Three Methods for Approximating Probability Distributions

Albert PEREZ Radim JIROUŠEK Milan STUDENÝ

Institute of Information Theory and Automation

DAR conference, December 19-20, 2005, Prague

Summary of the first part

The first part will be presented by M. Studený.

1. Some memories of Albert Perez
2. Considered situation - distributions with prescribed marginals
3. Possible use in multi-symptom diagnosis making
4. Dependence structure simplifications
5. Explicit expression approximation
6. Conclusions

Reference:

A. Perez, M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals

prepared for a special issue of *Kybernetika*.

Summary of the first part

The first part will be presented by M. Studený.

1. Some memories of Albert Perez
2. Considered situation - distributions with prescribed marginals
3. Possible use in multi-symptom diagnosis making
4. Dependence structure simplifications
5. Explicit expression approximation
6. Conclusions

Reference:

A. Perez, M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals

prepared for a special issue of *Kybernetika*.

Summary of the first part

The first part will be presented by M. Studený.

1. Some memories of Albert Perez
2. Considered situation - distributions with prescribed marginals
3. Possible use in multi-symptom diagnosis making
4. Dependence structure simplifications
5. Explicit expression approximation
6. Conclusions

Reference:

A. Perez, M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals

prepared for a special issue of *Kybernetika*.

Summary of the first part

The first part will be presented by M. Studený.

1. Some memories of Albert Perez
2. Considered situation - distributions with prescribed marginals
3. Possible use in multi-symptom diagnosis making
4. Dependence structure simplifications
5. Explicit expression approximation
6. Conclusions

Reference:

A. Perez, M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals

prepared for a special issue of *Kybernetika*.

Summary of the first part

The first part will be presented by M. Studený.

1. Some memories of Albert Perez
2. Considered situation - distributions with prescribed marginals
3. Possible use in multi-symptom diagnosis making
4. Dependence structure simplifications
5. Explicit expression approximation
6. Conclusions

Reference:

A. Perez, M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals

prepared for a special issue of *Kybernetika*.

Summary of the first part

The first part will be presented by M. Studený.

1. Some memories of Albert Perez
2. Considered situation - distributions with prescribed marginals
3. Possible use in multi-symptom diagnosis making
4. Dependence structure simplifications
5. Explicit expression approximation
6. Conclusions

Reference:

A. Perez, M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals

prepared for a special issue of *Kybernetika*.

Some memories of Albert Perez

The first part of this contribution is based on a paper prepared for a special issue of Kybernetika in honour of Albert Perez.

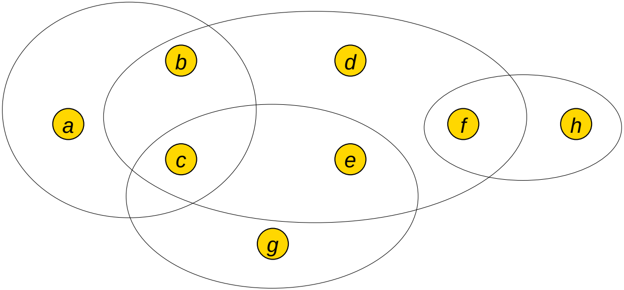
Perez was a founder of our research group interested in probabilistic and information-theoretical methods in decision-making.

He was a PhD supervisor of several research workers in the Institute of Information Theory and Automation.

The content of the first part of this contribution is based on the work Albert Perez did shortly before he passed away in December 2003.

Considered situation

- Let N be a finite non-empty set of variables,
- $\mathcal{S}$ a class of subsets of N such that $\bigcup \mathcal{S} = N$, and
- $\mathcal{M} = \{P_A; A \in \mathcal{S}\}$ a given system of discrete probability distributions



$$\mathcal{M} : P_{\{a,b,c\}}, P_{\{b,c,d,e,f\}}, P_{\{c,e,g\}}, P_{\{f,h\}}$$

Basic assumption

We assume that $\mathcal{M}$ is *strongly consistent* by which is meant that there exists at least one probability distribution P over N which has $\mathcal{M}$ as a system of marginal distributions.

Definition

The symbol $\mathcal{K}_{\mathcal{M}}$ will denote the system of discrete probability distributions over N that have the prescribed system of marginals $\mathcal{M}$.

We are interested in special approximations $\hat{P}$ of P, namely probability distributions "constructed" from $\mathcal{M}$ by means of "multiplication" of densities in a special way.
We will call these special approximations $\mathcal{M}$-constructs.

Basic assumption

We assume that $\mathcal{M}$ is *strongly consistent* by which is meant that there exists at least one probability distribution P over N which has $\mathcal{M}$ as a system of marginal distributions.

Definition

The symbol $\mathcal{K}_{\mathcal{M}}$ will denote the system of discrete probability distributions over N that have the prescribed system of marginals $\mathcal{M}$.

We are interested in special approximations $\hat{P}$ of P, namely probability distributions “constructed" from $\mathcal{M}$ by means of “multiplication" of densities in a special way.
We will call these special approximations $\mathcal{M}$-constructs.

Measuring quality of approximations

To compare the quality of approximations we use the relative entropy $H(P|\hat{P})$ as the measure of divergence of an approximation $\hat{P}$ from P.

The point is that the quality of an approximation $\hat{P}$ of the considered type actually does not depend on the choice of $P \in \mathcal{K}_{\mathcal{M}}$.

This is because, for any $P \in \mathcal{K}_{\mathcal{M}}$ and any $\mathcal{M}$-construct $\hat{P}$, one has

$$\boxed{H(P|\hat{P}) = I(P) - I_{\mathcal{M}}(\hat{P})}, \qquad (1)$$

where $I(P)$ is the multiinformation of P and

$I_{\mathcal{M}}(\hat{P})$ is an expression that only depends on $\hat{P}$ (and not on P).

The value of $I_{\mathcal{M}}(\hat{P})$ will be called the *multiinformation content* of $\hat{P}$.
Note that if $\hat{P} \in \mathcal{K}_{\mathcal{M}}$ then $I(\hat{P}) = I_{\mathcal{M}}(\hat{P})$.

Measuring quality of approximations

To compare the quality of approximations we use the relative entropy $H(P|\hat{P})$ as the measure of divergence of an approximation $\hat{P}$ from P.

The point is that the quality of an approximation $\hat{P}$ of the considered type actually does not depend on the choice of $P \in \mathcal{K}_{\mathcal{M}}$.

This is because, for any $P \in \mathcal{K}_{\mathcal{M}}$ and any $\mathcal{M}$-construct $\hat{P}$, one has

$$\boxed{H(P|\hat{P}) = I(P) - I_{\mathcal{M}}(\hat{P})}\,, \tag{1}$$

where $I(P)$ is the multiinformation of P and

$I_{\mathcal{M}}(\hat{P})$ is an expression that only depends on $\hat{P}$ (and not on P).

The value of $I_{\mathcal{M}}(\hat{P})$ will be called the *multiinformation content* of $\hat{P}$.

Note that if $\hat{P} \in \mathcal{K}_{\mathcal{M}}$ then $I(\hat{P}) = I_{\mathcal{M}}(\hat{P})$.

Possible use in multi-symptom diagnosis making I

Perez's motivation was to use approximations of this type in multi-symptom diagnosis making.

Assume that every variable $i \in N$ has a non-empty finite set of possible values X_i.

- Let $d \in N$ denote a *diagnostic variable*.
- The variables in $S \equiv N \setminus \{d\}$ will be called *symptom variables*.
- The decision should be based on an "observed configuration" of values $x_S \equiv [x_i]_{i \in S}$, where $x_i \in \mathsf{X}_i$ for $i \in S$.

On the basis of x_S, we would like to determine the most probable value of the diagnostic variable:

$$\hat{y} \rightleftharpoons \mathrm{argmax}\,\{\, P_{d|S}(y|x_S)\,;\ y \in \mathsf{X}_d \,\}.$$

Possible use in multi-symptom diagnosis making I

Perez's motivation was to use approximations of this type in multi-symptom diagnosis making.

Assume that every variable $i \in N$ has a non-empty finite set of possible values X_i.

- Let $d \in N$ denote a *diagnostic variable*.
- The variables in $S \equiv N \setminus \{d\}$ will be called *symptom variables*.
- The decision should be based on an "observed configuration" of values $x_S \equiv [x_i]_{i \in S}$, where $x_i \in \mathsf{X}_i$ for $i \in S$.

On the basis of x_S, we would like to determine the most probable value of the diagnostic variable:

$$\hat{y} \coloneqq \operatorname{argmax} \{ P_{d|S}(y|x_S) \, ; \; y \in \mathsf{X}_d \} .$$

Possible use in multi-symptom diagnosis making I

Perez's motivation was to use approximations of this type in multi-symptom diagnosis making.

Assume that every variable $i \in N$ has a non-empty finite set of possible values X_i.

- Let $d \in N$ denote a *diagnostic variable*.
- The variables in $S \equiv N \setminus \{d\}$ will be called *symptom variables*.
- The decision should be based on an "observed configuration" of values $x_S \equiv [x_i]_{i \in S}$, where $x_i \in \mathsf{X}_i$ for $i \in S$.

On the basis of x_S, we would like to determine the most probable value of the diagnostic variable:

$$\hat{y} \risingdotseq \mathrm{argmax}\, \{\, P_{d|S}(y|x_S)\, ;\; y \in \mathsf{X}_d \,\}.$$

Possible use in multi-symptom diagnosis making I

Perez's motivation was to use approximations of this type in multi-symptom diagnosis making.

Assume that every variable $i \in N$ has a non-empty finite set of possible values X_i.

- Let $d \in N$ denote a *diagnostic variable*.
- The variables in $S \equiv N \setminus \{d\}$ will be called *symptom variables*.
- The decision should be based on an "observed configuration" of values $x_S \equiv [x_i]_{i \in S}$, where $x_i \in \mathsf{X}_i$ for $i \in S$.

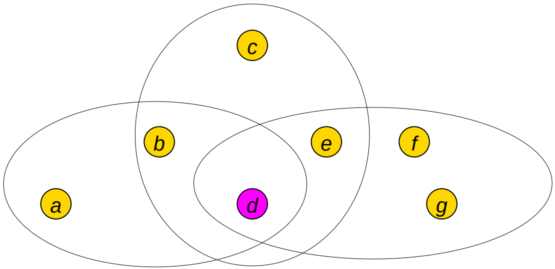
On the basis of x_S, we would like to determine the most probable value of the diagnostic variable:

$$\hat{y} \coloneqq \operatorname{argmax} \{ P_{d|S}(y|x_S) \, ; \; y \in \mathsf{X}_d \}.$$

Possible use in multi-symptom diagnosis making II

The complication is that we do not know the “actual" distribution P.

Thus, we try to replace P by its approximation $\hat{P}$ constructed from a given system of marginals $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ with $d \in A$ for every $A \in \mathcal{S}$.



$$\mathcal{S} = \{\ \{a, b, \boxed{d}\},\ \{b, c, \boxed{d}, e\},\ \{\boxed{d}, e, f, g\}\ \}$$

Two methodological approaches to multi-symptom diagnosis making

The first approach is based on *direct approximation* of P. We simply use an $\mathcal{M}$-construct $\hat{P}$ instead of P.

The respective estimator (of the value of the diagnostic variable) is then as follows:

$$
\begin{aligned}
\psi_1(x_S) &= \operatorname{argmax}\{\hat{P}_{d|S}(y|x_S);\ y \in \mathsf{X}_d\} \\
&\equiv \operatorname{argmax}\{\hat{P}([y, x_S])\,;\ y \in \mathsf{X}_d\}\,.
\end{aligned}
$$

Bayesian approach

The second option is a *Bayesian approach*. A prior distribution Q_d is given on X_d and one uses $Q_d \cdot \hat{P}_{S|d}$ instead of P, where $\hat{P}_{S|d}$ is an estimate of the respective conditional probability.

More specifically, for any fixed configuration $y \in \mathsf{X}_d$, consider the system of marginals of the “conditional probability" $P_{S|d}(\star|y)$:

$$\mathcal{M}[y] = \{P_{A\setminus\{d\}|d}(\star|y)\,;\,A \in \mathcal{S}\}.$$

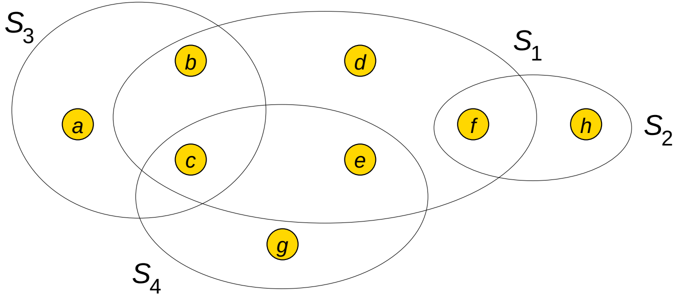
On basis of $\mathcal{M}[y]$, the respective approximation $\hat{P}_{[y]}$ of $P_{S|d}(\star|y)$ is computed. This leads to the following estimator:

$$\psi_2(x_S) \;=\; \mathrm{argmax}\,\{\,Q_d(y)\cdot \hat{P}_{[y]}(x_S)\,;\; y \in \mathsf{X}_d\,\}\,.$$

Dependence structure simplifications

This way of approximating measures from $\mathcal{K}_{\mathcal{M}}$ was already been proposed by Perez in the 1970s (and studied in PhD thesis of the speaker).

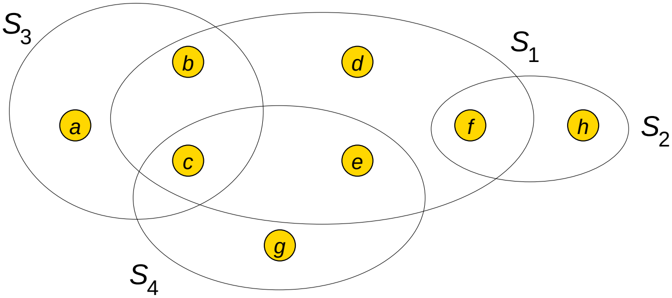
Let us choose a total ordering $\tau : S_1, \ldots, S_n$, $n \geq 1$ of elements of $\mathcal{S}$.



$S_1 = \{b, c, d, e, f\}$, $S_2 = \{f, h\}$, $S_3 = \{a, b, c\}$, $S_4 = \{c, e, g\}$

Dependence structure simplifications - example

Let us put put $F_j \equiv S_j \cap (\bigcup_{k<j} S_k)$ and $G_j \equiv S_j \setminus F_j$ for $1 \leq j \leq n$.

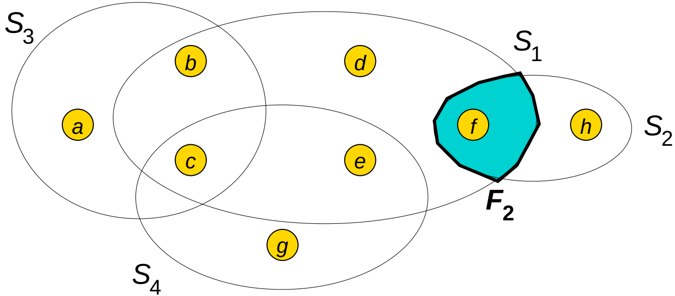


$F_1 = \emptyset$

$G_1 = \{b, c, d, e, f\}$

Dependence structure simplifications - example

Let us put put $F_j \equiv S_j \cap (\bigcup_{k<j} S_k)$ and $G_j \equiv S_j \setminus F_j$ for $1 \le j \le n$.

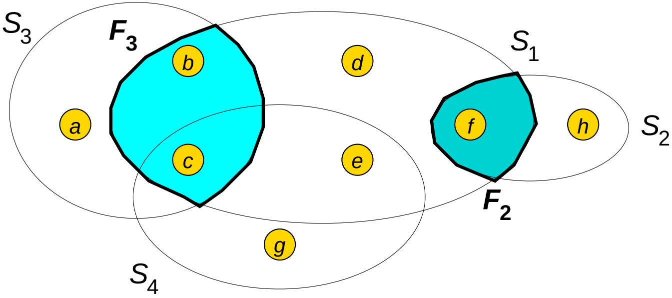


$F_1 = \emptyset$, $\boxed{F_2 = \{f\}}$

$G_1 = \{b, c, d, e, f\}$, $\boxed{G_2 = \{h\}}$

Dependence structure simplifications - example

Let us put put $F_j \equiv S_j \cap (\bigcup_{k<j} S_k)$ and $G_j \equiv S_j \setminus F_j$ for $1 \le j \le n$.

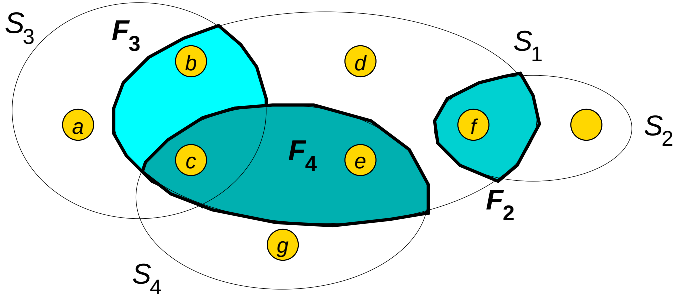


$F_1 = \emptyset$, $F_2 = \{f\}$, $F_3 = \{b, c\}$

$G_1 = \{b, c, d, e, f\}$, $G_2 = \{h\}$, $G_3 = \{a\}$

Dependence structure simplifications - example

Let us put put $F_j \equiv S_j \cap (\bigcup_{k<j} S_k)$ and $G_j \equiv S_j \setminus F_j$ for $1 \leq j \leq n$.



$F_1 = \emptyset$, $F_2 = \{f\}$, $F_3 = \{b, c\}$, $F_4 = \{c, e\}$

$G_1 = \{b, c, d, e, f\}$, $G_2 = \{h\}$, $G_3 = \{a\}$, $G_4 = \{g\}$

Definition of a dependence structure simplification

By a *choice* for $\mathcal{M}$ and τ we will understand a mapping ϑ which assigns a conditional density $p_{G_j|F_j}$ on X_{G_j} given X_{F_j} consistent with p_{S_j} to every $1 \le j \le n$. Here, $\mathsf{X}_A \equiv \prod_{i \in A} \mathsf{X}_i$ for $A \subseteq N$.

Definition

By a *dependence structure simplification* (DSS) for $\mathcal{M}$ determined by an ordering τ (and a choice ϑ) will be understood the probability measure $\overline{P}_{\tau,\vartheta}$ on X_N whose density $\overline{p}_{\tau,\vartheta}$ is given by

$$\overline{p}_{\tau,\vartheta}(x) = \prod_{j=1}^{n} p_{G_j|F_j}(x_{G_j}|x_{F_j}) \qquad \text{for every } x \in \mathsf{X}_N\,. \tag{2}$$

Remark on the definition

REMARK

Note that if $p_A > 0$ on X_A for every $A \in \mathcal{S}$ the the choice ϑ is unique and the respective DSS only depends on $\mathcal{M}$ and τ. Then

$$\overline{p}_\tau(x) = \prod_{j=1}^{n} \frac{p_{S_j}(x_{S_j})}{p_{F_j}(x_{F_j})} \qquad \text{for any } x \in \mathsf{X}_N \,, \tag{3}$$

where $p_\emptyset(x_\emptyset) \equiv 1$ by a convention.

An example of a DSS

$S_1 = \{b, c, d, e, f\}$, $S_2 = \{f, h\}$, $S_3 = \{a, b, c\}$, $S_4 = \{c, e, g\}$

The formula (2) gives

$$\bar{p}_\tau = p_{\{b,c,d,e,f\}} \cdot p_{h|f} \cdot p_{a|bc} \cdot p_{g|ce} \,.$$

More detailed way of writing it as follows:

$$\begin{aligned}\bar{p}_\tau(x_a, x_b, x_c, x_d, x_e, x_f, x_g, x_h) \quad = \quad & p_{\{b,c,d,e,f\}}(x_b, x_c, x_d, x_e, x_f) \\ & \cdot p_{h|f}(x_h|x_f) \\ & \cdot p_{a|bc}(x_a|x_b, x_c) \\ & \cdot p_{g|ce}(x_g|x_c, x_e) \,.\end{aligned}$$

The multiinformation content of a DSS

Lemma

The formula for the multiinformation content of a DSS $Q = \overline{P}_{\tau,\vartheta}$ is as follows:

$$I_{\mathcal{M}}(Q) = \sum_{A \in \mathcal{S}} I(P_A) - \sum_{j=2}^{n} I(P_{F_j})\,. \tag{4}$$

Observe that it only depends on the ordering τ, not on the choice ϑ.

Optimal dependence structure simplifications

Perez was interested in the problem of finding an optimal DSS.

Definition

Let $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures and $\mathcal{D}_{\mathcal{M}}$ denotes the class of DSS approximations. We say that $Q \in \mathcal{D}_{\mathcal{M}}$ is *optimal* relative to $P \in \mathcal{K}_{\mathcal{M}}$ if

$$H(P|Q) = \min\{\, H(P|Q');\ Q' \in \mathcal{D}_{\mathcal{M}}\}\,.$$

It follows from formula (1) that a DSS Q is optimal iff it maximizes $I_{\mathcal{M}}(Q)$ given by (4). In particular, the optimality of a DSS does not depend on a choice of $P \in \mathcal{K}_{\mathcal{M}}$

Explicit expression: non-normalized version

This is way of approximation was proposed by Perez in 2003.

Given $n \in \mathbb{N}$, let us denote

$$sg(n) = \begin{cases} +1 & \text{if } n \text{ is even,} \\ -1 & \text{if } n \text{ is odd.} \end{cases}$$

Definition

Let $\mathcal{M} = \{P_A;\, A \in \mathcal{S}\}$ be a strongly consistent collection of probability measures. We put

$$\mathrm{Exe}\,(x) = \prod_{\emptyset \neq \mathcal{A} \subseteq \mathcal{S}} p_{\bigcap \mathcal{A}}(x_{\bigcap \mathcal{A}})^{-sg(|\mathcal{A}|)} \qquad \text{for every } x \in \mathsf{X}_N\,, \tag{5}$$

where we accept the convention that $0^{-1} \equiv 0$.

Explicit expression: normalized version

Definition

The *norm* (of an explicit expression Exe) is the number

$$c = \sum_{x \in \mathsf{X}_N} \mathrm{Exe}\,(x)$$

It will be denoted by $|\mathrm{Exe}\,|$. Moreover, we put

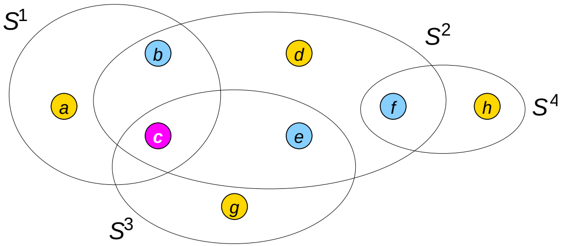
$$\overline{\mathrm{Exe}}(x) = c^{-1} \cdot \mathrm{Exe}\,(x) \quad \text{for every } x \in \mathsf{X}_N\,. \tag{6}$$

$\overline{\mathrm{Exe}}$ is a density of a probability measure on X_N, denoted by P_{exe}.

If we base our estimator on direct approximation of P by means of the explicit expression $\hat{P} = P_{exe}$, then it is *not necessary* to compute the norm $|\mathrm{Exe}\,|$. In this particular case, one has

$$\psi_1(x_S) = \operatorname{argmax}\,\{\mathrm{Exe}\,([y, x_S])\,;\ y \in \mathsf{X}_d\,\}\,.$$

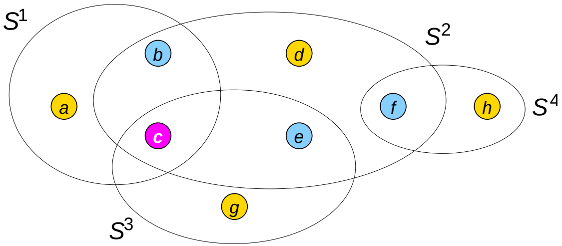
Explicit expression - example



As $p_\emptyset \equiv 1$ we limit our attention to classes $\mathcal{A} \subseteq \mathcal{S}$ with $\bigcap \mathcal{A} \neq \emptyset$.
$|\mathcal{A}| = 1$: $S^1 = \{a, b, c\}$, $S^2 = \{b, c, d, e, f\}$, $S^3 = \{c, e, g\}$, $S^4 = \{f, h\}$
$|\mathcal{A}| = 2$: $S^1 \cap S^2 = \{b, c\}$, $S^1 \cap S^3 = \{c\}$, $S^2 \cap S^3 = \{c, e\}$, $S^2 \cap S^4 = \{f\}$
$|\mathcal{A}| \geq 3$: only $S^1 \cap S^2 \cap S^3 = \{c\}$ is non-empty

$$\mathrm{Exe}(x) = \frac{p_{\{a,b,c\}} \cdot p_{\{b,c,d,e,f\}} \cdot p_{\{c,e,g\}} \cdot p_{\{f,h\}}}{p_{\{b,c\}} \cdot p_{\{c\}} \cdot p_{\{c,e\}} \cdot p_{\{f\}}} \cdot \frac{p_{\{c\}}}{1}$$

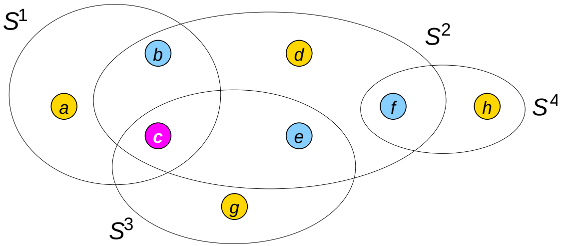
Explicit expression - example



As $p_\emptyset \equiv 1$ we limit our attention to classes $\mathcal{A} \subseteq \mathcal{S}$ with $\bigcap \mathcal{A} \neq \emptyset$.
$|\mathcal{A}| = 1$: $S^1 = \{a,b,c\}$, $S^2 = \{b,c,d,e,f\}$, $S^3 = \{c,e,g\}$, $S^4 = \{f,h\}$
$|\mathcal{A}| = 2$: $S^1 \cap S^2 = \{b,c\}$, $S^1 \cap S^3 = \{c\}$, $S^2 \cap S^3 = \{c,e\}$, $S^2 \cap S^4 = \{f\}$
$|\mathcal{A}| \geq 3$: only $S^1 \cap S^2 \cap S^3 = \{c\}$ is non-empty

$$\mathrm{Exe}\,(x) = \frac{p_{\{a,b,c\}} \cdot p_{\{b,c,d,e,f\}} \cdot p_{\{c,e,g\}} \cdot p_{\{f,h\}}}{p_{\{b,c\}} \cdot p_{\{c\}} \cdot p_{\{c,e\}} \cdot p_{\{f\}}} \cdot \frac{p_{\{c\}}}{1}$$

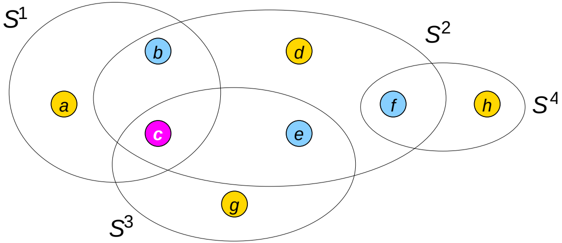
Explicit expression - example



As $p_\emptyset \equiv 1$ we limit our attention to classes $\mathcal{A} \subseteq \mathcal{S}$ with $\bigcap \mathcal{A} \neq \emptyset$.
$|\mathcal{A}| = 1$: $S^1 = \{a, b, c\}$, $S^2 = \{b, c, d, e, f\}$, $S^3 = \{c, e, g\}$, $S^4 = \{f, h\}$
$|\mathcal{A}| = 2$: $S^1 \cap S^2 = \{b, c\}$, $S^1 \cap S^3 = \{c\}$, $S^2 \cap S^3 = \{c, e\}$, $S^2 \cap S^4 = \{f\}$
$|\mathcal{A}| \geq 3$: only $S^1 \cap S^2 \cap S^3 = \{c\}$ is non-empty

$$\mathrm{Exe}\,(x) = \frac{p_{\{a,b,c\}} \cdot p_{\{b,c,d,e,f\}} \cdot p_{\{c,e,g\}} \cdot p_{\{f,h\}}}{p_{\{b,c\}} \cdot p_{\{c\}} \cdot p_{\{c,e\}} \cdot p_{\{f\}}} \cdot \frac{p_{\{c\}}}{1}$$

Explicit expression - example



After cancellation of the term $p_{\{c\}}$ we get:

$$\mathrm{Exe}\,(x) = \frac{p_{\{a,b,c\}} \cdot p_{\{b,c,d,e,f\}} \cdot p_{\{c,e,g\}} \cdot p_{\{f,h\}}}{p_{\{b,c\}} \cdot p_{\{c,e\}} \cdot p_{\{f\}}}$$

Hence,

$$|\mathrm{Exe}| = \sum_{x \in \mathsf{X}_N} \mathrm{Exe}\,(x) = 1 \quad \Rightarrow \quad \overline{\mathrm{Exe}} = Exe\,.$$

The multiinformation content of an explicit expression

Lemma

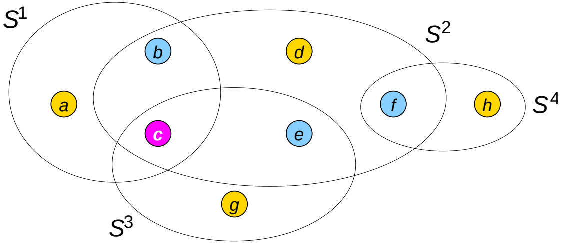
The formula for the multiinformation content of the normalized explicit expression approximation is as follows:

$$I_{\mathcal{M}}(P_{exe}) = -\ln|\mathrm{Exe}| + \sum_{B \in \mathcal{S}^{\downarrow}} \nu(B) \cdot I(P_B), \qquad (7)$$

where $\mathcal{S}^{\downarrow} = \{B;\ B \subseteq A \in \mathcal{S}\}$ *and*

$$\nu(B) = \sum \{-sg(|\mathcal{A}|);\ \emptyset \neq \mathcal{A} \subseteq \mathcal{S}, \bigcap \mathcal{A} = B\} \quad \text{for any } B \in \mathcal{S}^{\downarrow}. \qquad (8)$$

Multiinformation content of an explicit expression - example



In the preceding example we have:

$$\begin{aligned} I_{\mathcal{M}}(P_{exe}) &= I(P_{\{a,b,c\}}) + I(P_{\{b,c,d,e,f\}}) + I(P_{\{c,e,g\}}) + I(P_{\{f,h\}}) \\ &\quad - I(P_{\{b,c\}}) - I(P_{\{c,e\}}) - I(P_{\{f\}}) \end{aligned}$$

Note that $\ln|\mathrm{Exe}| = 0$ and $\nu(\{c\}) = 0$.

Conclusions of the first part

The original Perez's conjecture was that the explicit expression approximation is always better than any DSS approximation in the sense of the multiinformation content. We gave examples that, in general, none of these two methods is better than the other.

However, these approximations often coincide, for example if there exists an ordering τ of $\mathcal{S}$ satisfying the *running intersection property*:

$$\forall j > 2 \quad \exists \ell < j \quad F_j \equiv S_j \cap (\bigcup_{k<j} S_k) \subseteq S_\ell \,. \tag{9}$$

Note that the chosen distribution is then in concordance with the maximum entropy principle.

Perez was also interested in the question whether, in the considered situation with prescribed marginals, the maximum entropy principle coincides with his *barycenter principle*. We gave an example that it need not be the case.