples and pears of two kinds: expensive and inexpensive. Let A be the event that the next drawn object will be an apple and let B be the event that it will be an expensive apple.

Abe is about to learn the value of $\lambda = P(B|A)$. This means that the relative frequency of expensive apples among apples is λ.

If A is always decreasing under U with respect to Pr_0 and the family of constraints $\{P(B|A) = \lambda\}_\lambda$, then Abe is already in a position to know that the prior he chooses will assign to the event A a probability $\leq 1/2$!

Choosing a Prior

This is strange! Apparently, the mere knowledge that Abe will be informed of the value of λ implies a substantial bound on how likely it is that a given event will occur.

There are circumstances in which mere knowability can have substantial implications. For example, it may show that what can be known is not top secret, or that it lacks significance, in some sense of the word.

Choosing a Prior

This is strange! Apparently, the mere knowledge that Abe will be informed of the value of λ implies a substantial bound on how likely it is that a given event will occur.

There are circumstances in which mere knowability can have substantial implications. For example, it may show that what can be known is not top secret, or that it lacks significance, in some sense of the word.

Choosing a Prior

This is the logic which underlies. A commitment to infomin licenses an agent to assume that whatever new information may come to light, it is of minimal significance.

It is this assumption which leads to the always decreasing phenomenon with respect to infomin updating on both (E) and (C) constraints

Choosing a Prior

This is the logic which underlies. A commitment to infomin licenses an agent to assume that whatever new information may come to light, it is of minimal significance.

It is this assumption which leads to the always decreasing phenomenon with respect to infomin updating on both (E) and (C) constraints

Shiftiness

Definition

We call an updating method *(E)-shifty* (resp: *(C)-shifty*) if there are two families of (E) constraints (resp: (C) constraints), such that for some event $A \in \mathcal{B}$, A is always decreasing with respect to one family and always increasing with respect to the other.

The above theorems entail that for $|\Omega| \geq 4$, infomin prior selection is both (C)-shifty and (E)-shifty.

Shiftiness

Definition

We call an updating method *(E)-shifty* (resp: *(C)-shifty*) if there are two families of (E) constraints (resp: (C) constraints), such that for some event $A \in \mathcal{B}$, A is always decreasing with respect to one family and always increasing with respect to the other.

The above theorems entail that for $|\Omega| \geq 4$, infomin prior selection is both (C)-shifty and (E)-shifty.

Outline

Higher Order Probabilities

Let $\Omega = \{\omega_1, ..., \omega_n\}$. Then a probability function over $\mathcal{B}$ can be represented by a vector, $\overline{p} = (p_1, ..., p_n)$, where p_i is the probability of ω_i for $i = 1, \ldots, n$.

The space of all probabilities on $\mathcal{B}$ is given by the $n-1$ dimensional simplex:

$$\Delta = \{\overline{p} \in \mathbb{R}^n : \Sigma_{i=1}^{n} p_i = 1, p_i \geq 0, \text{ for } i = 1, \ldots, n\}$$

A *higher order probability* is a probability measure m defined over the σ-field of Borel subsets of Δ.

Higher Order Probabilities

Let $\Omega = \{\omega_1, ..., \omega_n\}$. Then a probability function over $\mathcal{B}$ can be represented by a vector, $\overline{p} = (p_1, ..., p_n)$, where p_i is the probability of ω_i for $i = 1, \ldots, n$.

The space of all probabilities on $\mathcal{B}$ is given by the $n-1$ dimensional simplex:

$$\Delta = \{\overline{p} \in \mathbb{R}^n : \Sigma_{i=1}^{n} p_i = 1, p_i \geq 0, \text{ for } i = 1, \ldots, n\}$$

A *higher order probability* is a probability measure m defined over the σ-field of Borel subsets of Δ.

Higher Order Probabilities

Let $\Omega = \{\omega_1, ..., \omega_n\}$. Then a probability function over $\mathcal{B}$ can be represented by a vector, $\overline{p} = (p_1, ..., p_n)$, where p_i is the probability of ω_i for $i = 1, \ldots, n$.

The space of all probabilities on $\mathcal{B}$ is given by the $n-1$ dimensional simplex:

$$\Delta = \{\overline{p} \in \mathbb{R}^n : \Sigma_{i=1}^{n} p_i = 1, p_i \geq 0, \text{ for } i = 1, \ldots, n\}$$

A *higher order probability* is a probability measure m defined over the σ-field of Borel subsets of Δ.

Higher Order Support

A higher order probability m *supports* an updating method U, if, for all $\lambda \in \Lambda$:

$$\begin{array}{ccc} m & \longrightarrow & m(\quad \|C_\lambda\|) \\ \downarrow & & \downarrow \\ Pr & \longrightarrow & U(Pr; C_\lambda) \end{array}$$

$P_{\overline{p}}$ = the probability function determined by $\overline{p}$.

$$\|C_\lambda\| =_{\mathrm{df}} \{\overline{p} \in \Delta : P_{\overline{p}} \text{ satisfies } C_\lambda\}$$

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Assumptions

For every $\Theta \subseteq \Lambda$, let $\|C_\Theta\| = \bigcup_{\theta \in \Theta} \|C_\theta\|$.

(1) For every $\overline{p} \in \|C_\Lambda\|$, there is a unique $\lambda \in \Lambda$ such that $\overline{p} \in \|C_\lambda\|$ and the function, f, which maps $\overline{p}$ to λ is continuous in $\overline{p}$.

(2) f maps $\|C_\Lambda\|$ onto Λ.

(3) $m(\|C_\Lambda\|) = 1$.

(4) If Θ is an open non-empty subset of Λ, then $m(\|C_\Theta\|) > 0$.

(5) For each Pr, if $U(Pr; C_\lambda) = P_{\overline{p}(\lambda)}$, then $\overline{p}(\lambda)$ is a continuous function of λ.

Higher Order Support

Definition

Let U be an updating method, let Pr be a probability on $\mathcal{B}$ and let $Pr_\lambda = U(Pr; C_\lambda)$, for all $\lambda \in \Lambda$. Then U is *supported* on Pr by the higher order probability m, if the following conditions hold:

(i) $$Pr = \int_\Delta Pr_{\overline{p}} dm(\overline{p})$$

and, for every $\Theta \subseteq \Lambda$ such that $\mu(\Theta) > 0$,

(ii) $$\frac{1}{\mu(\Theta)} \int_\Theta Pr_\lambda d\mu(\lambda) = \frac{1}{m(\|C_\Theta\|)} \int_{\|C_\Theta\|} P_{\overline{p}} dm(\overline{p})$$

Higher Order Support and Non-Deceptiveness

Theorem

If U is supported on Pr by m, then U is not deceptive with respect to Pr and the constraint family $\{C_\lambda\}_{\lambda\in\Lambda}$.

As we shall see, the converse of this theorem does not hold.

Higher Order Support and Non-Deceptiveness

Theorem

If U is supported on Pr by m, then U is not deceptive with respect to Pr and the constraint family $\{C_\lambda\}_{\lambda\in\Lambda}$.

As we shall see, the converse of this theorem does not hold.

Outline

Deceptiveness and Infomin: (C) Constraints

Theorem

Let $A, B \in \mathcal{B}$ be such that $A \neq \Omega$ and $A \cap B$ is a non-empty, proper subset of A. Then for any strictly positive prior Pr, A is always decreasing under infomin updating of Pr with respect to the family $\{P(B|A) = \lambda\}_\lambda$.

Deceptiveness and Infomin: (C) Constraints

We will assume that D is an arbitrary information measure, that is, a binary function that maps every pair (P, P^*) of probability functions to a real number $D(P, P^*)$, representing the 'cost' (in information) of updating from P to P^*.

We will state conditions on D that are sufficient to ensure deceptiveness of updating on (C) constraints, and which are satisfied by the KL divergence.

Deceptiveness and Infomin: (C) Constraints

We will assume that D is an arbitrary information measure, that is, a binary function that maps every pair (P, P^*) of probability functions to a real number $D(P, P^*)$, representing the 'cost' (in information) of updating from P to P^*.

We will state conditions on D that are sufficient to ensure deceptiveness of updating on (C) constraints, and which are satisfied by the KL divergence.

Deceptiveness and Infomin: (C) Constraints

We only consider strictly positive priors, i.e., probabilities in the set:

$$\Delta^0 = \{P \in \Delta : P(\omega) > 0, \text{ for all } \omega \in \Omega\}$$

(1) $D(P, P^*) \geq 0$, with equality holding iff $P = P^*$

(2) For every $P \in \Delta^0$ and for all $\lambda \in [0, 1]$, there exists a unique P^* which minimizes $D(P, P^*)$ under the constraint $P(B|A) = \lambda$, where P^* belongs to the set:

$$\Delta_A = \{P \in \Delta : 0 < P(A) < 1\}$$

Deceptiveness and Infomin: (C) Constraints

We only consider strictly positive priors, i.e., probabilities in the set:

$$\Delta^0 = \{P \in \Delta : P(\omega) > 0, \text{ for all } \omega \in \Omega\}$$

(1) $D(P, P^*) \geq 0$, with equality holding iff $P = P^*$

(2) For every $P \in \Delta^0$ and for all $\lambda \in [0, 1]$, there exists a unique P^* which minimizes $D(P, P^*)$ under the constraint $P(B|A) = \lambda$, where P^* belongs to the set:

$$\Delta_A = \{P \in \Delta : 0 < P(A) < 1\}$$

Deceptiveness and Infomin: (C) Constraints

We only consider strictly positive priors, i.e., probabilities in the set:

$$\Delta^0 = \{P \in \Delta : P(\omega) > 0, \text{ for all } \omega \in \Omega\}$$

(1) $D(P, P^*) \geq 0$, with equality holding iff $P = P^*$

(2) For every $P \in \Delta^0$ and for all $\lambda \in [0, 1]$, there exists a unique P^* which minimizes $D(P, P^*)$ under the constraint $P(B|A) = \lambda$, where P^* belongs to the set:

$$\Delta_A = \{P \in \Delta : 0 < P(A) < 1\}$$

Deceptiveness and Infomin: (C) Constraints

Let Pr be any probability function in Δ^0. We write $Pr|A$ for the conditional probability function $Pr(-|A)$ restricted to the algebra of subsets of A.

Any change in Pr can be fully described in terms of the following three sorts of alterations:

1. Changes in the value of $Pr(A)$.
2. Changes to the function $Pr|A$.
3. Changes to the function $Pr|\overline{A}$.

Deceptiveness and Infomin: (C) Constraints

Let Pr be any probability function in Δ^0. We write $Pr|A$ for the conditional probability function $Pr(-|A)$ restricted to the algebra of subsets of A.

Any change in Pr can be fully described in terms of the following three sorts of alterations:

1. Changes in the value of $Pr(A)$.
2. Changes to the function $Pr|A$.
3. Changes to the function $Pr|\overline{A}$.

Unimodality

Unimodality(UNI): Let D' be the restriction of D to the set $\{(P, P^*) \in \Delta^0 \times \Delta_A : P|A = P^*|A,\ P|\overline{A} = P^*|\overline{A}\}$. Then D' is independent of both $P|A$ and $P|\overline{A}$, and it is strictly decreasing as $Pr^*(A)$ approaches $Pr(A)$ from either the left or the right. Moreover, D' is differentiable with respect to $P^*(A)$ on the interval $(0, 1)$.

Conditional Monotonicity

Conditional Monotonicity (CM) Let D'' be the restriction of D to the set $\{(P, P^*) \in \Delta^0 \times \Delta_A : P(A) = P^*(A), P|\overline{A} = P^*|\overline{A}\}$. Then D'' is independent of $P|\overline{A}$. Moreover, if $P^*|A \neq P|A$, then D'' is differentiable with respect to $P^*(A)$ on the interval $(0, 1)$, and its derivative on this interval is bounded below by some positive number (which depends on $P^*|A$ and $P|A$).

How do the costs combine?

An updating of P to P^* can clearly be decomposed into the following three updating acts:

- $P \rightarrow P' = (P^*(A), P|A, P|\overline{A})$
- $P' \rightarrow P'' = (P^*(A), P^*|A, P|\overline{A})$
- $P'' \rightarrow P^*$.

Suppose

$$D(P, P^*) = F(P(A), P^*(A), D(P, P'), D(P', P''), D(P'', P^*))$$

Question: What is the form of F?

How do the costs combine?

An updating of P to P^* can clearly be decomposed into the following three updating acts:

- $P \rightarrow P' = (P^*(A), P|A, P|\overline{A})$
- $P' \rightarrow P'' = (P^*(A), P^*|A, P|\overline{A})$
- $P'' \rightarrow P^*$.

Suppose

$$D(P, P^*) = F(P(A), P^*(A), D(P, P'), D(P', P''), D(P'', P^*))$$

Question: What is the form of F?

How do the costs combine?

An updating of P to P^* can clearly be decomposed into the following three updating acts:

- $P \rightarrow P' = (P^*(A), P|A, P|\overline{A})$
- $P' \rightarrow P'' = (P^*(A), P^*|A, P|\overline{A})$
- $P'' \rightarrow P^*$.

Suppose

$$D(P, P^*) = F(P(A), P^*(A), D(P, P'), D(P', P''), D(P'', P^*))$$

Question: What is the form of F?

How do the costs combine?

If $D = D_{KL}$, then the updating rule is simple:

(ADD) $$F(s, t, u, v, w) = u + v + w$$

As it turns out, (ADD) (w/ (UNI)+(CM)) is sufficient to establish deceptiveness, but a much weaker condition than (ADD) is all that is needed.

How do the costs combine?

If $D = D_{KL}$, then the updating rule is simple:

(ADD) $$F(s, t, u, v, w) = u + v + w$$

As it turns out, (ADD) (w/ (UNI)+(CM)) is sufficient to establish deceptiveness, but a much weaker condition than (ADD) is all that is needed.

Cost Combination

Cost Combination (CCO)

(i) F is strictly increasing in each of u, v, w.

(ii) F has a total differential.

(iii) Let $f(t, u, v) = F(t, u, v, 0)$, then:

(iii.1) $\partial f / \partial t \geq 0$.

(iii.2) For some $c > 0$, $\partial f / \partial v > c$.

Another example

Here's another information measure that satisfies (CCO):

$$D^{\dagger}(Pr, Pr^*) = \left(\sum_{\omega \in \Omega} Pr^*(\omega) \left(\frac{Pr^*(\omega)}{Pr(\omega)} \right) \right) - 1$$

The combination rule for $D^{\dagger}$ is given by:

$$F(s, t, u, v, w) = u + \left(\frac{t}{s} \right) v + \left(\frac{1-t}{1-s} \right) w$$

$D^{\dagger}$ satisfies (UNI), (CM) and (CCO).

Another example

Here's another information measure that satisfies (CCO):

$$D^{\dagger}(Pr, Pr^*) = \left(\sum_{\omega \in \Omega} Pr^*(\omega) \left(\frac{Pr^*(\omega)}{Pr(\omega)} \right) \right) - 1$$

The combination rule for $D^{\dagger}$ is given by:

$$F(s, t, u, v, w) = u + \left(\frac{t}{s} \right) v + \left(\frac{1-t}{1-s} \right) w$$

$D^{\dagger}$ satisfies (UNI), (CM) and (CCO).

Another example

Here's another information measure that satisfies (CCO):

$$D^{\dagger}(Pr, Pr^*) = \left(\sum_{\omega \in \Omega} Pr^*(\omega) \left(\frac{Pr^*(\omega)}{Pr(\omega)} \right) \right) - 1$$

The combination rule for $D^{\dagger}$ is given by:

$$F(s, t, u, v, w) = u + \left(\frac{t}{s} \right) v + \left(\frac{1-t}{1-s} \right) w$$

$D^{\dagger}$ satisfies (UNI), (CM) and (CCO).

Theorem

Let D be any information measure satisfying (UNI), (CM) and (CCO), and let Pr be any prior probability function in Δ^0. If Pr_λ is the probability function which minimizes $D(Pr, Pr^*)$ under the constraint $Pr^*(B|A) = \lambda$, then:

$$Pr_\lambda(A) \leq Pr(A)$$

and the inequality is strict if $Pr_\lambda \neq Pr$.

Outline

Expected-Value Constraints

Question: For which events, $A \in \mathcal{B}$, is infomin updating of the uniform prior on $E(X)$ deceptive?

Notation: If A is a non-empty event, we write $E_0(X|A)$ for the mean value of X over A, i.e.:

$$E_0(X|A) = \frac{1}{|A|} \sum_{\omega \in A} x_\omega$$

We put $E_0(X) = E_0(X|\Omega)$.

Expected-Value Constraints

Question: For which events, $A \in \mathcal{B}$, is infomin updating of the uniform prior on $E(X)$ deceptive?

Notation: If A is a non-empty event, we write $E_0(X|A)$ for the mean value of X over A, i.e.:

$$E_0(X|A) = \frac{1}{|A|} \sum_{\omega \in A} x_\omega$$

We put $E_0(X) = E_0(X|\Omega)$.

A Sufficient Condition for Deceptiveness

Definition

A is a *mean* event for X if $E_0(X|A) = E_0(X)$.

Definition

A is a *interval* event iff there exist numbers a and b $(a \leq b)$, such that $A = \{\omega \in \Omega : a \leq x_\omega \leq b\}$.

Theorem

If $A \neq \Omega$ is a mean, interval event, then A is always decreasing under infomin updating with respect to the uniform prior and the family of constraints $\{E(X) = \lambda\}_\lambda$.

A Sufficient Condition for Deceptiveness

Definition

A is a *mean* event for X if $E_0(X|A) = E_0(X)$.

Definition

A is a *interval* event iff there exist numbers a and b $(a \leq b)$, such that $A = \{\omega \in \Omega : a \leq x_\omega \leq b\}$.

Theorem

If $A \neq \Omega$ is a mean, interval event, then A is always decreasing under infomin updating with respect to the uniform prior and the family of constraints $\{E(X) = \lambda\}_\lambda$.

A Sufficient Condition for Deceptiveness

Definition

A is a *mean* event for X if $E_0(X|A) = E_0(X)$.

Definition

A is a *interval* event iff there exist numbers a and b $(a \leq b)$, such that $A = \{\omega \in \Omega : a \leq x_\omega \leq b\}$.

Theorem

If $A \neq \Omega$ is a mean, interval event, then A is always decreasing under infomin updating with respect to the uniform prior and the family of constraints $\{E(X) = \lambda\}_\lambda$.

A Necessary Condition for Deceptiveness

Theorem

If an event A is always decreasing under infomin updating of the uniform prior with respect to the family $\{E(X) = \lambda\}_\lambda$, then A is a mean event for X.

Non-Deceptiveness and Higher Order Support

It follows from a previous result in *Shimony* (1973), that infomin updating of the uniform prior lacks higher order support, for any random variable X which takes more than two values.

However, since it is clear that we can select X so that no event in the algebra is a mean event (e.g., a three-sided die with faces $1, 2$ and 6), for some X infomin updating of the uniform prior on $E(X)$ is not-deceptive.

This shows that non-deceptiveness is a *strictly weaker* condition than higher order support.

Non-Deceptiveness and Higher Order Support

It follows from a previous result in *Shimony* (1973), that infomin updating of the uniform prior lacks higher order support, for any random variable X which takes more than two values.

However, since it is clear that we can select X so that no event in the algebra is a mean event (e.g., a three-sided die with faces $1, 2$ and 6), for some X infomin updating of the uniform prior on $E(X)$ is not-deceptive.

This shows that non-deceptiveness is a *strictly weaker* condition than higher order support.

Non-Deceptiveness and Higher Order Support

It follows from a previous result in *Shimony* (1973), that infomin updating of the uniform prior lacks higher order support, for any random variable X which takes more than two values.

However, since it is clear that we can select X so that no event in the algebra is a mean event (e.g., a three-sided die with faces $1, 2$ and 6), for some X infomin updating of the uniform prior on $E(X)$ is not-deceptive.

This shows that non-deceptiveness is a *strictly weaker* condition than higher order support.

Outline

Concluding Remarks

The fundamental connection that exists between probability and various notions of information does not by itself make clear the relevance of these latter notions for a theory of subjective probabilities.

If, as we argue, the unrestricted application of infomin methods leads to unacceptable consequences, then there must be objectionable moves in the proposed *a priori* arguments for these methods.

Concluding Remarks

The fundamental connection that exists between probability and various notions of information does not by itself make clear the relevance of these latter notions for a theory of subjective probabilities.

If, as we argue, the unrestricted application of infomin methods leads to unacceptable consequences, then there must be objectionable moves in the proposed *a priori* arguments for these methods.

Concluding Remarks

To provide a careful analysis of these arguments, in particular with regard to Paris & Vencovská (1997) – a work that proceeds directly from apparently plausible assumptions concerning subjective probabilities – can surely lead to a deeper understanding of the conceptual relationship between information and rational degrees of beliefs.

Csiszár, I. (1991), 'Why least squares and maximum entropy? an axiomatic approach to inference for linear inference problems', *The Annals of Statistics* **19**(4), 2032–2066.

Paris, J. & Vencovská, A. (1997), 'In defense of the maximum entropy inference process', *International Journal of Approximate Reasoning* **17**(1), 77–103.

Seidenfeld, T. (1986), 'Entropy and uncertainty', *Philosophy of Science* **53**(4), 467–491.

Shimony, A. (1973), 'Comment on the interpretation of inductive probabilities', *Journal of Statistical Physics* **9**(2), 187–191.

Shimony, A. (1985), 'The status of the principle of maximum entropy', *Synthese* **63**, 35–53.

Shore, J. & Johnson, R. (1980), 'Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy', *IEEE Trans. on Info. Theory* **IT-26**(1), 26–37.