SORT

Statistics and Operations Research Transactions

Volume 27
Number 2, July - December 2003

ISSN: 1696-2281

Generalitat de Catalunya
Institut d'Estadística
de Catalunya

Aims

SORT (*Statistics and Operations Research Transactions*) – formerly *Qüestiió* - is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-sponsored by the Universitat Politècnica de Catalunya (UPC), Universitat de Barcelona, Universitat Autònoma de Barcelona and Universitat de Girona and with the co-operation of the Spanish Region of the International Biometric Society. SORT promotes the publication of original articles of a methodological or applied nature on statistics, operations research, official statistics and biometrics.

The journal is described in the *Encyclopedia of Statistical Sciences*, and referenced in the *Current Index to Statistics*, the *Índice Español de Ciencia y Tecnología*, *Statistical Methods and Abstracts*, as well as MathSci of the American Mathematical Society (*Current Mathematical Publications and Mathematical Reviews*).

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

ISSN: 1696-2281
SORT 27 (2) July - December 139/190 (2003)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat de Catalunya
Institut d'Estadística
de Catalunya

SORT 27 (2) July-December 2003

Contents

Articles

PAPER ECOLÒGIC

SORT 27 (2) July-December 2003, 139-152

Partial cooperation and convex sets

J. Enrique Romero García*, Jorge J. López Vázquez

Universidad de Sevilla, Spain

Abstract

We consider games of transferable utility, those that deal with partial cooperation situations, made up of coalition systems, in which every unit coalition is feasible and every coalition of players can be expressed as a disjoint union of maximal feasible coalitions. These systems are named *partition systems* and cause restricted games. To sum up, we study feasible coalition systems defined by a partial order designed for a set of players and we analyze the characteristics of a feasible coalition system developed from a family of convex sets.

MSC: 90D12

Keywords: cooperative games, partial cooperation, convex sets

1 Partial cooperation

A system offeasible cooperations is defined by $(N, \mathcal{F})$, $\mathcal{F} \subseteq 2^N$, that proves the following axiom:
$(P1)$ *$\emptyset \in \mathcal{F}$, and the group $\{i\} \in \mathcal{F}$ $\forall\, i \in N$.*

Considering the given explanation, it results that any coalition $S \subseteq N$ can be expressed by a disjoint union of feasible coalitions, as

$$S = \bigcup_{a \in S} \{a\}.$$

However, this partition of S for *feasible coalitions* should not be unique. In general, we will denote $\mathcal{P}_{\mathcal{F}}(S)$ the set made up of partitions of $S \subseteq N$ in nonempty feasible

[*] *Address for correspondence:* José Enrique Romero García. romerogje@us.es. Facultad de Ciencias Económicas y Empresariales. Universidad de Sevilla. Avda. Ramón y Cajal, nº 1 41018-Sevilla. Spain.

Received: November 2001
Accepted: October 2003

coalitions. Obviously $\mathcal{P}_{\mathcal{F}}(\emptyset) = \{\emptyset\}$. The previous reasoning gives sense to and makes consistent the idea of a *restricted cooperation game*: *Define the triple* $(N, \mathcal{F}, v)$, *in which* $(N, \mathcal{F})$ *is a feasible coalition system and* (N, v) *a transferable utility game. Then the couple* $(N, v^{\mathcal{F}})$ *in which*

$$v^{\mathcal{F}} : 2^N \longrightarrow \mathbb{R}, \quad v^{\mathcal{F}}(S) = \max\left\{\sum_i v(T_i) \mid \{T_i\} \in \mathcal{P}_{\mathcal{F}}(S)\right\}.$$

is termed a game with restricted cooperation by the feasible coalition system $(N, \mathcal{F})$.

The supplied explanation for *game of restricted cooperation by a system of feasible coalitions* is for every coalition of players, an extension of the one by Faigle (1989) concerning games with restricted cooperations and by Bergantiños, Carreras and García–Jurado (1993) when using communication graphs to show incompatibility among some of the players. Indeed, it can be shown that $v^{\mathcal{F}}(S) \geq \sum_{i \in S} v(\{i\})$. Defined this way, the game is always superadditive.

Let $(N, \mathcal{F})$ *be a system of feasible coalitions. Let* $S \subseteq N$. *It is said that* T *is* $\mathcal{F}$–*component of* S *if it is proved that* $T \in \mathcal{F}$ *and* $T' \in \mathcal{F}$ *does not exist, as* $T \subset T' \subseteq S$. That is to say, the $S \subseteq N$ $\mathcal{F}$–components are the *maximal feasible coalitions* included in S and, for any $S \subseteq N$, the $\mathcal{F}$–components *of* S are a collection $\{T_k\}_k \subset 2^S$ such that

$$S = \bigcup_k T_k$$

But, the $\mathcal{F}$–components of $S \subseteq N$ are not necessarily a partition of S as its intersection can be nonempty.

It can be proved that if we consider $(N, \mathcal{F}, v)$, *where* $(N, \mathcal{F})$ *is a feasible coalition system,* (N, v) *a superadditive game and, for each coalition* $S \subseteq N$, *the* $\mathcal{F}$–*components of* S *are a partition of itself, then the restricted cooperation game* $(N, v^{\mathcal{F}})$ *verifies*

$$v^{\mathcal{F}}(S) = \sum_k v(T_k),$$

where $\{T_k\}_k \in \mathcal{P}_{\mathcal{F}}$, *the* S *partition for its maximal feasible coalitions* ($\mathcal{F}$–*components of* S).

Therefore, if the $\mathcal{F}$–components of any coalition are a partition of itself, and the game (N, v) is superadditive, then the restricted game by the system of feasible coalitions is determined by

$$v^{\mathcal{F}}(S) = \sum_k v(T_k),$$

in which $\{T_k\}_k$ is the S partition for maximal feasible coalitions. As the previous expression requires that maximal feasible coalitions must be disjointed, a new definition for a concrete feasible coalitions system has to be looked for. It will be named a *partition system* .

A partition system is the couple $(N,\mathcal{F})$, $\mathcal{F} \subseteq 2^N$ that verifies the following axioms:

$(P1)$ $\emptyset \in \mathcal{F}$, $\{i\} \in \mathcal{F}\ \forall\, i \in N$.

$(P2)$ $\forall\, S \subseteq N$, *the S maximal subsets in $\mathcal{F}$ ($\mathcal{F}$–components of S) are a partition of S, denoted by*

$$C_{\mathcal{F}}(S) = \{S_1, \ldots, S_k\}.$$

Evidently, a *partition system* is a *feasible coalitions system*, so, the $\mathcal{F}$ elements will not change their name.

Example 1 *Let $N = \{1, 2, \ldots, n\}$, a natural number n, and considering the collection $\mathcal{L}_n$ made of all the sets such as $[i, j] = \{i, i+1, \ldots, j-1, j\}$ for $1 \leq i \leq j \leq n$. This model represents a one-dimensional political election situation and the couple $(N, \mathcal{L}_n)$ is a partition system.*

Example 2 *A communication situation is the triple (N, G, v), in which (N, v) is a game and $G = (N, E(N))$ is a graph. This idea was first developed by Myerson (1977), and researched by Owen (1986) and Borm, Nouweland and Tijs (1992, 1993). It is easy to see that the couple $(N, \mathcal{F})$, in which*

$$\mathcal{F} = \{S \subseteq N \mid (S, E(S)) \text{ is a connected subgraph of } G\},$$

is a partition system. We must point out that the opposite is not always true, because every G graph is a collection of pairs $\{i, j\}$, and as a result, there must be feasible collections made up of two elements, but this might not happen.

The previous definitions come from an extension of communication situation and communication graph-restricted game, developed by Myerson (1977) and Owen (1986).

The following theorem shows a characterization of the concept of partition systems.

Theorem 1 *A feasible coalitions system $(N,\mathcal{F})$, $\mathcal{F} \subseteq 2^N$ is a partition system if and only if*

$$\forall A \in \mathcal{F}, B \in \mathcal{F}, \text{ con } A \cap B \neq \emptyset \implies A \cup B \in \mathcal{F}.$$

Proof. ($\Leftarrow$) Considering that the $\mathcal{F}$-components of $A \subseteq N$ form a recover, it is only necessary to prove that every pair of *$\mathcal{F}$–components* of A are disjointed. Let T_i, T_j ($i \neq j$) maximal feasible coalitions of A. If $T_i \cap T_j \neq \emptyset$,it would mean, hypothesizing, $T_i \cup T_j \in \mathcal{F}$ being $T_i \cup T_j \subset A$. This contradicts that T_i and T_j are maximal feasible coalitions of A.

($\Rightarrow$) Let $A \in \mathcal{F}$, $B \in \mathcal{F}$ with $A \cap B \neq \emptyset$. If $A \cup B \notin \mathcal{F}$, then

$$A \cup B = \bigcup_k T_k,$$

where $\{T_k\}$ is the partition of $A \cup B$ for maximal sets. As A and B are feasible coalitions contained in $A \cup B$, thus $A \subseteq T_j$, $B \subseteq T_p$ for every j and p. If $j \neq p$, then $T_j \cap T_p = \emptyset$

and, so, $A \cap B = \emptyset$ against the hypothesis; then $A \cup B \in \mathcal{F}$. If $j = p$ then $A \subseteq T_j \subseteq A \cup B$ and $B \subseteq T_j \subseteq A \cup B$, implies $A \cup B = T_j \in \mathcal{F}$. □

2 Partially ordered set restricted games

The aim of this section is to study a feasible coalition system defined by a partial order for all players. From this moment only posets $P = (N, \leq)$ will be considered and the feasible coalition system characteristics developed from the family of convex sets will be analyzed.

Let $P = (N, \leq)$ a poset. It is said that $A \subseteq N$ is convex in P if it is proved that

$$a \in A,\ b \in A \quad \text{and} \quad a \leq b \implies [a, b] \subseteq A.$$

If $P = (N, \leq)$ is a poset, we are interested in obtaining $P^* = (N, \leq)$, the dual of P, with

$$x \leq y \ \text{ en } P^* \iff y \leq x \ \text{ en } P.$$

It can be proved that $Co(P) \simeq Co(P^*)$, $\forall\ P$ (Birkoff and Bennett, 1985).The family of convex sets in P will be denoted

$$Co(P) = \{S \subseteq N \mid S \text{ is convex in } P\}.$$

This characterization implies, $\forall\ i \in N$, $\{i\} \in Co(P)$ so the couple $(N, Co(P))$ *is a feasible coalitions system.* Then, given a game (N, v), if there is an order relation among the players, it makes sense to take into consideration the triple $(N, Co(P), v)$ and the appropriate partial cooperation game,

$$v^{Co(P)} \mid 2^N \longrightarrow \mathbb{R},\ v^{Co(P)}(S) = \max\left\{\sum_i v(T_i) \mid \{T_i\} \in \mathcal{P}_{Co(P)}(S)\right\},$$

where $\mathcal{P}_{Co(P)}(S)$ is the family of partitions from the coalition S in convex sets in P.

It is easy to prove that $A, B \in Co(P)$, that $A \cap B \in Co(P)$, impliying $(N, Co(P))$ a closure space. Also, Edelman and Jamison (1985), Birkoff and Bennett (1985) think that $(N, Co(P))$ proves the Minkowski–Krein–Milman condition, and, therefore an atomic convex geometry, named order convex in N.

As $(N, Co(P))$ is a feasible coalition system, every subset in N can be expressed as a union of it maximal convex sets. In this particular case, the maximal convex definition of $S \subseteq N$ in P is equivalent to the one by Tijs (1993), which is due to the two $(N, Co(P))$ being a convex geometry: *Let $(N, Co(P))$ be a feasible coalition system and let $S \subseteq N$. If $T \in Co(P)$ and $T \subseteq N$, then T is maximal convex S in P if and only if, $\forall i \in S \setminus T$, $\quad T \cup \{i\} \notin Co(P)$.*

Notice that this characterization for maximal convex is certain in all convex geometry, and, in general, the feasible coalition system $(N, Co(P)$ is not a partition system.

Example 2 *Let $(N, \leq)$ be a poset, whose Hasse diagram is shown in Figure 1,*

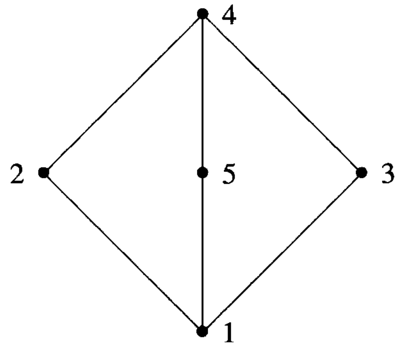
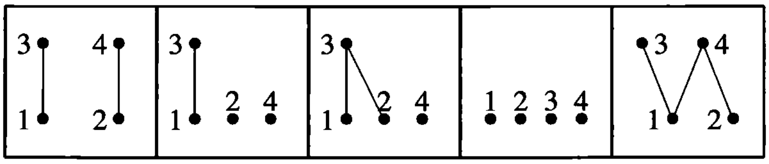


Figure 1



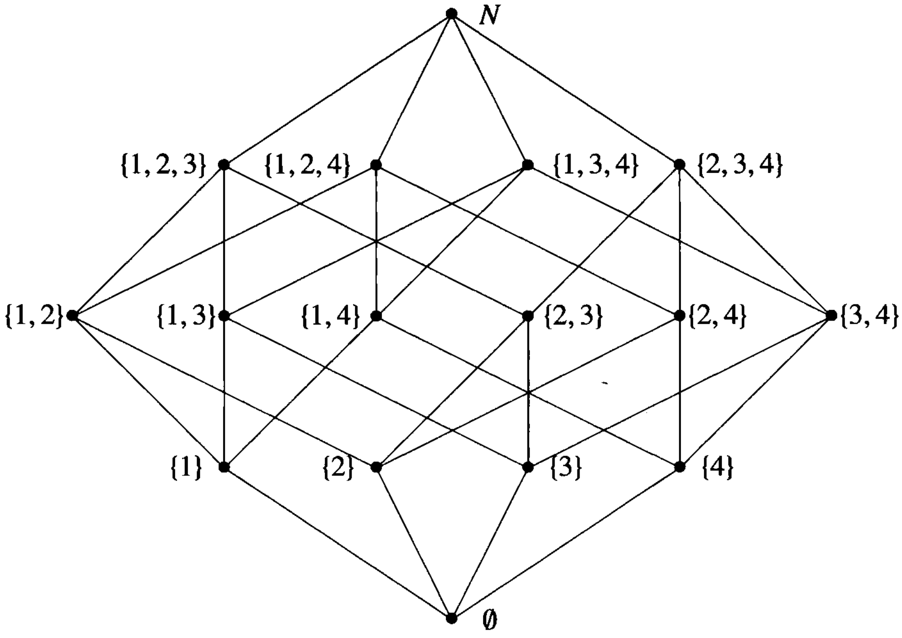


***Figure 2**: $(Co(P), \subseteq) \simeq (2^4, \subseteq)$*

The couple $(N, Co(P))$ is not a partition system, applying Theorem 1, because $\{1,3\} \in Co(P)$, $\{3,4\} \in Co(P)$, the intesection is not empty, however, $\{1,3\} \cup \{3,4\} \notin Co(P)$ due to $1 \leq 4$ y $[1,4] \not\subseteq \{1,3,4\}$.

Let $P = (N, \leq)$ be a poset whose range or length $l(P)$ might equal 1 or be less than 1. That is to say:

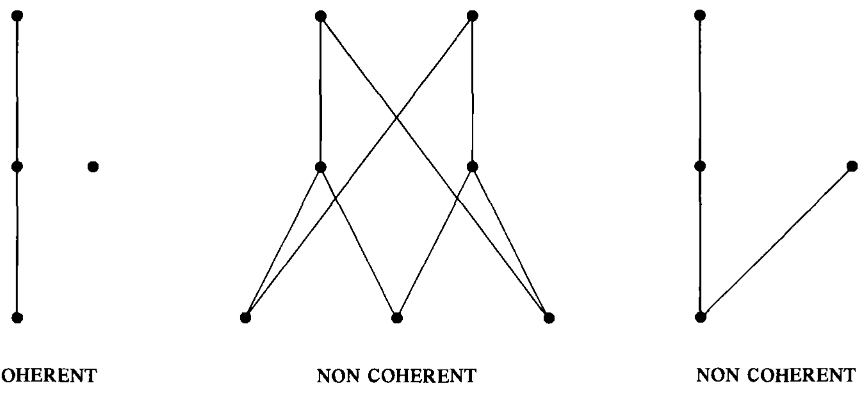
$$l(P) = \max\{l(C) \mid C \text{ is a chain in } P \text{ and } l(C) = |C| - 1\} \leq 1.$$

Then $(N, Co(P))$ is a partition convex geometry. As every subset in N is convex, either due to being an atom or a chain of two elements from N, it implies that $Co(P) \simeq 2^N$. For example, in Figure 2, $Co(P) \simeq 2^4$. If $l(P) \leq 1$ and if it is considered a partition system (or partition convex geometry) restricted $Co(P)$–game linked to the three $(N, Co(P), v)$, it verifies that $v^{Co(P)}(S) = v(S)$, $\forall S \in 2^N$ and, therefore restricted game and original game are the same.

It has been proved that if $l(P) \geq 2$, the atomic convex geometry $(N, Co(P))$ is not necessarily a partition system . This is the reason why only partially ordered sets with $l(P) \geq 2$ are taken into consideration, and we search for conditions to set $(N, Co(P))$ as a partition system. We will introduce the concept of completely coherent ordered sets as given by Birkoff and Bennett (1985).

A poset $P = (N, \leq)$ is *coherent* if it is connected and no maximal element from P covers any minimal element from P.

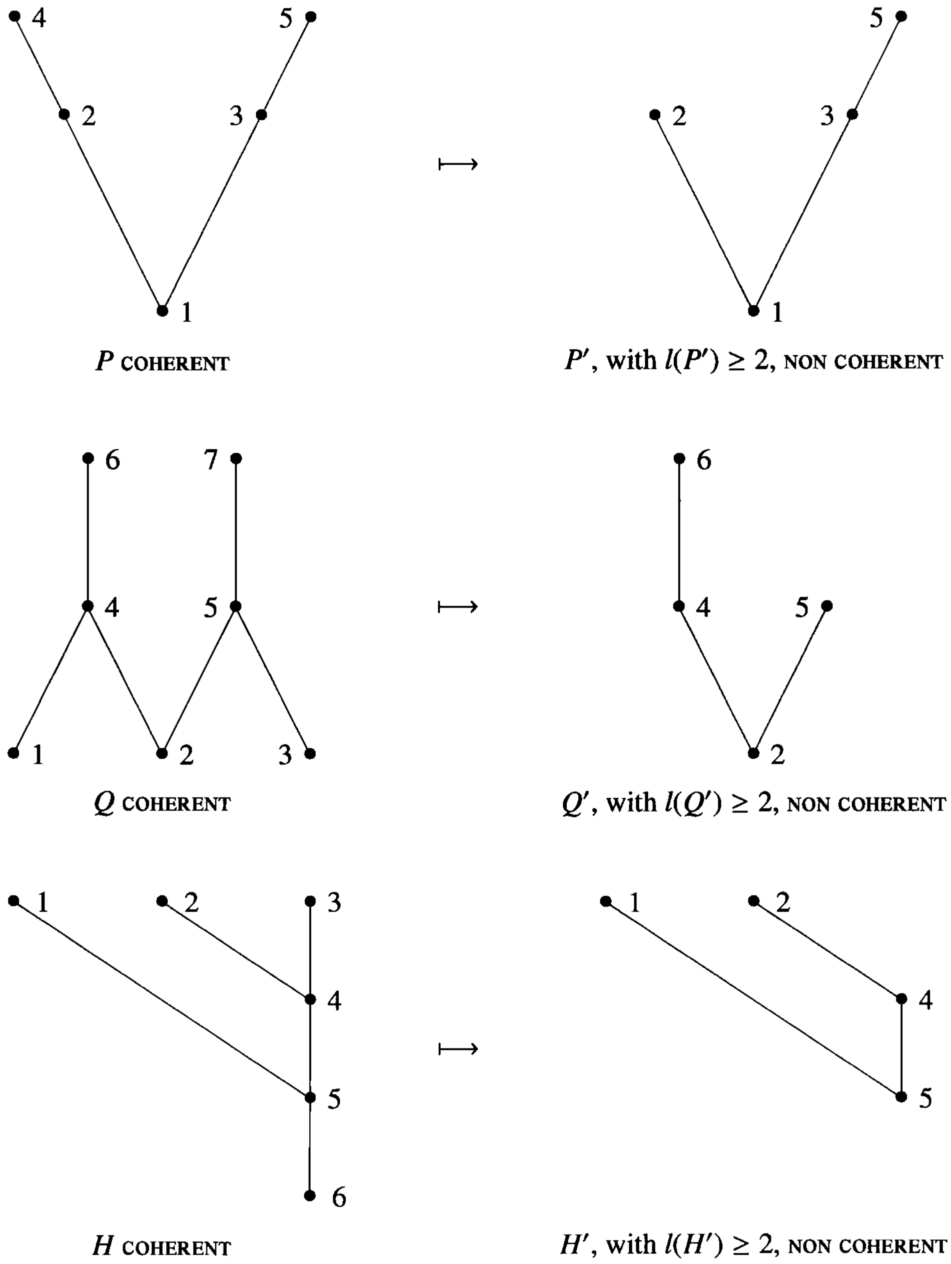
For example, the poset in example 3 (Figure 1) is *coherent*. Other possible situations are considered below:



(non connected) (maximals are about minimals)

Figure 3

A poset P, with $l(P) \geq 2$, is completely coherent if any subposet infered by P, P' with $l(P') \geq 2$, is coherent. The following figures illustrate this concept. Figure 4 shows diagrams of *coherent* posets that are not *completely coherent.* On the other hand, Figure 5, shows examples of *completely coherent* posets.



P COHERENT

P', with $l(P') \geq 2$, NON COHERENT

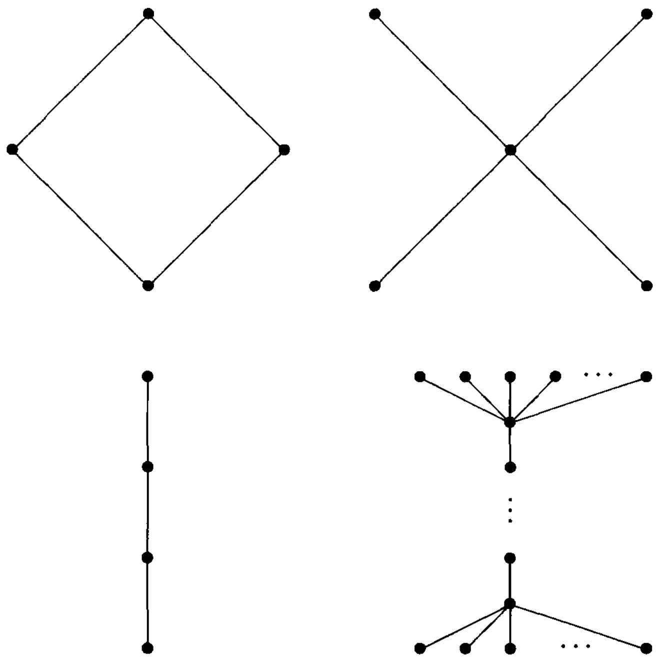
Q COHERENT

Q', with $l(Q') \geq 2$, NON COHERENT

H COHERENT

H', with $l(H') \geq 2$, NON COHERENT

Figure 4



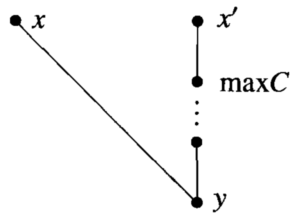
COMPLETELY COHERENT ORDERED SETS

Figure 5

Notice that completely coherent posets in Figure 5, except the first of them, verify that $P \setminus ex(P)$ is a chain. This property will be very important to prove that the couple $(N, Co(P))$ is a partition system.

Theorem 2 *Let $P = (N, \leq)$ be a completely coherent finite poset, as $P \backslash ex(P)$ is a chain C. Then, every maximal element from P covers the maximum in chain C and the minimal element from C covers every minimal element from P.*

Proof. If P is *coherent*, it is connected and its maximal elements do not cover any minimal. Therefore, if x is maximal, it follows that $y \in P$ is such that $x \succ y$ in which $y \notin ex(P)$ because set $ex(P)$ is the union of maximal and minimal elements. Then, $y \in C$ / $y \leq \max C$ exists.



If $y \neq \max C$, as $\max C$ is not maximal in P, there is $x' \succ \max C$. The induced subposet P', made up of the elements $\{y, \max C, x, x'\}$ verifies that $l(P') = 2$ and is not coherent, in opposition to the hypothesis. Consequently, $y = \max C$.

The reasoning for minimal elements is equivalent to the one above. □

The following theorem is the main result from this research. It establishes alternative characterization for the two $(N, Co(P))$ to be a partition system.

Theorem 3 *Let $P = (N, \leq)$ be a finite poset. The couple $(N, Co(P))$ is a partition system if and only if P is completely coherent and $P \setminus ex(P) = C$ is a chain.*

Proof. $(\Rightarrow)$ Consider that $(N, Co(P))$ is a partition system. We must prove that P is completely coherent and $P \setminus ex(P) = C$.

If $P \setminus ex(P) \neq C$, there are $a, b \in P \setminus ex(P)$ so that $\{a, b\}$ is an antichain. As $\{a, b\} \not\subseteq ex(P)$, consider the sets

$$m(a) = \{m \in P \mid m \prec a\}, \quad M(a) = \{m' \in P \mid a \prec m'\},$$

and, analagously, $m(b)$ and $M(b)$. Obviously, these are not empty sets, and it is easy to notice that $m(a) \cap M(b) = m(b) \cap M(a) = \emptyset$. However, $m(a) \cap m(b)$ and $M(a) \cap M(b)$, these intersections cannot be empty. So, these are the alternatives:

(1) $m(a) \cap m(b) \neq \emptyset$
(2) $M(a) \cap M(b) \neq \emptyset$
(3) $m(a) \cap m(b) = M(a) \cap M(b) = \emptyset$

Using the duality $Co(P) \simeq Co(P^*)$, we only need to pay attention to (1) and (3).
(1) Let $m \in m(a) \cap m(b)$, $m' \in M(a)$. If $b \not\leq m'$ (Figure 6), the set $\{m, b, m'\} \notin Co(P)$ and their maximal convexes $\{\{b, m'\}, \{m, b\}\}$ are not its partition. If $b \leq m'$ (Figure 7), $\{m, a, m'\} \notin Co(P)$ and their maximal convexes $\{\{a, m'\}, \{m, a\}\}$ are also not its partition.

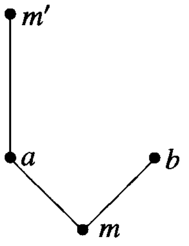


Figure 6

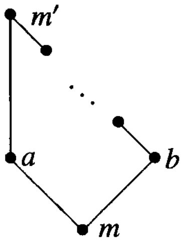


Figure 7

(3) Suppose that $m(a) \cap m(b) = M(a) \cap M(b) = \emptyset$ and let $m \in m(a)$ and $m' \in M(a)$. If there is no connection between b and elements m, m', then $\{m, b, m'\} \notin Co(P)$ and their maximal convexes$\{\{b, m'\}, \{m, b\}\}$ are not its partition (Figure 8). If there was connection it would be because, $m \leq b, b \leq m'$, one or both of them. In every situation, $m \notin m(b)$ and $m' \notin M(b)$ such that $m(a) \cap m(b) = M(a) \cap M(b) = \emptyset$. In all situations, we cannot find convex sets in which their maximal convexes are not a partition. Indeed, if $m \leq b$ there is a b_1 such that $m \leq b_1 \leq b$ (Figure 9) and, for $\{m, a, b\} \notin Co(P)$ their maximal convexes $\{\{m, a\}, \{a, b\}\}$ are not its partition. If $b \leq m'$ the reasoning is equivalent.

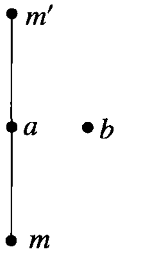


Figure 8

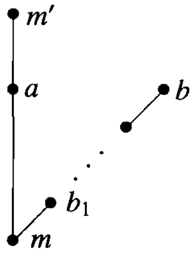


Figure 9

Thus, we have proved that if $P \setminus ex(P) \neq C$ then the hypothesis is not satisfied. Suppose that P is not completely coherent. Then, there is an inducted subset P', with $l(P') \geq 2$ that is not coherent, therefore P' is not connected nor does any maximal element from P' cover any minimal element from P'.

If P' is not connected, there are at least two connected components C_1, C_2 and all of them have to include a chain with length equal or bigger than 2. Suppose $l(C_1) \geq 2$. When we consider the first and last maximal chain C_1 element, indicated by $\{p, u\}$, together with any $a \in C_2$, there is for set $\{p, u, a\}$ the situation is analogue to the subposet in Figure 8, so there is a contradiction.

If P' is connected but any maximal element covers any minimal element, there are m and m' (minimal and maximal from P') such that $m \prec m'$. Nevertheless, that m' covers m in the subposet P' does not imply the same in P. So, there are two possibilities:

(1) $m \prec m'$ in P ($\{m, m'\} \in Co(P)$).
(2) $m \not\prec m'$ in P ($\{m, m'\} \notin Co(P)$).

(1), we consider the set $\{p, u, m, m'\}$ in which p and u are the first and the last elements included in a subposet maximal chain P' ($l(P') \geq 2$). As p and u are extreme elements in P', the three situations shown in Figures 10, 11 and 12 arise. There, $\{p, u, m, m'\} \notin Co(P)$ and its maximal convexes are not its partition. (Notice there is an unknown connection drawn between p and m', as well as between m and u).

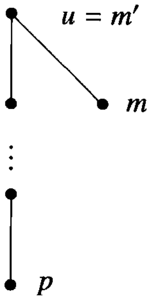


Figure 10

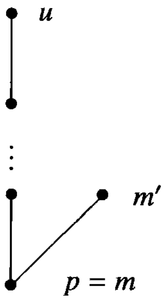


Figure 11

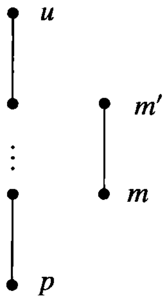


Figure 12

In (2), if $m \nless m'$ there is $p_1 \in P \setminus P'$ such as $m < p_1 < m'$. Let p and u be the first and the last elements from a subposet maximal chain P'. Then, there is $u_1 \in P'$ such as $p < u_1 < u$ ($l(P') \geq 2$). Evidently it cannot be $u = m'$ and $p = m$, because then $m \nless m'$ in P'. Therefore, we must take into consideration the situations in which $u \neq m'$ and $p \neq m$. Because of the duality, it is enough to study one of them. If $u \neq m'$, the situations where it originates (drawn in Figures 13, 14 and 15) are due to $\{u_1, p_1\}$ being an antichain or not.

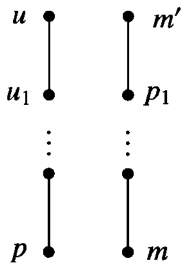


Figure 13

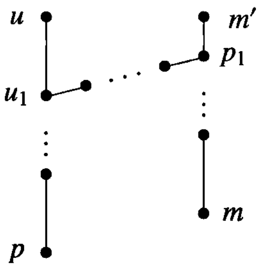


Figure 14

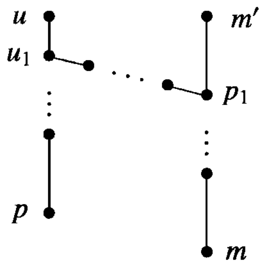


Figure 15

If $\{u_1, p_1\}$ is an antichain, there is a contradicion due to $P \setminus ex(P) \neq C$. If $u_1 < p_1$ or $p_1 < u_1$, we consider the sets $\{u, u_1, m'\} \notin Co(P)$, $\{u, p_1, m'\} \notin Co(P)$. In both cases, their maximal convexes are not their partition.

($\Leftarrow$) Notice that if $P = (N, \leq)$ is a completely coherent finite poset, such that $P \setminus ex(P)$ is a chain C, then $A \in Co(P)$, $B \in Co(P)$ and $A \cap B \neq \emptyset$ imply that $A \cup B \in Co(P)$. The set $A \cup B$ is convex if given $a \in A \cup B$, $b \in A \cup B$ with $a \leq b$, then $[a, b] \subseteq A \cup B$. The set $A \cup B$ is a disjoint union of $A \setminus B$, $A \cap B$ and $B \setminus A$; so, among the different possible alternatives for a and b, we only need to analyze a couple of them: $a \in A \setminus B$ and $b \in B \setminus A$, or $a \in B \setminus A$ and $b \in A \setminus B$. Furthermore, using the duality ($Co(P) \simeq Co(P^*)$), it is enough to analyze only one of the possibilities. Consequently, let $a \in A \setminus B$, $b \in B \setminus A$ with $a < b$.

It must be proved that $[a, b] \subseteq A \cup B$ and, by hypothesis, $A \cap B \neq \emptyset$. If there is an element $d \in A \cap B$ such that $d \in [a, b]$, then:

$$[a, b] = [a, d] \cup [d, b] \subseteq A \cup B,$$

as the intervals of P are always chains ($P \setminus ex(P) = C$), $\{a, d\} \subseteq A$, $\{d, b\} \subseteq B$ and $A, B \in Co(P)$.

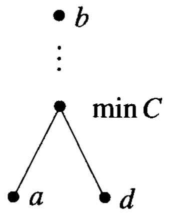
In the case that any $d \in A \cap B$ is not included in the interval $[a, b]$, the are four possible alternatives: 1) $d < a$, 2) $b < d$, 3) $\{a, d\}$ is an antichain and 4) $\{b, d\}$ is an antichain.

We are going to analyze:

(1) If $d < a < b$, then $[d, b] \subseteq B$. Therefore $a \in B$, instead of being $a \in A \setminus B$.

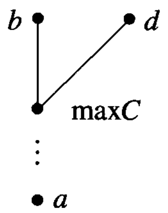
(2) If $a < b < d$, then $[a, d] \subseteq A$. Therefore, $b \in A$ which contradicts $b \in B \setminus A$.

(3) If $\{a, d\}$ is an antichain, then a and d are minimal elements (a is not maximal due to $a < b$ and the only possible antichains in P are made of maximal elements from P or of minimal elements).



Theorem 2 implies that minimal element from chain C, $\min C$ covers a and d. Then $d \prec \min C \leq b$ and $[d, b] \subseteq B$ as B is convex and $\{d, b\} \subseteq B$. Therefore,

$$[a, b] = \{a\} \cup [\min C, b] \subseteq \{a\} \cup [d, b] \subseteq A \cup B.$$



(4) Using an analogous reasoning, if $\{b, d\}$ is an antichain, both are maximals and it is deduced that $d \succ \max C \geq a$. Then, $[a, d] \subseteq A$ y

$$[a, b] = [a, \max C] \cup \{b\} \subseteq A \cup B.$$

□

Obviously, the results above have a theoretical interest. The knowledge of convex sets, and particulary those structures that lead to partition systems, have a practical interest, among other possibilities, in order to estimate power indexes —both Banzhaf's and Shapley's— in simple weighted voting games and in double-triple majority games, in which cooperation is restricted to a feasible coalition set. This application is discussed in more detail by Bilbao, Jiménez, López and Fernández (2000).

3 Bibliography

Bergantiños, G., Carreras, F. and García-Jurado, I. (1993). Cooperation when some players are incompatible. *ZOR-Methods and Models of Operations Research*, 38, 187-201.

Bilbao, J. M., Fernández, J., Jiménez, N. and López, J. (2000). Voting power in the European Union enlargement. *European Journal of Operational Research*, 143, 181-196.

Bilbao, J. M., Fernández, J. and López, J. (2003). Computing power indices in weighted multiple majority games. *Mathematical Social Science*. Accepted for publication.

Birkhoff, G. and Bennett, M. K. (1985). The convexity lattice of a poset. *Order*, 2, 223-242.

Borm, P., Owen, G. and Tijs, S. (1992). The position value for communication situations. *SIAM Journal on Discrete Mathematics*, 5, 305-320.

Borm, P., Nouweland, A., Owen, G. and Tijs, S. (1993). Cost allocation and communication. *Naval Research Logistics*, 40, 733-744.

Carreras, F. (1991). Restriction of simple games. *Mathematical Social Sciences*, 21, 245-260.

Edelman, P. H. and Jamison, R. E. (1985). The theory of convex geometries. *Geometriæ Dedicata*, 19, 247-270.

Edelman, P. H. (1996). A note on voting, *preprint*.

Faigle, U. (1989). Cores of games with restricted cooperation. *ZOR-Methods and Models of Operations Research*, 33, 405-402.

Myerson, R. B. (1977). Graphs and cooperation in games. *Mathematics of Operations Research*, 2, 225-229.

Owen, G. (1986). Values of graph-restricted games. *SIAM Journal of Algebraic and Discrete Methods*, 7, 210-220.

Stanley, R. P. (1986). *Enumerative Combinatorics*, vol. I, Wadsworth.

Tijs, S. and Otten, G. (1993). Compromise values in cooperative game theory. *TOP (Trabajos de Investigación Operativa)*, 1, 1-51.

Resum

Considerem jocs d'utilitat transferible que tracten amb situacions de cooperació parcial constituïdes per sistemes de coalicions, en els que tota coalició unitària és factible i tota coalició de jugadors es pot expressar com una unió disjunta de coalicions factibles maximals. Aquests sistemes reben el nom de sistemes de partició i donen lloc a jocs restringits. En particular, estudiem sistemes de coalició definits per un ordre parcial establert en el conjunt dels jugadors i analitzem les característiques de coalicions factibles construït a partir de la classe de conjunts convexos.

MSC: 90D12

Paraules clau: Jocs cooperatius, cooperació parcial, conjunts convexos

SORT 27 (2) July-December 2003, 153-164

On two matrix derivatives by Kollo and von Rosen

Heinz Neudecker*

University of Amsterdam, The Netherlands

Abstract

The article establishes relationships between the matrix derivatives of F with respect to X as introduced by von Rosen (1988), Kollo and von Rosen (2000) and the Magnus-Neudecker (1999) matrix derivative. The usual transformations apply and the Moore-Penrose inverse of the duplication matrix is used. Both X and F have the same dimension.

MSC: primary 15A69, secondary 62H10

Keywords: Matrix derivatives and differentials, vectorization, commutation and duplication matrices

1 Introduction

Von Rosen (1988) and Kollo and von Rosen (2000) study moments of the inverted Wishart distribution. For finding specific expressions they use two types of matrix derivatives. Unfortunately these are not easily accessible to the uninitiated reader. The two are obviously related, both being matrix representations of the Fréchet derivative. There is, however, a more accessible representation, namely the Magnus-Neudecker matrix derivative. In this article we shall link the three representations and consider some illustrative applications, most of these taken from the two quoted articles.

* *Address for correspondence:* Oosterstraat, 13. 1741 GH Schagen. The Netherlands. Postal address: NL 1741 GH 13. E-mail: heinz@fee.uva.nl

Received: April 2002
Accepted: June 2003

2 The three matrix derivatives

Von Rosen (1988) defined the matrix derivative

$$\frac{\partial F}{\partial X} = \sum_{ijkl} \varepsilon_{kl} \frac{\partial f_{ij}}{\partial x_{kl}} E_{ij} \otimes E_{kl} \tag{1}$$

where

$$\varepsilon_{kl} = \begin{cases} 1 & \text{if } k = l \\ \frac{1}{2} & \text{if } k \neq l, \end{cases} \qquad E_{ij} = e_i e_j'.$$

where e_i is the i^{th} column of the identity matrix I_p, $F = \left(f_{ij}\right)$ and $X = (x_{kl})$ are *symmetric* matrices of dimension p, and $i, j, k, l = 1, \ldots, p$. $F = F(X)$ is a function of X.

Kollo and von Rosen (2000) defined the matrix derivative

$$\frac{dF}{dX} = \sum_{ijkl} \varepsilon_{kl} \frac{\partial f_{ij}}{\partial x_{kl}} \left(\text{vec}\, E_{ij}\right) (\text{vec}\, E_{kl})'. \tag{2}$$

It is clear that

$$K_{p^2,p^2}\, \text{vec} \frac{dF}{dX} = C_2^p\, \text{vec} \frac{\partial F}{\partial X} \tag{3}$$

with $C_2^p = I_p \otimes K_{pp} \otimes I_p$, K denoting a commutation matrix. We used property (*ii*) of the Appendix.

Therefore

$$\text{vec} \frac{\partial F}{\partial X} = \left(K_{pp} \otimes K_{pp}\right) C_2^p\, \text{vec} \frac{dF}{dX}, \tag{4}$$

because

$$K_{p^2,p^2} = C_2^p \left(K_{pp} \otimes K_{pp}\right) C_2^p.$$

See, e.g. Ghazal and Neudecker (2000) for properties of C_2^p. Some will be reported in the Appendix.

In Section 3 we shall establish the identity

$$\frac{dF}{dX} = D_p \frac{\partial f}{\partial x'} D_p^+ \tag{5}$$

where $\frac{\partial f}{\partial x'}$ is the Magnus-Neudecker matrix derivative and D_p^+ is the Moore-Penrose inverse of the duplication matrix D_p. Further $f = D_p^+\, \text{vec}\, F$ and $x = D_p^+\, \text{vec}\, X$. Equivalently $f = v(F)$ and $x = v(X)$.

The combination of (4) and (5) would enable us to express $\frac{\partial F}{\partial X}$ in $\frac{\partial f}{\partial x'}$, in vectorized form. As this is not so fruitful, we shall not do it. We prefer to use (4) and subsequently devectorize the resulting $\text{vec} \frac{\partial F}{\partial X}$.

3 The link between $\frac{dF}{dX}$ and $\frac{\partial f}{\partial x'}$

The first thing to do is to establish

Lemma 1

$$K_{pp}\frac{d\,F}{d\,X} = \frac{d\,F}{d\,X}, \quad \frac{d\,F}{d\,X}K_{pp} = \frac{dF}{dX}. \tag{6}$$

Proof. We have $K_{pp}(e_j \otimes e_i) = e_i \otimes e_j$. Hence

$$K_{pp}\frac{\partial f_{ij}}{\partial x_{kl}}\left(e_j \otimes e_i\right) = \frac{\partial f_{ij}}{\partial x_{kl}}\left(e_i \otimes e_j\right) = \frac{\partial f_{ji}}{\partial x_{kl}}\left(e_i \otimes e_j\right),$$

and

$$K_{pp}\sum_{ij}\frac{\partial f_{ij}}{\partial x_{kl}}\left(e_j \otimes e_i\right) = \sum_{ij}\frac{\partial f_{ji}}{\partial x_{kl}}\left(e_i \otimes e_j\right) = \sum_{ij}\frac{\partial f_{ij}}{\partial x_{kl}}\left(e_j \otimes e_i\right)$$

by interchanging the indices i and j. Hence $K_{pp}\frac{dF}{dX} = \frac{dF}{dX}$. Similarly we find that

$$\frac{dF}{dX}K_{pp} = \frac{dF}{dX}. \qquad \square$$

Having found these basic properties of $\frac{dF}{dX}$ we shall prove

Lemma 2

$$\frac{dF}{dX}d\,vec\,X = d\,vec\,F.$$

Proof. Using the definition of $\frac{dF}{dX}$ we get

$$\begin{aligned}
\frac{dF}{dX} d\operatorname{vec} X &= \sum_{ijkl} \varepsilon_{kl} \frac{\partial f_{ij}}{\partial x_{kl}} (e_j \otimes e_i)(e_l \otimes e_k)' d\operatorname{vec} X \\
&= \sum_{ijkl} \varepsilon_{kl} \frac{\partial f_{ij}}{\partial x_{kl}} \left(e_j e_l' \otimes e_i e_k'\right) d\operatorname{vec} X \\
&= \sum_{ijkl} \varepsilon_{kl} \frac{\partial f_{ij}}{\partial x_{kl}} d\operatorname{vec} e_i e_k' X e_l e_j' \\
&= \sum_{ijkl} \varepsilon_{kl} \left(\frac{\partial f_{ij}}{\partial x_{kl}} d x_{kl}\right) \operatorname{vec} e_i e_j' \\
&= \sum_{ijkl} \varepsilon_{kl} \frac{\partial \operatorname{vec} f_{ij} e_i e_j'}{\partial x_{kl}} d x_{kl} \\
&= \sum_{kl} \varepsilon_{kl} \frac{\partial \operatorname{vec} F}{\partial x_{kl}} d x_{kl} = d\operatorname{vec} F.
\end{aligned}$$ □

Having established this result we shall prove

Theorem 3

$$D_p \frac{\partial f}{\partial x'} = \frac{dF}{dX} D_p.$$

Proof. We rewrite the result of lemma 2 as

$$\frac{dF}{dX} D_p dx = D_p df = D_p \frac{\partial f}{\partial x'} dx.$$

Omitting the arbitrary dx we obtain the result. □

The basic result follows as a corollary, namely.

Corollary 4

$$\frac{dF}{dX} = D_p \frac{\partial f}{\partial x'} D_p^+.$$

Proof. Postmultiplication of the result of theorem 3 by D_p^+ yields, by virtue of lemma 1,

$$D_p \frac{\partial f}{\partial x'} D_p^+ = \frac{dF}{dX} D_p D_p^+ = \frac{1}{2} \frac{dF}{dX} \left(I_{p^2} + K_{pp}\right) = \frac{dF}{dX}.$$

□

4 Some applications of $\frac{dF}{dX}$

We shall examine three cases.

(1) Kollo & von Rosen (2000) find

$$\frac{dF}{dX} = \frac{1}{2}\left(I_{p^2} + K_{pp}\right),$$

for $F = X$ (their result 2.9).

This can be derived succinctly in the following way. Differentiation of F yields $dF = dX$, which leads to $d\,\mathrm{vec}\,F = d\,\mathrm{vec}\,X$ and subsequently to $df = dx$. Hence $\frac{dF}{dX} = D_p D_p^+ = \frac{1}{2}\left(I_{p^2} + K_{pp}\right)$ by corollary 4.

□

(2) $$\frac{dF}{dX} = -\tfrac{1}{2}\left(I_{p^2} + K_{pp}\right)\left(X^{-1} \otimes X^{-1}\right) \text{ for } F = X^{-1}.$$

Proceeding as before we get

$$\begin{aligned} dF &= -X^{-1}(dX)X^{-1}, \\ d\,\mathrm{vec}\,F &= -\left(X^{-1} \otimes X^{-1}\right) d\,\mathrm{vec}\,X, \\ df &= -D_p^+\left(X^{-1} \otimes X^{-1}\right) D_p\, dx \quad \text{and} \\ \frac{\partial f}{\partial x'} &= -D_p^+\left(X^{-1} \otimes X^{-1}\right) D_p. \end{aligned}$$

Corollary 4 yields then

$$\begin{aligned} \frac{dF}{dX} &= -D_p D_p^+\left(X^{-1} \otimes X^{-1}\right) D_p D_p^+ = -\tfrac{1}{4}\left(I_{p^2} + K_{pp}\right)\left(X^{-1} \otimes X^{-1}\right)\left(I_{p^2} + K_{pp}\right) = \\ &= -\tfrac{1}{4}\left(I_{p^2} + K_{pp}\right)^2\left(X^{-1} \otimes X^{-1}\right) = -\tfrac{1}{2}\left(I_{p^2} + K_{pp}\right)\left(X^{-1} \otimes X^{-1}\right). \end{aligned}$$

□

(3) $$\frac{dF}{dX} = \tfrac{1}{2}\left(I_{p^2} + K_{pp}\right)\left(I_p \otimes X + X \otimes I_p\right) \text{ for } F = X^2.$$

We get

$$\begin{aligned} dF &= (dX)X + X\,dX, \\ d\,\mathrm{vec}\,F &= \left(I_p \otimes X + X \otimes I_p\right) d\,\mathrm{vec}\,X, \\ df &= D_p^+\left(I_p \otimes X + X \otimes I_p\right) D_p\, d\,x. \end{aligned}$$

Hence

$$\begin{aligned}\frac{\partial f}{\partial x'} &= D_p^+\left(I_p \otimes X + X \otimes I_p\right) D_p,\\ \frac{dF}{dX} &= D_p D_p^+\left(I_p \otimes X + X \otimes I_p\right) D_p D_p^+ =\\ &= \tfrac{1}{4}\left(I_{p^2} + K_{pp}\right)\left(I_p \otimes X + X \otimes I_p\right)\left(I_{p^2} + K_{pp}\right) =\\ &= \tfrac{1}{4}\left(I_{p^2} + K_{pp}\right)^2\left(I_p \otimes X + X \otimes I_p\right) =\\ &= \tfrac{1}{2}\left(I_{p^2} + K_{pp}\right)\left(I_p \otimes X + X \otimes I_p\right).\end{aligned}$$

□

In the following section we shall give some applications of $\frac{\partial F}{\partial X}$.

5 Some applications of $\frac{\partial F}{\partial X}$

(1) $\frac{\partial X}{\partial X} = \tfrac{1}{2}\left(\operatorname{vec} I_p\right)\left(\operatorname{vec} I_p\right)' + \tfrac{1}{2} K_{pp}$.

See von Rosen (1988, Lemma 2.1, *d, i*).

Derivation: Use $\frac{dX}{dX} = \tfrac{1}{2}\left(I_{p^2} + K_{pp}\right)$.

See section 4 (1). It is easy to see that $C_2^p \operatorname{vec} I_{p^2} = C_2^p \operatorname{vec}\left(I_p \otimes I_p\right) = \operatorname{vec} I_p \otimes \operatorname{vec} I_p$. Further $C_2^p \operatorname{vec} K_{pp} = \operatorname{vec} K_{pp}$. For these properties we refer to (*ii*) and (*iii*) in the Appendix.

Hence

$$\begin{aligned}\operatorname{vec}\frac{\partial X}{\partial X} &= \tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right)\left(\operatorname{vec} I_p \otimes \operatorname{vec} I_p + \operatorname{vec} K_{pp}\right)\\ &= \tfrac{1}{2}\left(\operatorname{vec} I_p \otimes \operatorname{vec} I_p + \operatorname{vec} K_{pp}\right)\\ &= \tfrac{1}{2}\operatorname{vec}\left[\left(\operatorname{vec} I_p\right)\left(\operatorname{vec} I_p\right)' + K_{pp}\right],\end{aligned}$$

from which (1) follows.

□

(2) $\frac{\partial X^{-1}}{\partial X} = -\tfrac{1}{2}\left(\operatorname{vec} X^{-1}\right)\left(\operatorname{vec} X^{-1}\right)' - \tfrac{1}{2} K_{pp}\left(X^{-1} \otimes X^{-1}\right)$.

See von Rosen (1988, Lemma 2.2*i*).

Derivation: Use $\dfrac{dX^{-1}}{dX} = -\frac{1}{2}\left(I_{p^2} + K_{pp}\right)\left(X^{-1} \otimes X^{-1}\right)$.

See section 4 (2). Then

$$\begin{aligned}
\operatorname{vec}\frac{\partial X^{-1}}{\partial X} &= -\tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right) C_2^p \operatorname{vec}\left[\left(I_{p^2} + K_{pp}\right)\left(X^{-1} \otimes X^{-1}\right)\right] \\
&= -\tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right)\left(\operatorname{vec} X^{-1} \otimes \operatorname{vec} X^{-1}\right) \\
&\quad -\tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right) \operatorname{vec} K_{pp}\left(X^{-1} \otimes X^{-1}\right) \\
&= -\tfrac{1}{2} \operatorname{vec} X^{-1} \otimes \operatorname{vec} X^{-1} - \tfrac{1}{2} \operatorname{vec} K_{pp}\left(X^{-1} \otimes X^{-1}\right) \\
&= -\tfrac{1}{2} \operatorname{vec}\left[\left(\operatorname{vec} X^{-1}\right)\left(\operatorname{vec} X^{-1}\right)'\right] - \tfrac{1}{2} \operatorname{vec} K_{pp}\left(X^{-1} \otimes X^{-1}\right),
\end{aligned}$$

from which (2) follows.

We also used property 18 in Neudecker (2000). For a simple proof see (*iv*) in the Appendix.

□

(3) $\dfrac{\partial X^2}{\partial X} = \frac{1}{2} K_{pp}\left(I_p \otimes X + X \otimes I_p\right) + \frac{1}{2}(\operatorname{vec} X)\left(\operatorname{vec} I_p\right)' + \frac{1}{2}\left(\operatorname{vec} I_p\right)(\operatorname{vec} X)'$.

Derivation: Use $\dfrac{dX^2}{dX} = \frac{1}{2}\left(I_{p^2} + K_{pp}\right)\left(I_p \otimes X + X \otimes I_p\right)$.

See section 4 (3). Then

$$\begin{aligned}
\operatorname{vec}\frac{\partial X^2}{\partial X} &= \tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right) C_2^p \operatorname{vec}\left[\left(I_{p^2} + K_{pp}\right)\left(I_p \otimes X + X \otimes I_p\right)\right] \\
&= \tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right)\left(\operatorname{vec} I_p \otimes \operatorname{vec} X + \operatorname{vec} X \otimes \operatorname{vec} I_p\right) \\
&\quad + \tfrac{1}{2}\left(K_{pp} \otimes K_{pp}\right) \operatorname{vec}\left[K_{pp}\left(I_p \otimes X + X \otimes I_p\right)\right] \\
&= \tfrac{1}{2}\left(\operatorname{vec} I_p \otimes \operatorname{vec} X + \operatorname{vec} X \otimes \operatorname{vec} I_p\right) + \tfrac{1}{2} \operatorname{vec}\left[K_{pp}\left(I_p \otimes X + X \otimes I_p\right)\right] \\
&= \tfrac{1}{2} \operatorname{vec}\left[(\operatorname{vec} X)\left(\operatorname{vec} I_p\right)' + \left(\operatorname{vec} I_p\right)(\operatorname{vec} X)'\right] \\
&\quad + \tfrac{1}{2} \operatorname{vec}\left[K_{pp}\left(I_p \otimes X + X \otimes I_p\right)\right].
\end{aligned}$$

Devectorization yields (3).

□

(4) $\dfrac{\partial F^{-1}}{\partial X} = -\left(F^{-1} \otimes I_p\right)\dfrac{\partial F}{\partial X}\left(F^{-1} \otimes I_p\right)$, where $F = F(X)$.

See von Rosen (1988, Lemma 2.1, *c*, *iii*).

Derivation: It is known that $dF^{-1} = -F^{-1}(dF)F^{-1}$, hence

$$\begin{aligned} dv\left(F^{-1}\right) &= D_p^+ d\,\mathrm{vec}\,F^{-1} = -D_p^+\left(F^{-1}\otimes F^{-1}\right)d\,\mathrm{vec}\,F \\ &= -D_p^+\left(F^{-1}\otimes F^{-1}\right)D_p\,d\,v(F) = -D_p^+\left(F^{-1}\otimes F^{-1}\right)D_p\frac{\partial f}{\partial x'}dx \end{aligned}$$

and finally

$$\frac{\partial v\left(F^{-1}\right)}{\partial x'} = -D_p^+\left(F^{-1}\otimes F^{-1}\right)D_p\frac{\partial f}{\partial x'}.$$

By Corollary 4 and Lemma 1 we have

$$\begin{aligned} \frac{dF^{-1}}{dX} &= -D_pD_p^+\left(F^{-1}\otimes F^{-1}\right)D_p\frac{\partial f}{\partial x'}D_p^+ \\ &= -\tfrac{1}{2}\left(I_{p^2}+K_{pp}\right)\left(F^{-1}\otimes F^{-1}\right)\frac{dF}{dX} \\ &= -\tfrac{1}{2}\left(F^{-1}\otimes F^{-1}\right)\left(I_{p^2}+K_{pp}\right)\frac{dF}{dX} \\ &= -\left(F^{-1}\otimes F^{-1}\right)\frac{dF}{dX}. \end{aligned}$$

Application of Section 2 (4) yields

$$\begin{aligned} \mathrm{vec}\,\frac{\partial F^{-1}}{\partial X} &= -\left(K_{pp}\otimes K_{pp}\right)C_2^p\,\mathrm{vec}\left(F^{-1}\otimes F^{-1}\right)\frac{dF}{dX} \\ &= -\left(K_{pp}\otimes K_{pp}\right)C_2^p\left(I_{p^2}\otimes F^{-1}\otimes F^{-1}\right)\mathrm{vec}\,\frac{dF}{dX} \\ &= -\left(K_{pp}\otimes K_{pp}\right)C_2^p\left(I_p\otimes I_p\otimes F^{-1}\otimes F^{-1}\right)\mathrm{vec}\frac{dF}{dX} \\ &= -\left(K_{pp}\otimes K_{pp}\right)\left(I_p\otimes F^{-1}\otimes I_p\otimes F^{-1}\right)C_2^p\,\mathrm{vec}\frac{dF}{dX} \\ &= -\left(F^{-1}\otimes I_p\otimes F^{-1}\otimes I_p\right)\left(K_{pp}\otimes K_{pp}\right)C_2^p\,\mathrm{vec}\frac{dF}{dX} \\ &= -\left(F^{-1}\otimes I_p\otimes F^{-1}\otimes I_p\right)\mathrm{vec}\frac{\partial F}{\partial X} \\ &= -\mathrm{vec}\left[\left(F^{-1}\otimes I_p\right)\frac{\partial F}{\partial X}\left(F^{-1}\otimes I_p\right)\right]. \end{aligned}$$

Hence

$$\frac{\partial F^{-1}}{\partial X} = -\left(F^{-1}\otimes I_p\right)\frac{\partial F}{\partial X}\left(F^{-1}\otimes I_p\right).$$

□

6 Appendix

The following matrix properties have been used in this article. The first five involve $C_2^p = I_p \otimes K_{pp} \otimes I_p$.

(*i*) $C_2^p (A \otimes B \otimes C \otimes D) C_2^p = A \otimes C \otimes B \otimes D$, for $(p \times p) A, B, C, D$.

(*ii*) $C_2^p \operatorname{vec}(A \otimes B) = \operatorname{vec} A \otimes \operatorname{vec} B$, for $(p \times p) A$ and B.

Proof.

$$\begin{aligned}
C_2^p \operatorname{vec}(A \otimes B) &= \sum_{ij} \left(I_p \otimes E_{ij} \otimes E_{ji} \otimes I_p\right) \operatorname{vec}(A \otimes B) \\
&= \sum_{ij} \operatorname{vec}\left[\left(E_{ji} \otimes I_p\right)(A \otimes B)\left(I_p \otimes E_{ji}\right)\right] \\
&= \sum_{ij} \operatorname{vec}\left(E_{ji} A \otimes B E_{ji}\right) = \sum_{ij} \operatorname{vec}\left(e_j A_{i.} \otimes B_{.j} e_i'\right) \\
&= \sum_{ij} \operatorname{vec}\left[\left(e_j \otimes B_{.j}\right)(A_{i.} \otimes e_i')\right] = \\
&= \operatorname{vec}\left[\left(\sum_j \operatorname{vec} B_{.j} e_j'\right)\left(\sum_i \operatorname{vec} e_i A_{i.}\right)'\right] \\
&= \operatorname{vec}\left[(\operatorname{vec} B)(\operatorname{vec} A)'\right] = \operatorname{vec} A \otimes \operatorname{vec} B.
\end{aligned}$$

As usual $A_{i.}$ is the i^{th} row of A, $A_{.j}$ is the j^{th} column of A.

□

(*iii*) $C_2^p \operatorname{vec} K_{pp} = \operatorname{vec} K_{pp}$.

Proof.

$$\begin{aligned}
C_2^p \operatorname{vec} K_{pp} &= \sum_{ij} C_2^p \operatorname{vec}\left(E_{ij} \otimes E_{ji}\right) \\
&= \sum_{ij} \left(\operatorname{vec} E_{ij} \otimes \operatorname{vec} E_{ji}\right) = \operatorname{vec} \sum_{ij} \left(\operatorname{vec} E_{ji}\right)\left(\operatorname{vec} E_{ij}\right)' \\
&= \operatorname{vec} \sum_{ij} \left(e_i \otimes e_j\right)\left(e_j' \otimes e_i'\right) = \operatorname{vec} \sum_{ij} \left(e_i e_j' \otimes e_j e_i'\right) \\
&= \operatorname{vec} \sum_{ij} \left(E_{ij} \otimes E_{ji}\right) = \operatorname{vec} K_{pp}.
\end{aligned}$$

□

(*iv*) $C_2^p \operatorname{vec} K_{pp} (A \otimes B) = \operatorname{vec} K_{pp} (A \otimes B')$ for $(p \times p) A$ and B.

Proof.

$$
\begin{aligned}
C_2^p \operatorname{vec} K_{pp}(A \otimes B) &= C_2^p \left(A' \otimes B' \otimes I_p \otimes I_p\right) \operatorname{vec} K_{pp} \\
&= \left(A' \otimes I_p \otimes B' \otimes I_p\right) C_2^p \operatorname{vec} K_{pp} = \left(A' \otimes I_p \otimes B' \otimes I_p\right) \operatorname{vec} K_{pp} \\
&= \operatorname{vec}\left(B' \otimes I_p\right) K_{pp} \left(A \otimes I_p\right) = \operatorname{vec} K_{pp} \left(I_p \otimes B'\right)\left(A \otimes I_p\right) \\
&= \operatorname{vec} K_{pp}(A \otimes B') .
\end{aligned}
$$

□

(v) $K_{p^2,p^2} = C_2^p \left(K_{pp} \otimes K_{pp}\right) C_2^p$.

Proof. Consider the $\left(p^2,\, p^2\right)$ matrix

$$Q = \sum_{kl} (E_{kl} \otimes Q_{kl})$$

or equivalently

$$Q = [Q_{kl}] \quad (k, l = 1, \ldots, p)$$

with $(p \times p)$ matrix Q_{kl}.

Then $K_{p^2,p^2} \operatorname{vec} Q = \operatorname{vec} Q'$ and

$$
\begin{aligned}
C_2^p \left(K_{pp} \otimes K_{pp}\right) C_2^p \operatorname{vec} Q &= \sum_{ijst} C_2^p \left(E_{ij} \otimes E_{ji} \otimes E_{st} \otimes E_{ts}\right) C_2^p \operatorname{vec} Q \\
&= \sum_{ijstkl} \left(E_{ij} \otimes E_{st} \otimes E_{ji} \otimes E_{ts}\right) \operatorname{vec}(E_{kl} \otimes Q_{kl}) \\
&= \operatorname{vec} \sum_{ijstkl} \left(E_{ji} \otimes E_{ts}\right) (E_{kl} \otimes Q_{kl}) \left(E_{ji} \otimes E_{ts}\right) \\
&= \operatorname{vec} \sum_{ijstkl} \left(E_{ji} E_{kl} E_{ji} \otimes E_{ts} Q_{kl} E_{ts}\right) \\
&= \operatorname{vec} \sum_{ijst} \left(E_{ji} \otimes E_{ts} Q_{ij} E_{ts}\right) = \operatorname{vec} \sum_{ijst} \left(Q_{ij}\right)_{st} \left(E_{ij} \otimes E_{st}\right)' \\
&= \operatorname{vec} \sum_{ij} \left(E_{ij} \otimes Q_{ij}\right)' = \operatorname{vec} Q' ,
\end{aligned}
$$

where $\left(Q_{ij}\right)_{st}$ is the (s, t) element of Q_{ij}.

□

We further used the standard properties:

(vi) $K_{pp} \operatorname{vec} A = \operatorname{vec} A'$ for $(p \times p)$ matrix A.

(vii) $D_p D_p^+ = \frac{1}{2}\left(I_{p^2} + K_{pp}\right)$, $D_p^+ D_p = I_{p^*}$ with $2p^* = p(p+1)$.

$(viii)$ $\text{vec}\, A\, B\, C = (C' \otimes A)$ $\text{vec}\, B$ for compatible matrices A, B and C.

7 References

Ghazal, G. A. and Neudecker, H. (2000). On second-order and fourth-order moments of jointly distributed random matrices: a survey. *Linear Algebra and its Applications*, 321, 61-93.

Kollo, T. and von Rosen, D. (2000). Distribution and density approximation of the covariance matrix in the growth curve model. *Statistics*, 35, 1-22.

Magnus, J. R. and Neudecker, H. (1980). The elimination matrix: some lemmas and applications. *SIAM Journal on Algebraic and Discrete Methods*, 1, 422-449.

Magnus, J. R. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. John Wiley, Chichester, England.

Neudecker, H. (2000). On expected values of fourth-degree matrix products of a multinormal matrix variate. *New Trends in Probability and Statistics. Proceedings of the 6th Tartu Conference on Multivariate Statistics.* TEV Vilnius/Utrecht.

Neudecker, H. and Wansbeek, T.
(1983). Some results on commutation matrices with statistical applications. *Canadian Journal of Statistics*, 11, 221-31.

von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, 15, 97-109.

Resum

Aquest article estableix les relacions entre las derivades matricials de *F* respecte de *X* introduïdes per von Rosen (1988), Kollo i von Rosen (2000) i les derivades matricials de Magnus i Neudecker (1999). Les operacions vectorials de duplicació i transformació en vectors són les usuals i les inverses de las matrius duplicades són les de Moore-Penrose. Ambdues *X* i *F = F(X)* tenen la mateixa dimensió.

MSC: primary 15A69, secondary 62H10

Paraules clau: Derivades i diferencials de matrius, vectorització, commutació i duplicació de matrius

SORT 27 (2) July-December 2003, 165-174

Asymptotic study of canonical correlation analysis: from matrix and analytic approach to operator and tensor approach

Jeanne Fine*

Université Paul Sabatier, France

Abstract

Asymptotic study of canonical correlation analysis gives the opportunity to present the different steps of an asymptotic study and to show the interest of an operator and tensor approach of multidimensional asymptotic statistics rather than the classical, matrix and analytic approach. Using the last approach, Anderson (1999) assumes the random vectors to have a normal distribution and the non zero canonical correlation coefficients to be distinct. The new approach we use, Fine (2000), is coordinate-free, distribution-free and permits to have no restriction on the canonical correlation coefficients multiplicity order. Of course, when vectors have a normal distribution and when the non zero canonical correlation coefficients are distinct, it is possible to find again Anderson's results but we diverge on two of them. In this methodological presentation, we insist on the analysis frame (Dauxois and Pousse, 1976), the sampling model (Dauxois, Fine and Pousse, 1979) and the different mathematical tools (Fine, 1987, Dauxois, Romain and Viguier, 1994) which permit to solve problems encountered in this type of study, and even to obtain asymptotic behavior of the analyses random elements such as principal components and canonical variables.)

MSC: 62E20, 62H20, 62H25, 47N30

Keywords: multivariate analysis, canonical correlation analysis, asymptotic study, operator, coordinate-free, distribution-free

1 Classical approach

1.1 Population canonical correlation analysis

Let X and Y be two random vectors, p and q dimensional respectively ($p \leq q$) defined on a same probability space $(\Omega, \mathcal{A}, P)$, centered and admitting order 4 moments. We

* *Address for correspondence:* Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex, France. e-mail: fine@cict.fr.

Received: October 2003
Accepted: December 2003

assume the matrix covariance V_X of X to be non-singular and we denote by H_X the vector space of real-valued random variables (r.r.v.) linear combinations of X components. We introduce similarly V_Y and H_Y and we denote by V_{XY} the cross covariance the components of X with the ones of Y.

The aim of canonical correlation analysis (CCA) of (X, Y) is to measure the relationships between X and Y. CCA may be defined as the search for f_1 and g_1, r.r.v. of H_X and H_Y with unit variance and maximal correlation ρ_1 then, iteratively, for $j = 2, ..., r$, $(r \leq p)$, as the search of f_j and g_j, r.r.v. of H_X and H_Y with unit variance, uncorrelated with the $(f_k)_{k<j}$ and $(g_k)_{k<j}$ and with maximal correlation ρ_j. The r.r.v. f_j and g_j are called *j^{th} canonical variables* and the real ρ_j of [0,1] is called *j^{th} canonical correlation coefficient.*

Let $R_X = V_X^{-\frac{1}{2}} V_{XY} V_Y^{-1} V_{YX} V_X^{-\frac{1}{2}}$ and the same for R_Y permuting X and Y roles. It is easy to verify that R_X and R_Y have the same non-zero eigenvalues denoted by $(\rho_j^2)_{j=1,...,r}$ when written in a decreasing order. We set : $\lambda_j = \rho_j^2$, for $j = 1, ..., r$, and, except in the particular case $r = p = q$, we set $\lambda_j = \rho_j = 0$ for $j > r$. For $j > r$ we define f_j and g_j as r.r.v. of H_X and H_Y respectively with unit variance and uncorrelated with the $(f_k)_{k<j}$ and $(g_k)_{k<j}$.

CCA of (X, Y) is then :

$$((\rho_j)_{j=1,...,r+1}, (f_j)_{j=1,...,p}, (g_j)_{j=1,...,q}). \tag{1}$$

CCA of (X, Y) depends only on H_X and H_Y, which are also generated by components of $X' := V_X^{-\frac{1}{2}} X$ and $Y' := V_Y^{-\frac{1}{2}} Y$ respectively. We show that, if $(u_j)_{j=1,...,p}$ and $(v_j)_{j=1,...,q}$ denote unit eigenvectors bases of R_X and R_Y associated with $(\lambda_j)_{j=1,...,p}$ and $(\lambda_j)_{j=1,...,q}$ respectively, we can obtain canonical variables f_j and g_j as linear combinations of X' and Y' components, using u_j and v_j components as coefficients, that is, by setting: $f_j = \langle u_j, X' \rangle_p$ et $g_j = \langle v_j, Y' \rangle_q$, where $\langle ., . \rangle_p$ and $\langle ., . \rangle_q$ denote $\mathbb{R}^p$ and $\mathbb{R}^q$ usual scalar products.

Decomposition (1) is not unique because each canonical variable associated with a simple eigenvalue may be replaced by its opposite and the set of canonical variables associated with a multiple eigenvalue may be replaced by an other set according to the choice of R_X and R_Y eigenvectors associated with this eigenvalue.

1.2 Sample canonical correlation analysis

Let $(X_l, Y_l)_{l=1,...,n}$ be a n-sample i.i.d. as (X, Y). We index by n the elements defined previously and calculated on the sample : $\mu_X^n, \mu_Y^n, V_X^n, V_Y^n, V_{XY}^n, R_X^n, R_Y^n$.

Let $(\lambda_j^n)_{j=1,...,p}$ be the decreasing sequence of the p eigenvalues of R_X^n (and of the p largest eigenvalues of R_Y^n, the other ones, if $q > p$, being null), $(u_j^n, v_j^n)_{j=1,...,p}$ a sequence of associated unit eigenvectors of R_X^n and of R_Y^n and $(f_j^n, g_j^n)_{j=1,...,p}$ the canonical variables sequence, vectors of $\mathbb{R}^n$, obtained by :

$$\forall l \in \{1,...,n\} \quad (f_j^n)_l = \langle u_j^n, (V_X^n)^{-\frac{1}{2}}(X_l - \mu_X^n)\rangle_p \quad \text{and} \quad (g_j^n)_l = \langle v_j^n, (V_Y^n)^{-\frac{1}{2}}(Y_l - \mu_Y^n)\rangle_q.$$

If the case arises ($q > p$), we let $\lambda_{p+1}^n = 0$, we complete $(v_j^n)_{j=1,...,p}$ with R_Y^n eigenvectors in order to obtain an orthonormal basis $(v_j^n)_{j=1,...,q}$ of $\mathbb{R}^q$ and we define the canonical variables associated.

At last, for all j in $\{1, ..., p+1\}$ let $\rho_j^n = \sqrt{\lambda_j^n}$. Sample CCA of (X, Y) is then:

$$((\rho_j^n)_{j=1,...,p+1}, (f_j^n)_{j=1,...,p}, (g_j^n)_{j=1,...,q}). \tag{2}$$

1.3 Asymptotic study

Asymptotic study of CCA consists in establishing a.s. convergence of the canonical elements sequences of the sample CCA (2) to the corresponding canonical elements of the population CCA (1) and in establishing convergence in distribution of the standardized canonical elements sequences.

Difficulties are numerous : canonical variables are estimated ("predicted") by $\mathbb{R}^n$ vectors, the space dimension increasing with sample size. Then the use is to restrict asymptotic study to R_X^n and R_Y^n eigenvectors : $(u_j^n)_{j=1,...,p}$ and $(v_j^n)_{j=1,...,q}$ respectively, called *canonical vectors* and also to the $\mathbb{R}^p$ and $\mathbb{R}^q$ vectors defined by: $x_j^n = (V_X^n)^{-\frac{1}{2}}u_j^n$ and $y_j^n = (V_Y^n)^{-\frac{1}{2}}v_j^n$ respectively, called *canonical factors*; these vectors permit to obtain directly canonical variables by:

$$\forall l \in \{1,\ldots,n\} \quad (f_j^n)_l = \langle x_j^n, X_l - \mu_X^n\rangle_p \quad \text{et} \quad (g_j^n)_l = \langle y_j^n, Y_l - \mu_Y^n\rangle_q.$$

Multiple eigenvalues case is difficult to process because the eigenvectors associated with are not uniquely defined. Then the use is to restrict asymptotic study to the case where all eigenvalues are simple. Uniqueness is then verified by choosing systematically the unit vector (between the two ones) which has the first non null coordinate in respect with the canonical basis positive.

As for all multidimensional analyses, covariance matrices of sample random matrices V_X^n, V_{XY}^n, R_X^n, ... are "super-matrices" (that is, matrices of matrices). Tools such as the "vec" operator which transforms matrix into vector, have been introduced in order to handle theses super-matrices; the difficulty comes from the necessity of fixing the order of lines and columns elements.

In other respects, we know that the sequence ($\sqrt{n}(V_X^n - V_X)$) converges in distribution to a centered normal variable, the covariance super-matrix of which is known in some special cases, when X has a normal or elliptical distribution for example.

In the CCA frame work, we need to study convergence in distribution of the sequence ($\sqrt{n}(V_Z^n - V_Z)$) with $Z = (X, Y)$, then to study convergence in distribution of the sequence ($\sqrt{n}(R_Z^n - R_Z)$) with $R_Z = (R_X, R_Y)$ and $R_Z^n = (R_X^n, R_Y^n)$, before studying convergence of the R_Z^n eigenelements sequences and convergence of the sample canonical elements sequences. This asymptotic study is much more complex than the one of Principal Component Analysis (PCA) because principal values and principal vectors of the X PCA are eigenelements of the X covariance matrix.

It is only in 1999 that Anderson publishes a CCA asymptotic study when (X, Y) has a normal distribution and when all non zero eigenvalues are simple. Canonical factors components and canonical correlation coefficients of the population CCA are differentiable functions of V_Z. Results are then obtained from Taylor expansions. So, this classical approach may be qualified as matricial and analytic.

In order to simplify calculations, we propose to change variables from (X, Y) to (X', Y'), which is equivalent to changing the basis in $\mathbb{R}^p$ and $\mathbb{R}^q$. We then have: $V_{X'} = I_p$, and $R_X = R_{X'} = V_{X'Y'}V_{Y'X'}$, and similarly for $V_{Y'}$ and R_Y.

2 Operator and tensor approach

2.1 Introduction

Difficulties previously described are ensued only from the fact that matricial tool is not convenient. Working directly on linear operators in Euclidean spaces avoid indices problems and can be easily extended to an Hilbertian frame. Moreover, instead of studying eigenvectors associated with simple eigenvalues, it is possible to study eigenprojectors associated with multiple eigenvalues. Eaton (1983) also advices a "vector space approach" of the multidimensional statistics.

Dauxois and Pousse (1976) enlarge the PCA definition of a $\mathbb{R}^p$ random vector to a Hilbert random variable and even to a Hilbert random function, that is, a Hilbert random variable depending on a parameter in order to process temporal or spatial data. They redefine each factorial analysis (PCA, CCA, Correspondence Analysis, Discriminant Analysis, ...) in an operatorial and stochastic frame, that leads them to define, between others, nonlinear analyses.

The first asymptotic study in this frame has been realized by Romain (1979) for a Hilbert random function PCA (see also Dauxois, Pousse and Romain, 1982), study completed by Arconte (1980) who also started on CCA asymptotic study but all tools were not available to continue the study. Dauxois, Romain and Viguier (1994) propose to use some tensor products and establish a dictionary between matricial and operatorial formula. This work permits to compare common results obtained in both frames, but also to obtain more easily complex results; writings in respect with eigenvectors basis are established after concise formulations with operators.

These new tools permit to realize in Fine (2000) the CCA asymptotic study without restriction, that is, without assumption on the (X, Y) distribution, in the general case where eigenvalues may be multiple and without excluding canonical variables asymptotic study (CCA random elements). Therefore, our approach may be qualified as an operator and tensor approach. We give below the different steps of the CCA asymptotic analysis, some tools used and some examples of results.

2.2 Different steps of the CCA asymptotic study, tools, results

1) Population CCA

First, the matter is to define CCA of a pair of Euclidean random variables (population CCA). Again, we use classical approach notations substituting $(\mathbb{R}^p, \langle .,. \rangle_p)$ and $(\mathbb{R}^q, \langle .,. \rangle_q$ for p and q dimensional Euclidean spaces $(\mathcal{X}, \langle .,. \rangle_{\mathcal{X}})$ and $(\mathcal{Y}, \langle .,. \rangle_{\mathcal{Y}})$ respectively. Obviously, we work without reference to any basis (free coordinate).

Let $L^2(P)$ the Hilbert space of r.r.v. defined on $(\Omega, \mathcal{A}, P)$ and admitting order 2 moments, scalar product of which associates $\mathbb{E}(fg)$ to (f, g).

The operator Φ_X from $\mathcal{X}$ to $L^2(P)$ which associates $\langle x, X \rangle_{\mathcal{X}}$ to x plays an essential role in the operator approach of multidimensional statistics. In particular, X is a normal Euclidean random variable if, and only if, $\forall x \in \mathcal{X}, \langle x, X \rangle_{\mathcal{X}}$ is a normal r.r.v..

The expected value of X is the unique element of $\mathcal{X}$ (Riesz theorem), denoted by $\mathbb{E}(X)$, verifying: $\forall x \in \mathcal{X}, \langle x, \mathbb{E}(X) \rangle_{\mathcal{X}} = \mathbb{E}(\langle x, X \rangle_{\mathcal{X}})$.

For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we denote by $x \otimes y$ the operator from $\mathcal{X}$ to $\mathcal{Y}$ which associates $\langle x', x \rangle_{\mathcal{X}}\, y$ to x' ; it is an element of the Hilbert space $\sigma_2(\mathcal{X}, \mathcal{Y})$ of operators from $\mathcal{X}$ to $\mathcal{Y}$ with the scalar product : $\langle A, B \rangle_2 = tr(AB^*)$. Due to the Riesz theorem, we may then define covariance operators V_X of X, V_Y of Y, and crossed covariance operators V_{XY} and V_{YX} of X and Y: $V_X = \mathbb{E}((X - \mu_X) \otimes (X - \mu_X))$, ...

As in the CCA classical approach (§1.1), X and Y are assumed to be centered. The adjoint operator Φ_X^* of Φ_X is the operator from $L^2(P)$ to $\mathcal{X}$ which associates $\mathbb{E}(fX)$ to f and then we have : $\Phi_X^* \circ \Phi_X = V_X$, $\Phi_X^* \circ \Phi_Y = V_{XY}$, ...

It is convenient to represent operators relationships in the following commutative diagram, also called a duality scheme; here, each space is identified with its dual space. The H_X and H_Y spaces are image spaces of Φ_X and Φ_Y respectively and the orthogonal projectors of $L^2(P)$ on these subspaces are: $\Pi_X = \Phi_X \circ V_X^{-1} \circ \Phi_X^*$ and $\Pi_Y = \Phi_Y \circ V_Y^{-1} \circ \Phi_Y^*$.

$$\begin{tikzcd}
\mathcal{X} \arrow[d, shift right, "V_X^{-1}"'] & L^2(P) \arrow[l, "\Phi_X^*"'] \arrow[r, "\Phi_Y^*"] & \mathcal{Y} \arrow[d, shift left, "V_Y^{-1}"] \\
\mathcal{X} \arrow[u, shift right, "V_X"'] \arrow[r, "\Phi_X"'] & L^2(P) \arrow[u, "I"'] & \mathcal{Y} \arrow[l, "\Phi_Y"] \arrow[u, shift left, "V_Y"]
\end{tikzcd}$$

Operators R_X and R_Y, and also CCA of (X, Y), are defined as previously (symbols $\circ$ are deleted in order to reduce notation).

As in the classical approach, in order to facilitate calculations, we change the scalar product on $\mathcal{X}$ so that the covariance operator of X is the identity of $\mathcal{X}$, and similarly for $\mathcal{Y}$. We then have: $R_X = V_{XY}V_{YX}$ and $R_Y = V_{YX}V_{XY}$.

2) Sample model and sample CCA

We use a sample model (Dauxois, Fine and Pousse, 1979) establishing a link between the sample used in Data Analysis and the i.i.d. sample of Statistics. A sample $(X_l, Y_l)_{l\in\mathbb{N}^*}$ i.i.d. as (X, Y) is built from an element ω of $\Omega^{\mathbb{N}^*}$ setting, for all l of $\mathbb{N}^*$ (π_l denoting the l^{th} projection of $\Omega^{\mathbb{N}^*}$ onto Ω) : $X_l = X \circ \pi_l$ and $Y_l = Y \circ \pi_l$ that is $X_l(\omega) = X(\omega_l)$ and $Y_l(\omega) = Y(\omega_l)$.

We then provide $L^2(P)$ with the scalar product (random scalar product as it depends on ω):

$$\forall (f, g) \in L^2(P) \times L^2(P), \quad \mathbb{E}_n(fg) = \frac{1}{n}\sum_{l=1}^{n} f(\omega_l)g(\omega_l).$$

Then we have:

$$\mathbb{E}_n(X) = \mu_X^n, \ \Phi_X^n = \langle ., X - \mu_X^n\rangle_{\mathcal{X}}, \ V_X^n = \frac{1}{n}\sum_{l=1}^{n}(X_l - \mu_X^n) \otimes (X_l - \mu_X^n), \ldots$$

This sample model is the clue to distinguish randomness implied by the model ($L^2(P)$ elements) from randomness implied by sampling. It permits to obtain the canonical variables asymptotic distribution.

The duality scheme of sample CCA is the same as the population one after substituting $L^2(P)$ for $(L^2(P), \mathbb{E}_n)$ and indexing operators by n.

Sample operators R_X^n and R_Y^n and sample CCA are defined as previously.

3) Convergence of sample operators sequence

Limit theorems in Euclidean or Hilbert spaces permit to obtain a.s. convergence and convergence in distribution of the covariance operators sequence without assumption on the distribution of (X, Y) except the existence of order 4 moments. For CCA, we obtain (remind we let $Z = (X, Y)$):

$$W_Z^n := \sqrt{n}(V_Z^n - V_Z) \quad \xrightarrow{\mathcal{D}} \quad W_Z \sim N(0\,; \mathbb{K}_Z),$$

where $\mathbb{K}_Z$ is the covariance operator of $Z \otimes Z$.

In what concerns the sample operators $R_Z^n = (R_X^n, R_Y^n)$, elements of $\sigma_2(\mathcal{Z})$ (with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$), a.s. convergence derives from the fact that it is possible to write R_X^n and R_Y^n as continuous function of V_Z^n.

Let $U_Z^n = \sqrt{n}(R_Z^n - R_Z)$ $(:= (U_X^n, U_Y^n))$.

We write $U_X^n = \Psi_X^n(W_Z^n)$ where (Ψ_X^n) is a sequence of random operators from $\sigma_2(\mathcal{Z})$ to $\sigma_2(\mathcal{X})$ a.s. converging to Ψ_X. We then deduce the convergence in distribution of (U_X^n) to $U_X = \Psi_X(W_Z)$, centered normal variable, covariance operator of which being $\mathbb{L}_X = \Psi_X \circ \mathbb{K}_Z \circ \Psi_X^*$, and the same result for (U_Y^n) permuting X and Y roles. The proposition used here, is easy to prove from classical results in metric spaces (Billingsley, 1968). We obtain for example:

$$U_X = -\frac{1}{2}(W_X R_X + R_X W_X) + W_{XY} V_{YX} + V_{XY} W_{YX} - V_{XY} W_Y V_{YX} \sim N(0\,; \mathbb{L}_X)$$

4) Convergence of eigenelements and CCA elements sequences

Whatever may be the "factorial" method, which is an analysis or a model obtained from a spectral (or singular-value) decomposition, all results concerning eigenelements (eigenvalues, eigenprojectors, eigenvectors associated with simple eigenvalues, ...) are easily obtained thanks to perturbation theory of linear operators (Kato, 1980). In Fine (1987), this theory has been adapted to bounded perturbations that permits to use it, due to the iterated logarithm law, in the asymptotic study frame. So we obtain a.s. expansions of eigenelements of a symmetric positive operators sequence.

We may also consult Dossou-Gbete and Pousse (1991) for limit results but, for the convergence in distribution of some CCA elements, limit results are not sufficient when perturbation expansions permit to conclude.

For example, for the canonical factors associated to a simple eigenvalue λ_i, we have: $x_i = u_i$ because $V_X = I_X$ and $x_i^n = (V_X^n)^{-\frac{1}{2}} u_i^n$ so:

$$\sqrt{n}(x_i^n - x_i) = -(V_X^n)^{-\frac{1}{2}}((V_X^n)^{\frac{1}{2}} + I_X)^{-1}[\sqrt{n}(V_X^n - I_X)]u_i^n + [\sqrt{n}(u_i^n - u_i)].$$

We know that $(\sqrt{n}(V_X^n - I_X))$ converges in distribution to W_X and $(\sqrt{n}(u_i^n - u_i))$ to $S_{X_i} U_X x_i$ (with $S_{X_i} = (R_X - \lambda_i I_X)^-$) but, thanks to perturbation expansions, it is possible to establish:

$$\sqrt{n}(x_i^n - x_i) \quad \xrightarrow{\mathcal{D}} \quad \frac{1}{2} W_X x_i + S_{X_i} U_X x_i \sim N(0\,; \mathbb{L}_{Xi})$$

5) Asymptotic covariance operators in the elliptical case

We have already seen that the asymptotic covariance operator of $(\sqrt{n}(R_X^n - R_X))$ is $\mathbb{L}_X = \Psi_X \circ \mathbb{K}_Z \circ \Psi_X^*$ where $\mathbb{K}_Z$ is the asymptotic covariance operator of $(\sqrt{n}(V_Z^n - V_Z))$ and where the operator Ψ_X from $\sigma_2(\mathcal{Z})$ to $\sigma_2(\mathcal{X})$ can be written explicitly. All the distribution limits of eigenelements or CCA elements sequences are centered normal variables (or function of centered normal variables), covariance operator of which being written as function of $\mathbb{K}_Z$ in the same way.

Now, we may write explicitly these asymptotic covariance operators in the case where Z has an elliptical distribution with mean μ_Z, covariance operator V_Z and kurtosis κ (real parameter, which, when it is null, leads to a $N(\mu_Z, V_Z)$ distribution). We then know that $\mathbb{K}_Z$ is the operator from $\sigma_2(\mathcal{Z})$ to itself which associates to T:

$$\mathbb{K}_Z(T) = (1 + \kappa)V_Z(T + T^*)V_Z + \kappa\langle V_Z, T\rangle_2 V_Z.$$

At this step, we need more algebraic tools. The tensor product in spaces of type σ_2 is denoted by $\tilde{\otimes}$. For example:

$$\forall (A, B) \in \sigma_2(\mathcal{Z}) \times \sigma_2(\mathcal{Z}), \quad \forall T \in \sigma_2(\mathcal{Z}), \quad A \tilde{\otimes} B(T) = \langle T, A \rangle_2 B.$$

We define also product $\overset{\ell}{\otimes}$ in spaces of type σ_2. For example:

$$\forall (A, B) \in \sigma_2(\mathcal{Z}) \times \sigma_2(\mathcal{Z}), \quad \forall T \in \sigma_2(\mathcal{Z}), \quad A \overset{\ell}{\otimes} B(T) = BTA^*.$$

We define the commutation operator C which associates to an operator T its adjoint T^*. At last, we substitute the space product $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ for the Hilbertian sum $\mathcal{Z} = \mathcal{X} \oplus \mathcal{Y}$; this permits to plunge all the operators into $\sigma_2(\mathcal{Z})$ in order to simplify notation. Projector P_X from $\sigma_2(\mathcal{Z})$ onto $\sigma_2(\mathcal{X})$ becomes in this frame a symmetric operator of $\sigma_2(\mathcal{Z})$.

The operator $\mathbb{K}_Z$ from $\sigma_2(\mathcal{Z})$ to itself may be written as:

$$\mathbb{K}_Z = (1 + \kappa) V_Z \overset{\ell}{\otimes} V_Z (I + C) + \kappa V_Z \tilde{\otimes} V_Z.$$

Let $(x_i)_{i=1,\dots,p}$ be an orthonormal basis formed by canonical factors of $\mathcal{X}$, then $(x_i \otimes x_j)_{i,j=1,\dots,p}$ is an orthonormal basis of $\sigma_2(\mathcal{X})$ and $((x_i \otimes x_j) \tilde{\otimes} ((x_k \otimes x_l))_{i,j,k,l=1,\dots,p}$ is an orthonormal basis of $\sigma_2(\sigma_2(\mathcal{X}))$.

After calculations obtained in a concise way, it is easy to decompose operators in respect with this type of basis. For example, we obtain for the asymptotic covariance operator of ($\sqrt{n}(R_X^n - R_X)$):

$$\mathbb{L}_X = (1 + \kappa)(I + C)[-\frac{3}{4} R_X^2 \overset{\ell}{\otimes} I_X + R_X^2 \overset{\ell}{\otimes} R_X + R_X \overset{\ell}{\otimes} I_X - \frac{5}{4} R_X \overset{\ell}{\otimes} R_X](I + C)$$

and, in respect with the basis of canonical factors (remember that $(\lambda_j)_{j=1,\dots,p}$ is the decreasing sequence of eigenvalues of R_X):

$$\mathbb{L}_X = \frac{1}{2}(1 + \kappa) \sum_{j=1}^{p} \sum_{k=1}^{p} \left(-\frac{3}{4}\lambda_j^2 - \frac{3}{4}\lambda_k^2 + \lambda_j^2 \lambda_k + \lambda_j \lambda_k^2 + \lambda_j + \lambda_k - \frac{5}{2}\lambda_j \lambda_k \right)$$

$$(x_j \otimes x_k + x_k \otimes x_j) \tilde{\otimes} (x_j \otimes x_k + x_k \otimes x_j)$$

When (X, Y) has a normal distribution and when all eigenvalues are simple, it is possible to rediscover Anderson's results but we diverge on two of them.

6) Convergence of CCA random elements sequences

As previously announced (§ 2.1.2) the sample model permits to obtain a.s. convergence and convergence in distribution of canonical variables sequences. We have for example, for the canonical variable associated with a simple eigenvalue λ_i:

$$\sqrt{n}(f_i^n - f_i)) \quad \xrightarrow{\mathcal{D}} \quad \langle \frac{1}{2} W_X x_i + S_{Xi} U_X x_i, X \rangle_X \sim N(0\,; \mathbb{M}_{Xi})$$

with, in the particular case where (X, Y) has an elliptical distribution:

$$\mathbb{M}_{Xi} = \frac{1}{4}(2 + 3\kappa)f_i \otimes f_i + (1 + \kappa)\sum_{j \neq i}(1 - \lambda_i)(\lambda_i + \lambda_j - 2\lambda_i\lambda_j)(\lambda_i - \lambda_j)^{-2} f_j \otimes f_j$$

7) Inferential applications and conclusion

These results on CCA asymptotic study permit to tackle easily inferential applications (confidence interval estimation, statistical tests, ...) which imply CCA elements, particularly the proximity measures built on canonical correlation coefficients. See Anderson (1999) and Dauxois and Nkiet (2002).

Further aspects and results may be consulted in Fine (2000). This methodological presentation shows that the operator approach performs quite well in solving asymptotic problems in multivariate statistics.

3 References

Anderson, T. W. (1999). Asymptotic Theory for Canonical Correlation Analysis. *Journal of Multivariate Analysis*, 70, 1-29.

Arconte, A. (1980). *Étude asymptotique de l'analyse en composantes principales et de l'analyse canonique.* Thèse de 3ème cycle, Université de Pau et des Pays de l'Adour.

Billingsley, P. (1968). *Convergence of probability measures.* Wiley, New York.

Dauxois, J., Fine, J. and Pousse A. (1979). Échantillonnage en segmentation, étude de la convergence. *Statistique et Analyse des Données,* 3, 45-53.

Dauxois, J. and Nkiet, G. M. (2002). Measures of Association for Hilbertian subspaces and some applications. *Journal of Multivariate Analysis*, 82, 263-298.

Dauxois, J. and Pousse, A. (1976). *Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthétique.* Thèse de Doctorat d'État, Université Paul Sabatier, Toulouse.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function; some applications to statistical inference. *Journal of Multivariate Analysis*, 12, 136-154.

Dauxois, J., Romain, Y. and Viguier, S. (1994). Tensor products and statistics. *Linear Algebra and its Applications*, 210, 59-88.

Dossou-Gbete, S. and Pousse, A. (1991). Asymptotic study of eigenelements of a sequence of random self adjoint operators. *Statistics,* 22, 479-491.

Eaton, M. L. (1983). *Multivariate statistics. A vector space approach.* Wiley, New York.

Fine, J. (1987). On the validity of the perturbation method in asymptotic theory. *Statistics,* 18, 401-414.

— (2000). Étude Asymptotique de l'Analyse Canonique. *Pub. Inst. Stat. Univ. Paris,* 44, 2-3, 21-72.

Kato, T. (1980). *Perturbation theory for linear operators.* Springer-Verlag, New York.

Romain, Y. (1979). *Étude Asymptotique des approximations par échantillonnage de l'analyse en composantes principales d'une fonction aléatoire. Quelques applications.* Thèse 3ème cycle. Université Paul Sabatier. Toulouse.

Resum

L'estudi asimptòtic de l'Anàlisi de la Correlació Canònica ens permet presentar els diferents passos de les propietats asimptòtiques i mostrar l'interès del plantejament amb operadors i tensors dels estadístics multivariants en comptes del plantejament clàssic, matricial i analític. Emprant aquesta aproximació clàssica, Anderson (1999) suposa que els vectors aleatoris segueixen la distribució normal i que els coeficients de correlacions canòniques no nuls són diferents. Fem servir un nou plantejament a lliure distribució (Fine, 2000) que també és lliure de les coordenades i que no té restriccions sobre l'ordre de multiplicitat de les coeficients de correlacions canòniques Tanmateix, quan els vectors aleatoris segueixen la distribució normal i quan les coeficients de correlacions canòniques no nul·les són diferents, podem recuperar els resultats d'Anderson, però no coincidim en dues situacions. En aquesta presentació metodològica, insistim en l'estructura analítica (Dauxois and Pousse, 1976), els models d'obtenció de mostres (Dauxois, Fine and Pousse, 1979) i diferents eines matemàtiques (Fine, 1987, Dauxois, Romain and Viguier, 1994), que permeten resoldre problemes que apareixen en aquest tipus d'estudi, i fins i tot obtenir el comportament asimptòtic dels aspectes aleatoris d'altres elements (components principals, variables canòniques, ...).

MSC: 62E20, 62H20, 62H25, 47N30

Paraules clau: Anàlisi multivariant, anàlisi de correlació canònica, estudi asimptòtic, operadors, lliure de coordenades, lliure distribució

SORT 27 (2) July-December 2003, 175-190

A posteriori disclosure risk measure for tabular data based on conditional entropy**

Anna Oganian and Josep Domingo-Ferrer[a*]

[a] *Universitat Rovira i Virgili, Spain*

Abstract

Statistical database protection, also known as Statistical Disclosure Control (SDC), is a part of information security which tries to prevent published statistical information (tables, individual records) from disclosing the contribution of specific respondents. This paper deals with the assessment of the disclosure risk associated to the release of tabular data. So-called sensitivity rules are currently being used to measure the disclosure risk for tables. These rules operate on an *a priori* basis: the data are examined and the rules are used to decide whether the data can be released as they stand or should rather be protected. In this paper, we propose to complement *a priori* risk assessment with *a posteriori* risk assessment in order to achieve a higher level of security, that is, we propose to take the protected information into account when measuring the disclosure risk.

The proposed *a posteriori* disclosure risk measure is compatible with a broad class of disclosure protection methods and can be extended for computing disclosure risk for a set of linked tables. In the case of linked table protection via cell suppression, the proposed measure allows detection of secondary suppression patterns which offer more protection than others.

MSC: 62P99

Keywords: statistical disclosure control, statistical databases, tabular data, security

1 Introduction

Statistical database protection is a part of information security called inference control in the classical book by Denning (1982). The most typical output offered by national statistical agencies is tabular data. Tables are central in official statistics: many survey and census data are categorical in nature, so that their representation as cross-classifications or tables is a natural reporting strategy. Tabular data being thus aggregate

** Work partly funded by the European Union under project "CASC" IST-2000-25069. The first author is supported by a grant from Generalitat de Catalunya.

* *Address for correspondence:* {aoganian, jdomingo}@etse.urv.es. Universitat Rovira i Virgili, Dept. of Computer Eng. and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain.

Received: April 2003

Accepted: September 2003

data, one is tempted to think they are not supposed to contain information that can reveal the contribution of particular respondents. However, as noted in Giessing (2001), in many cases table cells do contain information on a single or very few respondents, which implies a disclosure risk for the data of those respondents. In these cases, disclosure control methods must be applied to the tables prior to their release.

A number of disclosure control methods to protect tabular data have been proposed (see Willenborg and de Waal (2001), Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001) for a survey). Next we list the main principles underlying those methods:

Cell suppression If a table cell is deemed sensitive, then it is suppressed from the released table (primary suppression). If marginal totals or other linked tables are also to be published, then it may be necessary to remove additional table cells (secondary suppressions) to prevent primary suppressions from being computable. Secondary suppressions should be chosen in a way such that the utility of the resulting table is maximized.

Rounding A positive integer b (rounding base) is selected and all table cells are rounded to an integer multiple of b. Controlled rounding is a variant of rounding in which table additivity is preserved (*i.e.* rounded rows and columns still sum to their rounded marginals).

Table redesign Categories used to tabulate data are recoded into different (often more general) categories, so that the resulting tabulation does not contain sensitive cells any more. A simple redesign could be to combine two rows containing sensitive cells to obtain a new row without sensitive cells.

Sampling A table is released which is based on a sample of the units on which the original table was built.

Swapping and simulation In data swapping, units are swapped so that the table resulting from the swapped data set still preserves all k-dimensional margins of the original table. A more elaborate version of swapping was proposed in Fienberg, Makov and Steele (1998), whereby the original table is replaced by a random draw from the exact distribution under the log-linear model whose minimal sufficient statistics correspond to the margins of the original table. Further extensions of this idea would lead to drawing a synthetic table from the full distribution of all possible tables with the same margins of the original table.

As noted by Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001), any attempt to compare methods for tabular data protection should focus on two basic attributes:

1. *Disclosure risk:* a measure of the risk to respondent confidentiality that the data releaser (typically a statistical agency) would experience as a consequence of releasing the table.
2. *Data utility:* a measure of the value of the released table to a legitimate data user.

In this paper, we concentrate on the assessment of disclosure risk. Up to now, disclosure risk assessment for tables was usually performed *a priori*, that is, before applying any protection methods to the table. The standard approach is to use a *sensitivity rule* to decide whether a particular table cell can safely be released.

However, *a priori* measures do not actually measure the disclosure risk incurred once a particular table is released. In this paper, we propose to complement *a priori* risk assessment with a *posteriori* risk assessment, which takes protected information into account. The proposed measure applies to a broad class of disclosure protection methods and is computable in practice.

Section 2 describes existing disclosure risk measures, which are *a priori* by their nature. In Section 3, an *a posteriori* measure based on the reciprocal of conditional entropy is proposed as a complement to *a priori* measures. Section 4 describes an application of the proposed *a posteriori* measure to different table protection methods, both for simple tables and for linked tables. In the case of cell suppression methods, the proposed measure turns out to be useful to detect suppression patterns which offer more protection than others. Section 5 is a conclusion.

2 Background on *a priori* disclosure risk measures

A priori disclosure risk measures used by statistical agencies for tabular data protection are also called sensitivity rules. For magnitude tables (normally related to economic data), there are two widely accepted sensitivity rules:

$n - k$-dominance In this rule, n and k are two parameters with values to be specified. A cell is called sensitive if the sum of the contributions of n or fewer respondents represents more than a fraction k of the total cell value.

pq-rule The prior-posterior rule is another rule gaining increasing acceptance. It also has two parameters p and q. It is assumed that, prior to table publication, each respondent can estimate the contribution of each other respondent to within less than q percent. A cell is considered sensitive if, posterior to the publication of the table, someone can estimate the contribution of an individual respondent to within less than p percent. A special case is the $p\%$-rule: in this case, no knowledge prior to table publication is assumed, *i.e.* the pq-rule is used with $q = 100$.

For tables of counts or frequencies (normally related to demographic data), a so-called **threshold rule** is used. A cell is defined to be sensitive if the number of respondents is less than a threshold k.

These sensitivity rules have received critiques for failing to adequately reflect the risk of disclosure, but these were mostly limited to numerical counterexamples for particular choices of the parameters of these rules. Recently, it was shown in Domingo-Ferrer and Torra (2002) through general counterexamples that releasing a cell declared non-sensitive by these rules can imply higher disclosure risk than releasing a cell declared sensitive. It was proposed to use Shannon's entropy of relative contributions to a table cell as a better alternative to (n, k)-dominance, pq-rule and $p\%$-rule. Formally speaking,

$$H(X) = -\sum_{i=1}^{N} (x_i/x) \log_2 (x_i/x) \tag{1}$$

where $x = x_1 + x_2 + \cdots + x_N$ is the value of a table cell and xi are contributions to that cell.

A cell is considered sensitive by the above rule if $H(X)/\log_2 N < t$, where $t \in [0, 1]$ is a parameter; otherwise, the cell is declared non-sensitive.

3 An *a posteriori* disclosure risk measure based on conditional entropy

We have mentioned above that using only *a priori* measures may be insufficient for table protection. Now we want to illustrate this on the following examples.

Example 1 *Suppose that the person or entity who wants to guess secret information about how much a particular respondent contributed to some cell of the table is someone who also contributed to that cell. So he obviously knows his own contribution to that cell. He also may know some additional information, for example, how many respondents have contributed to that cell, who they are, etc. This internal intruder is in a better position than an outsider to estimate the contribution of his interest. This kind of information is not taken into account by a priori measures. According to these, the disclosure risk is the same for all types of intruders and that is not true.*

The information held by an intruder does not only depend on her being internal or external; it clearly depends also on what information has previously been published and on how that information has been protected.

Example 2 *Assume we have an n-dimensional table whose cells are deemed sensitive, and therefore cannot be released. Only some 2-dimensional (or $(n - i)$-dimensional) tables are released, which have been obtained as projections of the n-dimensional table. Due to their origin, the released tables are linked tables, so the uncertainty about a cell value in the n-dimensional table is conditional to the particular tables released so far.*

Therefore, we propose to complement *a priori* risk assessment provided by sensitivity rules with *a posteriori* risk assessment. The latter is performed *after* data have been protected and takes protected data into account to compute bounds on cells labeled as sensitive by a sensitivity rule.

Our proposal for *a posteriori* measure is to use the reciprocal of Shannon's conditional entropy [Shannon (1948)] to express the disclosure risk in a natural and unified manner.

Entropy-based measures were already discussed in Willenborg and de Waal (2001) for computing information loss at the table level, but not for computing disclosure risk. However, the authors of Willenborg and de Waal (2001) do not believe entropy is a practical information loss measure. We support their opinion with the following example.

Example 3 *Assume we use rounding with integer base b to protect a table. The entropy-based information loss measure defined in Willenborg and de Waal (2001) is the reciprocal of the number of original tables whose rounded version matches the published rounded table (i.e. the number of original tables "compatible" with the published one). The number of compatible tables depends on the rounding base b, but is independent on how close the published rounded values are to the original values. Thus, the entropy-based information loss measure is the same when the original table exactly corresponds to the rounded table (which happens when all cell values in the original table are multiples of b) and when all differences between corresponding cell values in the original and rounded tables are close to b/2. This does not seem to adequately reflect the utility of the published data.*

In Duncan, Fienberg, Krishnan, Padman, and Roehrig (2001), the reciprocal of Shannon's entropy (not conditional entropy) was suggested as measure of disclosure risk at the cell level. What was not clear there is how to compute the probabilities, that is, what distribution should be chosen. In fact, as we noted above, the particular distribution for an intruder depends on the knowledge of that intruder.

The above discussion suggests that the most natural *a posteriori* measure for disclosure risk is the reciprocal of conditional entropy

$$DR(X) = 1/H(X|Y = y) = 1/\left(-\sum_{x} p(x/y) \log_2 p(x/y)\right) \qquad (2)$$

where X is a variable representing an original cell and Y is a variable representing the intruder's knowledge (which is supposed to be equal to some specific value y). The intuitive notion behind Expression (2) is that, the more uncertainty about the value of the original cell X (which depends on the constraints $Y = y$), the less disclosure risk (and conversely).

There are two practical problems in computing Expression (2):

1. Finding the set $S_y(X)$ of possible values of X given the constraints y.
2. Estimating the probabilities $p(x|y)$, *i.e.* the probability of the cell X being x given that Y is y.

As noted by Willenborg and de Waal (2001) when discussing entropy-based information loss measures, taking the uniform probability distribution over the set $S_y(X)$ can make sense for some disclosure control methods. Using the uniform distribution, Expression (2) is simplified to

$$DR_{unif}(X) = 1 / \log_2 m\left(S_y(X)\right) \tag{3}$$

where $m\left(S_y(X)\right)$ is the number of possible values of the cell in $S_y(X)$.

Table 1: *A table with suppressed cells.*

Economic activity	*Size class*					*Total*
	4	5	6	7	8	
2,3	80	253	54	0	0	387
4	641	3694	2062	746	0	7143
5	592	x_1	329	x_2	1440	3898
6	57	x_3	946	x_4	2027	4281
7	78	0	890	1719	1743	4430
Total	1148	4353	4281	4847	5210	20139

Note 1 (On $m(S_y(X))$) *We assume in what follows that table cells take values in a discrete domain: either integer values or real values with a fixed number of decimal positions. This is the usual case in published statistical tables: count tables consist of integer values and magnitude tables consist of either integer values or real values with limited precision. Thus the set $S_y(X)$ of possible values is enumerable and it makes sense to speak of $m\left(S_y(X)\right)$ as the number of cell values in $S_y(X)$.*

4 Application to several table protection scenarios

We show in this Section how to compute Expression (3) for several disclosure control methods for tables; the case of linked tables will also be discussed.

4.1 Cell suppression

The disclosure risk computation for cell suppression is illustrated by extending an example provided in Willenborg and de Waal (2001). Let Table 1 be a table from which four cells x_1, x_2, x_3 and x_4 have been suppressed. Assume that the suppressed values are integer.

According to the definition given in Section 3, the disclosure risk for each suppressed cell is the reciprocal of one of the following conditional entropies:

$$\begin{aligned} &H(x_1|x_1 + x_2 = 1537,\ x_1 + x_3 = 406,\ x_i \geqslant 0)\\ &H(x_2|x_1 + x_2 = 1537,\ x_2 + x_4 = 2382,\ x_i \geqslant 0)\\ &H(x_3|x_1 + x_3 = 406,\ x_3 + x_4 = 1251,\ x_i \geqslant 0)\\ &H(x_4|x_2 + x_4 = 2382,\ x_3 + x_4 = 1251,\ x_i \geqslant 0) \end{aligned} \qquad (4)$$

***Table 2**: A table with two rows combined.*

Economic activity	*Size class*					*Total*
	4	5	6	7	8	
2,3	80	253	54	0	0	387
4	641	3694	2062	746	0	7143
5,6	649	406	1275	2382	3467	8179
7	78	0	890	1719	1743	4430
Total	1148	4353	4281	4847	5210	20139

Expressions (4) contain constraints yi for each suppressed cell xi which allow $m(S_{yi}(x_i))$ to be computed by solving two linear programming (LP) problems (one maximization and one minimization) and subtracting the solutions. In the case of Table 1, minimizations and maximizations bound every cell as follows: $0 \leqslant x_1 \leqslant 406$, $1131 \leqslant x_2 \leqslant 1537$, $0 \leqslant x_3 \leqslant 406$ and $845 \leqslant x_4 \leqslant 1251$. By subtracting the bounds we obtain $m(S_{yi}(x_i)) = 407$ for $i = 1, 2, 3, 4$. Using Expression (3), we can compute $DR_{unif}(x_i) = 1/\log_2 407 = 0.115$ for every cell.

4.2 Rounding

When the table is protected by rounding, the cell entropy conditional to the rounded table depends on the rounding base b. In a rounded table without marginals, if the value of a cell x_i' is $n_i b$ (*i.e.* n_i times the rounding base), then we know that the original cell x_i must lie in the interval $I_i = [(n_i - 1/2)b, (n_i + 1/2)b)$. Thus, $DR_{unif}(x_i) = 1/\log_2 m(I_i)$, where $m(I_i)$ is the number of possible cell values in I_i (keep in mind that cell values are either integer or with a fixed number of decimal positions).

4.3 Table redesign

This case is very similar to cell suppression. Imagine that the sensitive cells in Table 1 are protected by combining rows with *Economic_activity* = 5 or 6. This yields Table 2.

Let us label the six cells in the original row with *Economic_activity* = 5 as x_1 through x_6 and the six cells in *Economic_activity* = 6 as x_7 through x_{12} (x_6 is the marginal of the first row and x_{12} is the marginal of the second row).

Then the following equalities hold:

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 + x_5 - x_6 &= 0 \\ x_7 + x_8 + x_9 + x_{10} + x_{11} - x_{12} &= 0 \\ x_1 + x_7 &= 649 \\ x_2 + x_8 &= 406 \\ x_3 + x_9 &= 1275 \\ x_4 + x_{10} &= 2382 \\ x_5 + x_{11} &= 3467 \\ x_6 + x_{12} &= 8179 \\ x_i &\geqslant 0 \quad \text{for} \quad i = 1, \ldots, 12 \end{aligned} \tag{5}$$

From the above, $m(S_{yi}(x_i))$ and $DR_{unif}(x_i)$ are computed in a way analogous to the case of cell suppression.

4.4 Linked tables

We will show the application of conditional entropy as *a posteriori* disclosure risk measure for linked tables with an example.

Let us consider the three-dimensional table ASR formed by cells $z_{a_i s_j r_k}$, where each cell denotes the total turnover of businesses with activity a_i and size s_j in region r_k. Assume that table ASR is not released because every cell in it is considered sensitive. Instead of ASR, some of the following tables obtained by bidimensional projection are released: $AS = \{z_{a_i s_j}\}$, which breaks down turnover by activity and business size, $AR = \{z_{a_i r_k}\}$, which breaks down turnover by activity and region, and $SR = \{z_{s_j r_k}\}$, which breaks down turnover by size and region. Assume three scenarios: 1) only AS is released; 2) AS and AR are released; 3) AS, AR and SR are released. The disclosure risk of cell $z_{a_i s_j r_k}$ in each scenario can be expressed as:

$$DR_{unif}\left(z_{a_i s_j r_k} | AS\right) = 1/H\left(z_{a_i s_j r_k} | z_{a_i s_j} = \sum_k z_{a_i s_j r_k}\right) \tag{6}$$

$$DR_{unif}\left(z_{a_i s_j r_k} | AS, AR\right) \\ = 1/H\left(z_{a_i s_j r_k} | z_{a_i s_j} = \sum_k z_{a_i s_j r_k}, z_{a_i r_k} = \sum_j z_{a_i s_j r_k}\right) \tag{7}$$

$$DR_{unif}\left(z_{a_is_jr_k}|AS,AR,SR\right)$$
$$= 1/H\left(z_{a_is_jr_k}|z_{a_is_j} = \sum_k z_{a_is_jr_k}, z_{a_ir_k} = \sum_j z_{a_is_jr_k}, z_{s_jr_k} = \sum_i z_{a_is_jr_k}\right) \tag{8}$$

The released tables impose constraints on the possible cell values of the table ASR. Such constraints actually determine the simplexes $S_{AS}(z_{a_is_jr_k})$, $S_{AS,AR}(z_{a_is_jr_k})$ or $S_{AS,AR,SR}(z_{a_is_jr_k})$ where $z_{a_is_jr_k}$ should lie. By solving one LP maximization and one LP minimization for each $z_{a_is_jr_k}$, an interval where the cell lies can be determined. Then, the cell disclosure risk is computed using Expression (3). If a cell is too closely bounded, then its disclosure risk is too high and disclosure control methods must be used.

When the disclosure control method chosen is cell suppression, it is important to notice that linked tables have the property that there are sets of linearly dependent constraints, so that one of the constraints in each such set may be suppressed without decreasing the rank of the whole constraint system. This will influence the quality of the protection offered by different suppression patterns: the best pattern is the one decreasing most the system rank, which results in more degrees of freedom, and thus more cell entropy and lower disclosure risk. We show this with in the following section.

Table 3: Constraint matrix imposed by Table AS. Here $i \in \{1,\ldots,3\}$, $j \in \{1,\ldots,4\}$, $k \in \{1,\ldots,3\}$.

1	1	1																		$z_{a_1s_1}$
			1	1	1															$z_{a_1s_2}$
						1	1	1												$z_{a_1s_3}$
									1	1	1									$z_{a_1s_4}$
																				$\vdots$
																	1	1	1	$z_{a_3s_4}$

4.5 Minimizing disclosure risk in linked table release

For the sake of concreteness, we will resume the example of three linked tables used in the previous section. We want to estimate the disclosure risk of cells in ASR depending on the released tables. Assume that $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,m\}$ and $k \in \{1,\ldots,t\}$.

4.5.1 One table released out of three linked tables

If only the table AS is released, the constraint matrix is shown in Table 3, where there is a column for each cell $z_{a_is_jr_k}$. The matrix rank is nm, as all equations are linearly independent. Every choice for secondary suppressions causes the same rank decrease and consequently has an equivalent impact on the disclosure risk. Therefore, there is no room for optimization (unless there are specific additional constraints specified by the data protector).

4.5.2 Two tables released out of three linked tables

If tables AS and AR are released, the constraint matrix is shown in Table 4, where there is a column for each cell $z_{a_i s_j r_k}$. This matrix consists of n submatrices of size $(t+m) \times mt$ with rank $(t+m-1)$, that is, one constraint in each submatrix is a linear combination of the remaining constraints in the submatrix. That is, using Gaussian elimination we have:

$$\sum_{k=1}^{t} z_{a_i r_k} - \sum_{j=1}^{m} z_{a_i s_j} = 0 \quad \text{for} \quad 1 \leqslant i \leqslant n \tag{9}$$

Therefore, n constraints can be suppressed and the matrix rank will not change, nor will change $H(z_{a_i s_j r_k}|AS,AR)$ nor the disclosure risk. Note that only one row per submatrix can be suppressed for disclosure risk to stay unchanged, which, in terms of tables AS and AR, means only one cell per two corresponding rows in AS and AR (*e.g.* for the submatrix related to Expression (9), the two rows are those in AS and AR with subscript a_i).

From the above discussion, we can state the following proposition:

Proposition 1 *When two out three linked tables are released, the entropy increase and the disclosure risk decrease are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same columns and rows for both released tables.*

Proof. Assume that tables AS and AR are released. Now assume that $z_{a_i s_j r_k}$ in supertable ASR has a high disclosure risk which makes it necessary to increase its entropy. So, if a cell suppression is used, a natural option is to suppress the cells in the released tables which refer to $z_{a_i s_j r_k}$. These suppressions will be called primary suppressions and will be $z_{a_i s_j}$ and $z_{a_i r_k}$ in the tables AS and AR, respectively. The suppression of these two cells will decrease the rank by 1 (see the discussion above).

Secondary suppressions will be the following:

- Two cells, say $z_{a_i s_l}$ and $z_{a_f s_j}$ in the table AS in order to prevent $z_{a_i s_j}$ from being computable (in what follows, we will say —"to protect $z_{a_i s_j}$"), which decrease the rank by 1,
- The cell $z_{a_f s_l}$ in AS to protect $z_{a_f s_j}$, which decreases the rank by 1.
- Two cells in the table AR: one in the row a_i and other in the column r_k. If $z_{a_i r_v}$ is the cell in row a_i, its suppression decreases the rank by 1. When we choose the candidate for the suppression in column r_k, we should take into account that, if we choose the cell in a row other than a_f, this will not decrease the rank (because the rows where we have already suppressed cells are a_i and a_f). So, in order to

increase the entropy and to decrease disclosure risk, we have to choose row a_f, that is we have to suppress the cell $z_{a_f r_k}$.

- Finally, in order to protect $z_{a_f r_k}$, we have to suppress $z_{a_f r_v}$, which will decrease the rank by 1.

So, from the above argument, we have that choosing the same rows for secondary suppressions in both tables will decrease the rank by 1 more than if we chose different rows. Therefore, this is the strategy to maximize rank decrease, which is equivalent to to maximizing entropy increase and disclosure risk decrease. QED

Table 4: *Constraint matrix imposed by tables AS and AR. Here $i \in \{1,\dots,3\}, j \in \{1,\dots,4\}, k \in \{1,\dots,3\}$.*

1	1	1																		$z_{a_1 s_1}$
1			1			1			1											$z_{a_1 r_1}$
	1			1			1			1										$z_{a_1 r_2}$
		1			1			1			1									$z_{a_1 r_3}$
			1	1	1															$z_{a_1 s_2}$
						1	1	1												$z_{a_1 s_3}$
									1	1	1									$z_{a_1 s_4}$
													⋮						⋮	
																	1	1	1	$z_{a_3 s_4}$

4.5.3 All three linked tables released

Now let us consider the case where three tables AS, AR, SR are released. Again, using Gaussian elimination, we obtain several sets of constraints which are linearly dependent (exactly one constraint in each set is a linear combination of the rest). The linear combinations are as follows:

$$\sum_{k=1}^{t} z_{a_i r_k} - \sum_{j=1}^{m} z_{a_i s_j} = 0 \quad \text{for} \quad 1 \leqslant i \leqslant n \tag{10}$$

$$\sum_{k=1}^{t} z_{s_j r_k} - \sum_{j=1}^{n} z_{a_i s_j} = 0 \quad \text{for} \quad 1 \leqslant j \leqslant s-1 \tag{11}$$

$$\sum_{j=1}^{m} z_{s_j r_k} - \sum_{i=1}^{n} z_{a_i r_k} = 0 \quad \text{for} \quad 1 \leqslant k \leqslant t-1 \tag{12}$$

$$z_{s_m r_t} - \sum_{j=1}^{m-1} \sum_{k=1}^{t-1} z_{s_j r_k} + \sum_{i=1}^{n} \sum_{j=1}^{m-1} z_{a_i s_j} - \sum_{i=1}^{n} z_{a_i r_t} = 0 \tag{13}$$

Furthermore, up to $n+m+t-1$ cells can be suppressed without changing the entropy nor the disclosure risk. But this is heavily dependent on which cells are suppressed. A suppression pattern of maximal size which does not change the rank of the system may be the following: $z_{a_1 s_m}$, $z_{a_2 s_m}, \ldots, z_{a_n s_m}$, $z_{s_1 r_t}$, $z_{s_2 r_t}, \ldots, z_{s_{m-1} r_t}$, $z_{s_m r_1}$, $z_{s_m r_2}, \ldots, z_{s_m r_t}$. Now, assume that the cell $z_{a_1 s_1 r_1}$ has za1s1 = zs a a high disclosure risk which makes it necessary to increase its entropy. If cell suppression is used, a natural option is to suppress every cell in the three tables which refers to $z_{a_1 s_1 r_1}$. These suppressions will be called primary suppressions and will be $z_{a_1 s_1}$ from Table AS, $z_{a_1 r_1}$ from Table AR and $z_{s_1 r_1}$ from Table SR. Note that with these suppressions the rank of the system will decrease by 1. If $z_{a_1 s_1}$ is suppressed, the rank does not decrease because, by Expression (11), the suppressed cell is a linear combination

$$z_{a_1 s_1} = \sum_{k=1}^{t} z_{s_1 r_k} - \sum_{i \neq 1} z_{a_i s_1} \tag{14}$$

If $z_{a1,1}$ is suppressed next, the rank does not change either. By Expression (12), we can express the suppressed cell as a linear combination of cells which have not yet been suppressed:

$$z_{a_1 r_1} = \sum_{j=1}^{m} z_{s_j r_1} - \sum_{i \neq 1} z_{a_i r_1} \tag{15}$$

If $z_{s_1 r_1}$ is our third suppression, then the rank will decrease by 1, because that cell appears in Equations (14) and (15) and it is easy to see that there is no way to use the above equations to express that cell as a combination of the cells which have not yet been suppressed.

If the table is released with marginals, then a set of secondary suppressions is required to prevent primary suppressions from being computable. At this moment, it is important to choose the necessary strategy for secondary suppressions because the rank of the system and consequently the cell entropy will vary depending on what secondary suppressions are made. Let us analyze what happens with the secondary suppressions corresponding to each primary suppression:

1. Assume that, to protect $z_{a_1 s_1}$, we choose as secondary suppressions $z_{a_1 s_3}$, $z_{a_3 s_1}$ and $z_{a_3 s_3}$. Suppressing $z_{a_1 s_3}$ does not change the rank, because by Equation (11) we have

$$z_{a_1 s_3} = \sum_{k=1}^{t} z_{s_3 r_k} - \sum_{i \neq 1} a_{a_i s_3} \tag{16}$$

where the cells on the right-hand side have not yet been suppressed. Suppressing $z_{a_3 s_1}$ does not change the rank either, because by Equation (10):

$$z_{a_3 s_1} = \sum_{k=1}^{t} z_{a_3 r_k} - \sum_{j \neq 1} z_{a_3 s_j} \tag{17}$$

Suppression of $z_{a_3s_3}$ causes the rank to decrease by 1, because there is no way to express the suppressed cell as combination of other cells which have not yet been suppressed: if Equation (11) is used, $z_{a_1s_3}$ is necessary but has been suppressed already; if Equation (10) is used, then $z_{a_3s_1}$ is necessary which has been suppressed; for a similar reason we cannot use Equation (13). Note also that, once the suppression process has started, Equation (13) is not very useful to obtain linear combinations, because it depends on nearly all cells.

2. Now assume that, to protect $z_{a_1r_1}$, we choose as secondary suppressions $z_{a_1r_3}$, $z_{a_2r_1}$ and $z_{a_2r_3}$. When $z_{a_1r_3}$ is suppressed, the rank does not change, because by Equation (12)

$$z_{a_1r_3} = \sum_{j=1}^{m} z_{s_jr_3} - \sum_{i\neq 1} z_{a_ir_3} \tag{18}$$

Suppressing $z_{a_2r_1}$ does not change the rank either, because by Equation (10):

$$z_{a_2r_1} = \sum_{j=1}^{m} z_{a_2s_j} - \sum_{k\neq 1} z_{a_2r_k} \tag{19}$$

Note that, if we now choose to suppress $z_{a_3r_1}$, the rank would decrease by 1, because there is no way to express it as a combination of other cells not yet suppressed (if Equation (12) was used, $z_{a_1r_1}$ would be necessary and, if Equation (10) was used, then $z_{a_3s_1}$ would be necessary). So, *choosing for this suppression any row in Table AR other than row 3 (which was used in Table AS) does not decrease the rank and, consequently, adds hardly any protection.* Finally, using similar arguments, it is not difficult to see that suppression of $z_{a_2r_3}$ decreases the rank by 1 (this suppression is inevitable in order to protect $z_{a_2r_1}$ and $z_{a_1r_3}$).

3. As to the third primary suppression, assume that, to protect $z_{s_1r_1}$, we choose as secondary suppressions $z_{s_1r_2}$, $z_{s_4r_1}$ and $z_{s_4r_2}$. Suppressing $z_{s_1r_2}$ does not change the rank, because, by Equation (12):

$$z_{s_1r_2} = \sum_{i=1}^{n} z_{a_ir_2} - \sum_{j\neq 1} z_{s_jr_2} \tag{20}$$

Note that, if we chose $z_{s_1r_3}$, the rank would decrease by 1. *So, choosing for this suppression any column other than column 3 (which was used for Table AR) does not decrease the rank and adds virtually no protection.* Suppressing $z_{s_4r_1}$ does not change the rank either, because, by Equation (11):

$$z_{s_4r_1} = \sum_{i=1}^{n} z_{a_is_4} - \sum_{k\neq 1} z_{s_4r_k} \tag{21}$$

If we chose $z_{s_3r_1}$ instead of $z_{s_4r_1}$, then the rank would decrease by 1. *So, choosing for this suppression any row other than row 3 (which was used for Table AS when $z_{a_3s_1}$ was suppressed) does not decrease the rank and adds no real protection.* Finally, we have to suppress $z_{s_4r_2}$, which causes the rank to decrease by 1.

We have performed 12 suppressions altogether (primary and secondary), which decrease the rank of the system by 4. From the above, it is clear that the strategy we followed was to choose the suppression pattern which minimizes the decrease of the system rank and consequently minimizes the increase of entropy (and of protection!).

From the previous discussion, we can infer the following general result, whose proof is analogous to the proof of Proposition 1:

Proposition 2 *The entropy increase and the disclosure risk decrease in three linked tables are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same columns and rows for all three tables.*

A weaker necessary condition for maximizing the decrease of disclosure risk in the case of three tables is as follows:

Proposition 3 *Proposition 3 The entropy increase and the disclosure risk decrease in three linked tables are maximized if the suppression patterns are chosen so that the secondary suppressions are in the same row for tables with the same first variable —e.g. AS, AR—, the same column for tables with the same second variable —e.g. AR, SR—, and the same row for tables which share a variable in a different position —e.g. AS, SR—).*

Using such optimal patterns we can decrease the rank of the system of 3 linked tables by 3 more units (up to an overall rank decrease of 7).

4.5.4 Disclosure risk by internal intruders

Finally, a point we have to take into account here is that disclosure risk is different for different users. Let us imagine that, when solving one LP maximization and one LP minimization for $z_{a_i s_j r_k}$, we find that $995 \leqslant z_{a_j s_j r_k} \leqslant 1004$. Then, if company A is the second largest contributor to this cell with a turnover of, say, 400, then company A knows that the largest contributor (company B) has a turnover between 401 and 604. Thus, company A is able to estimate the turnover of company B within 50% of its value. However, the uncertainty of an external intruder about the turnover of company B is roughly 200% of its value: the external intruder only knows that the turnover of the largest contributor is between $\varepsilon > 0$ and 1004. Therefore, for an internal intruder (respondent contributing to the cell), the measure of disclosure risk $1/H(z_{a_i s_j r_k}|\text{released tables})$ should be replaced by:

$$DR(X) = 1/H(z_{a_i s_j r_k}|\text{released tables, intruder's contribution}) \qquad (22)$$

5 Conclusion

Due to the limitations of *a priori* disclosure risk assessment, *a posteriori* risk assessment has been proposed as a complement to *a priori* measures. That is, we propose to measure disclosure risk not only before the application of protection methods, but also after that. We have shown that the reciprocal of Shannon's conditional entropy (conditioned to the knowledge of the intruder) may be used as such a measure. While Shannon's entropy may not be suitable to evaluate the impact of disclosure control on table utility, it turns out to be extremely useful to quantify disclosure risk. As shown in Section 4, computing disclosure risk in this way can easily be done for different disclosure control methods, both with simple tables and linked tables. In the case of cell suppression methods applied to linked table protection, the proposed measure allows detection of secondary suppression patterns which offer more protection than others do. The strategy for choosing the best candidates for secondary suppressions has been outlined in the paper.

6 Acknowledgments

Thanks go to Sarah Giessing for useful comments on earlier versions of this paper.

7 References

Cox, L. H. (2001). Disclosure risk for tabular economic data, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 167-183.

Denning, D. E. (1982). *Cryptography and Data Security*. Reading, MA: Addison- Wesley.

Domingo-Ferrer, J. and Torra, V. (2002). A critique of the sensitivity rules usually employed for statistical table protection, in *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 5, 545-556.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 135-166.

Felsö, F., Theeuwes, J. and Wagner, G. G. (2001). Disclosure limitation methods in use: Results of a survey, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North- Holland, 17-42.

Fienberg, S. E., Makov, U. E. and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data, *Journal of Official Statistics*, 14, 485-512.

Giessing, S. (2001). Nonperturbative disclosure control methods for tabular data, in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, 185-213.

Holvast, J. (1999). Statistical dissemination, confidentiality and disclosure, in *Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality*. Luxembourg: Eurostat, 191-207.

Luige, T. and Meliskova, J. (1999). Confidentiality practices in the transition countries, in *Proceedings of the Joint Eurostat/UNECE Work Session on Statistical Data Confidentiality*. Luxembourg: Eurostat, 287-319.

Robertson, D. and Ethier, R. (2002). Cell suppression: Theory and experience, in *Inference Control in Statistical Databases*, LNCS 2316, ed. J. Domingo-Ferrer. Berlin: Springer-Verlag, 9-21.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, 27, 379-423, 623-656, July and Oct. 1948.

Willenborg, L. and de Waal, T. (2001). *Statistical Disclosure Control in Practice*. New York: Springer-Verlag.

Resum

La protecció de dades estadístiques, també coneguda com a Control de Revelació Estadística (SDC), és una part de la seguretat de la informació que intenta evitar la publicació d'informació estadística (taules, registres individuals) que reveli la contribució de responents específics. Aquest article tracta de la valoració del risc de revelació associat a la difusió de dades tabulades. Les anomenades regles de sensibilitat estan sent utilitzades actualment per tal de mesurar el risc de revelació en taules. Aquestes regles operen sobre una base *a priori*: les dades són examinades i les regles s'utilitzen per decidir si les dades poden ser difoses tal com s'han elaborat o bé han de ser protegides. En aquest article, proposem complementar la mesura de *risc a priori* amb una mesura de risc *a posteriori* per tal d'aconseguir un nivell de seguretat més alt, és a dir, proposem tenir en compte la informació protegida quan es mesura el risc de revelació.

La mesura del risc de revelació *a posteriori* proposada és compatible amb una àmplia classe de mètodes de protecció de revelació i pot ser aplicada al càlcul del risc de revelació d'un grup de taules vinculades. En el cas de protecció de taules vinculades a través de la supressió de cel·les, la mesura proposada permet la detecció de patrons de supressió secundària els quals ofereixen més protecció que d'altres.

MSC: 62P99

Paraules clau: control de revelació estadística, bases de dades estadístiques, dades tabulades, seguretat

Book reviews

SURVIVAL ANALYSIS. TECHNIQUES FOR CENSORED AND TRUNCATED DATA (2nd ed.)

John P. Klein & Melvin L. Moeschberger

Springer-Verlag, NewYork, 2003
535 pages

This book is a second edition of a good reference on survival analysis. It combines theoretical concepts with real data sets helping to understand key concepts defined in each section. Moreover there are two special subsections at the end of sections: *Practical Notes* and *Theoretical Notes.* The first one is about examples used in the literature related with the main issue of each section, or software indications and program code for techniques not include in standard software. The second one includes some theoretical extensions of the key concepts defined in the section. At the end of each chapter there is a good collection of exercises with a selection of solutions in Appendix E, a new section of this second edition of the book.

Because all of these points emphasized above, the book is suitable for teaching specialized courses on survival analysis and as a support in practical research, mainly in biology and medicine.

The book includes five major themes:

- Basic concepts and terminology
- Estimation of summary survival statistics based on censored and /or truncated data
- Hypothesis testing
- Regression analysis for censored and/or truncated data
- Multivariate models for survival data

These issues are divided into thirteen chapters summarized as follows:

The first chapter contains a brief introduction to censoring and presents 19 datasets of survival data used throughout the book.

Chapter 2 defines the basic tools used in modeling survival data as well as common parametric models for time and regression models for survival data with covariates. A new section about models for competing risks has been added to this second edition of the book.

Chapter 3 deals with the issued of censoring and truncation. Various categories of censoring are introduced, mainly centered on types of left and right censoring schemes. Truncation is also defined as a feature of survival data. The last two sections are about some theoretical results of survival analysis: likelihood construction for censored and truncated data and counting processes.

Chapter 4 is about nonparametric estimation of the distribution of time to some event, based on right-censored data. Apart from the known Kaplan-Meier curve, we emphasize the sections about confidence intervals for survival function and the point and interval estimates of mean and median survival time. This second edition of the book includes a new section on nonparametric estimation of time for the case of competing risks.

Basic tools for other types of censoring such as left, double or interval censored are introduced in chapter 5.

In chapter 6 there are two issues for the univariate estimation of the survival time: how crude estimates of the hazard rate can be smoothed to provide a better estimator of the hazard rate, and a Bayesian nonparametric approach as an alternative to the classical approach to estimating survival curves.

Hypothesis testing for survival and hazard functions are introduced in chapter 7. There is a detailed list of statistics used for one-sample tests and two and more sample tests to compare hazard rates and survival curves. The last section is new with respect to the previous edition. It introduces tests for comparing survival curves at a predetermined fixed point in time

Chapters 8 and 9 are about the proportional hazards model. Here we emphasize the second and the sixth sections in chapter 8. Section two is about coding and interpreting qualitative variables as covariates in a proportional hazards model. Section 6 is about discretizing continuous variables in order to draw conclusions like qualitative variables.

Main differences of this edition are in chapter 10 about additive hazards regression models. In this edition two models are presented: Aalen's nonparametric additive hazard model and Lin and Ying's additive hazard model. In this chapter there are included sections about additive hazards models of chapter 11 of the first edition.

Chapter 11 introduces methods to obtain regression diagnostics for the Cox model based on residual plots: checking the adequacy of the proportional hazards assumption, checking the accuracy of the proportional hazards model for predicting the survival of a given subject and, examining the influence that each subject has on the model fit.

Alternative regression models to Cox's proportional hazards are introduced in chapter 12. Apart from the usual accelerated failure-time models a linear model in log-time is also considered.

The last chapter deals with multivariate survival analysis. The starting point is on frailty models as a method to control the association between individual survival times. The final section gives a very brief introduction to marginal modeling for each individual.

The book finishes with appendices about specialized issues.

Anna Espinal
Servei d'Estadística
Universitat Autònoma de Barcelona
Spain

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.es) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a LaTeX 2_ε.

In any case, upon request the journal secretary will provide authors with LaTeX 2_ε templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

The publishing rights for SORT belong to the Institut d'Estadística de Catalunya since 1992 through express concession by the Technical University of Catalonia. The journal's current legal registry can be found under the reference number DL B-46.085-1977 and its ISSN is 1696-2281. Partial or total reproduction of this publication is expressly forbidden, as is any manual, mechanical, electronic or other type of storage or transmission without prior written authorisation from the Institut d'Estadística de Catalunya.

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9 (1), 73-84.

Books
Platek, R., Rao, J. N. K., Sărndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium.* New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.es

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name ____
____
Organisation ____
____
Street Address ____
____
Zip/Postal code ____________________ City ____________________
State/Country ____________________ Tel. ____________________
Fax ______________ NIF/VAT Registration Number ______________
E-mail ____
Date ____________________

Signature

I wish to subscribe to **SORT** ***(Statistics and Operations Research Transactions)*** for the year 2003 (volume 27)

Annual subscription rates:
- Spain: €22 (VAT included)
- Other countries: €25 (VAT included)

Price for individual issues (current and back issues):
- Spain: €9/issue (VAT included)
- Other countries: €11/issue (VAT included)

Method of payment:

□ Bank transfer to account number 2013-0100-53-0200698577

□ Automatic bank withdrawal from the following account number
□□□□ □□□□ □□ □□□□□□□□□□

□ Check made payable to the Institut d'Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

SORT ***(Statistics and Operations Research Transactions)***
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for
SORT ***(Statistics and Operations Research Transactions)***

The undersigned ____

authorises Bank/Financial institution ___________________________

located at (Street Address) ___________________________________

Zip/postal code ____________________ City ________________

Country ____

to draft the subscription to **SORT** ***(Statistics and Operations Research Transactions)*** from my account

number □□□□ □□□□ □□ □□□□□□□□□□

Date ________________________

Signature

SORT ***(Statistics and Operations Research Transactions)***
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Institut d'Estadística de Catalunya

The Institut d'Estadística de Catalunya (Idescat) is the statistical office of the Government of Catalonia. Its main duty is the management of the Catalan Statistical System by planning, coordinating and standardizing the statistical activity as well as providing statistical technical assistance.

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 412 09 24 - 93 412 00 88
Fax. +34 - 93 412 31 45
E-mail: sort@idescat.es

Publisher: Institut d'Estadística de Catalunya (Idescat)
© Institut d'Estadística de Catalunya
ISSN 1696-2281
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
http://www.idescat.es/sort

SORT

Volume 27 (2), July - December 2003
Formerly Qüestiió

Contents

Articles

ISSN: 1696-2281
http://www.idescat.es/sort