(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0168473
(43) 공개일자 2023년12월14일

(51) 국제특허분류(Int. Cl.)
G06F 17/15 (2006.01) *G06F 17/16* (2006.01)
G06F 17/18 (2006.01)
(52) CPC특허분류
G06F 17/15 (2013.01)
G06F 17/16 (2013.01)
(21) 출원번호 10-2022-0068967
(22) 출원일자 2022년06월07일
심사청구일자 2022년06월07일

(71) 출원인
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
유재근
서울특별시 서초구 서운로 62, 105동 2801호(서초동, 래미안 리더스원)
김보영
인천광역시 연수구 아카데미로 23(송도동)
(74) 대리인
특허법인 무한

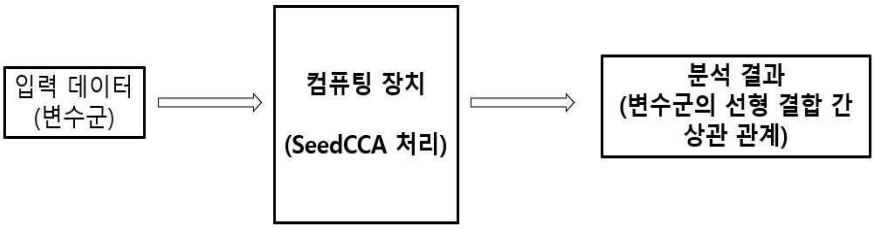
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **표준 상관 분석 및 부분 최소 제곱을 위한 SeedCCA 처리 방법 및 상기 방법을 수행하는 컴퓨팅 장치**

(57) 요 약

표준 상관 분석 및 부분 최소 제곱을 위한 SeedCCA 처리 방법 및 상기 방법을 수행하는 컴퓨팅 장치가 개시된다. SeedCCA 처리 방법은 제1 변수 집합과 제2 변수 집합을 식별하는 단계; 상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하는 단계; 상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 표준 CCA를 적용하는 단계를 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 17/18 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711147515
과제번호 2019R1F1A1050715
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 복합적 심화 충분차원축소 방법의 개발
기 여 율 1/1
과제수행기관명 이화여자대학교
연구기간 2021.03.01 ~ 2022.02.28

명 세 서

청구범위

청구항 1

컴퓨팅 장치가 수행하는 SeedCCA 처리 방법에 있어서,

제1 변수 집합과 제2 변수 집합을 식별하는 단계;

상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하는 단계;

상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 표준 CCA를 적용하는 단계

를 포함하는 SeedCCA 처리 방법.

청구항 2

제1항에 있어서,

상기 표준 CCA를 적용하는 단계는,

상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 대한 정규 변량쌍을 결정하는 단계;

상기 정규 변량쌍에 기초한 피어슨 상관 관계를 결정하는 단계;

상기 피어슨 상관 관계를 최대화하기 위한 표준 계수 행렬을 결정하는 단계

를 포함하는 SeedCCA 처리 방법.

청구항 3

제1항에 있어서,

제1 변수 집합의 차원이 p이고, 제2 변수 집합의 차원이 r이며, $\max(p,r) > n$ 인 경우에 SeedCCA가 적용되고,

상기 제1 변수 집합과 제2 변수 집합의 차원은 m>>n이 되도록 m차원으로 축소되는 SeedCCA 처리 방법.

청구항 4

제1항에 있어서,

상기 표준 CCA를 적용하는 단계는,

상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하는 초기화 과정 이후에 수정 및 직교성을 위해 표준 CCA를 적용하는 SeedCCA 처리 방법.

청구항 5

컴퓨팅 장치가 수행하는 SeedCCA 처리 방법에 있어서,

SeedCCA 함수의 type값을 식별하는 단계;

type값에 따라 표준 CCA, 보통 최소 제곱, 부분 최소 제곱, SeedCCA 중 어느 하나로 제1 변수 집합과 제2 변수 집합을 처리하는 단계;

상기 처리된 결과를 통해 서브 클래스를 리턴하는 단계

를 포함하는 SeedCCA 처리 방법.

청구항 6

제5항에 있어서,

상기 처리하는 단계는,

상기 SeedCCA 함수의 옵션을 고려하여 제1 변수 집합과 제2 변수 집합을 처리하는 SeedCCA 처리 방법.

청구항 7

제5항에 있어서,

상기 처리하는 단계는,

type값이 cca이고 max(p, r) < n 인 경우, 제1 변수 집합과 제2 변수 집합을 표준 CCA로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임)

상기 서브 클래스를 리턴하는 단계는, "finalCCA" 서브 클래스를 리턴하는 SeedCCA 처리 방법.

청구항 8

제5항에 있어서,

상기 처리하는 단계는,

type값이 cca이고 max(p, r) < n 및 min(p, r) = 1인 경우, 제1 변수 집합과 제2 변수 집합을 보통 최소 제곱으로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임)

상기 서브 클래스를 리턴하는 단계는, "seedols" 서브 클래스를 리턴하는 SeedCCA 처리 방법.

청구항 9

제5항에 있어서,

상기 처리하는 단계는,

type값이 seed1이고 max(p, r) ≥ n 인 경우, 제1 변수 집합과 제2 변수 집합을 SeedCCA로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임)

상기 서브 클래스를 리턴하는 단계는, "finalCCA" 서브 클래스를 리턴하는 SeedCCA 처리 방법.

청구항 10

제5항에 있어서,

상기 처리하는 단계는,

type값이 seed2이고 max(p, r) ≥ n 인 경우, 제1 변수 집합과 제2 변수 집합을 SeedCCA로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임)

상기 서브 클래스를 리턴하는 단계는, "finalCCA" 서브 클래스를 리턴하는 SeedCCA 처리 방법.

청구항 11

제5항에 있어서,

상기 처리하는 단계는,

type값이 pls이고 p ≥ n and r < n 인 경우, 제1 변수 집합과 제2 변수 집합을 부분 최소 제곱으로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임)

상기 서브 클래스를 리턴하는 단계는, "seedpls" 서브 클래스를 리턴하는 SeedCCA 처리 방법.

청구항 12

SeedCCA 처리 방법을 수행하는 컴퓨팅 장치는,

프로세서를 포함하고,

상기 프로세서는,

제1 변수 집합과 제2 변수 집합을 식별하고,

상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하고,

상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 표준 CCA를 적용하는 컴퓨팅 장치.

청구항 13

SeedCCA 처리 방법을 수행하는 컴퓨팅 장치는,

프로세서를 포함하고,

상기 프로세서는, SeedCCA 함수의 type값을 식별하고,

type값에 따라 표준 CCA, 보통 최소 제곱, 부분 최소 제곱, SeedCCA 중 어느 하나로 제1 변수 집합과 제2 변수 집합을 처리하고,

상기 처리된 결과를 통해 서브 클래스를 리턴하는 컴퓨팅 장치.

발명의 설명

기 술 분 야

배 경 기 술

[0001] 설명 연구는 특정 모델을 개발하기 전에 데이터의 패턴과 특수 구조를 식별하는데 중요하다. p차원 확률변수 X(X ∈ Rp)와 r차원 확률변수 Y(Y ∈ Rr)의 두 세트 간의 연구가 주요 관심사일 때 인기 있는 설명 통계 방법 중 하나는 표준 상관 분석(CCA)이다.

[0002] 표준 상관 분석(Canonical Correlation Analysis: CCA)은 고차원 데이터 분석에서 설명적인 통계 방법(explanatory statistical method) 중 하나이다. 표준 상관 분석은 화학계량학(chemometrics), 패턴 인식(pattern recognition), 게놈 염기서열 분석(genomic sequence analysis) 등 많은 분야에 적용되어 왔다.

[0003] CCA의 주요 목표는 두 개의 변수 집합 간의 연관성(association)을 측정함으로써 두 개의 변수 집합의 차원 축소이다. 피어슨 상관을 최대화함으로써 두 개의 변수 간의 선형 결합의 쌍이 구성된다. CCA는 고차원 데이터에 대한 차원 축소 도구로 사용될 뿐만 아니라 최소 제공 추정기와 연결될 수 있다. 그래서, CCA는 설명 및 차원 축소 방법 뿐만 아니라 최소 제공 추정의 대안으로 활용될 수 있다.

[0004] 이러한 CCA에서 특정 조건일 때에도 처리할 수 있는 방안이 필요하다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 표준 상관 분석과 부분 최소 제곱을 함께 고려할 수 있는 SeedCCA를 처리할 수 있는 방법 및 컴퓨팅 장치를 제공한다.

과제의 해결 수단

[0006] 본 발명의 일실시예에 따른 컴퓨팅 장치가 수행하는 SeedCCA 처리 방법은 제1 변수 집합과 제2 변수 집합을 식별하는 단계; 상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하는 단계; 및 상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 표준 CCA를 적용하는 단계를 포함한다.

[0007] 상기 표준 CCA를 적용하는 단계는, 상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 대한 정규 변량쌍을 결정하는 단계; 상기 정규 변량쌍에 기초한 피어슨 상관 관계를 결정하는 단계; 상기 피어슨 상관 관계를 최대화하기 위한 표준 계수 행렬을 결정하는 단계를 포함할 수 있다.

[0008] 제1 변수 집합의 차원이 p이고, 제2 변수 집합의 차원이 r이며, $\max(p,r) > n$ 인 경우에 SeedCCA가 적용되

고, 상기 제1 변수 집합과 제2 변수 집합의 차원은 m>>n이 되도록 m차원으로 축소될 수 있다.

[0009] 상기 표준 CCA를 적용하는 단계는, 상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하는 초기화 과정 이후에 수정 및 직교성을 위해 표준 CCA를 적용할 수 있다.

[0010] 본 발명의 다른 실시예에 따른 컴퓨팅 장치가 수행하는 SeedCCA 처리 방법은 SeedCCA 함수의 type값을 식별하는 단계; type값에 따라 표준 CCA, 보통 최소 제곱, 부분 최소 제곱, SeedCCA 중 어느 하나로 제1 변수 집합과 제2 변수 집합을 처리하는 단계; 상기 처리된 결과를 통해 서브 클래스를 리턴하는 단계를 포함할 수 있다.

[0011] 상기 처리하는 단계는, 상기 SeedCCA 함수의 옵션을 고려하여 제1 변수 집합과 제2 변수 집합을 처리할 수 있다.

[0012] 상기 처리하는 단계는, type값이 cca이고 max(p, r) < n 인 경우, 제1 변수 집합과 제2 변수 집합을 표준 CCA로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임) 상기 서브 클래스를 리턴하는 단계는, "finalCCA" 서브 클래스를 리턴할 수 있다.

[0013] 상기 처리하는 단계는, type값이 cca이고 max(p, r) < n 및 min(p, r) = 1인 경우, 제1 변수 집합과 제2 변수 집합을 보통 최소 제곱으로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임) 상기 서브 클래스를 리턴하는 단계는, "seedols" 서브 클래스를 리턴할 수 있다.

[0014] 상기 처리하는 단계는, type값이 seed1이고 max(p, r) ≥ n 인 경우, 제1 변수 집합과 제2 변수 집합을 SeedCCA로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임) 상기 서브 클래스를 리턴하는 단계는, "finalCCA" 서브 클래스를 리턴할 수 있다.

[0015] 상기 처리하는 단계는, type값이 seed2이고 max(p, r) ≥ n 인 경우, 제1 변수 집합과 제2 변수 집합을 SeedCCA로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임) 상기 서브 클래스를 리턴하는 단계는, "finalCCA" 서브 클래스를 리턴할 수 있다.

[0016] 상기 처리하는 단계는, type값이 pls이고 p ≥ n and r < n 인 경우, 제1 변수 집합과 제2 변수 집합을 부분 최소 제곱으로 처리하고, (여기서, p는 제1 변수 집합의 차원이고, r은 제2 변수 집합의 차원이며, n은 임의의 수임) 상기 서브 클래스를 리턴하는 단계는, "seedpls" 서브 클래스를 리턴할 수 있다.

[0017] 본 발명의 일실시예에 따른 SeedCCA 처리 방법을 수행하는 컴퓨팅 장치는, 프로세서를 포함하고, 상기 프로세서는, 제1 변수 집합과 제2 변수 집합을 식별하고, 상기 제1 변수 집합과 제2 변수 집합의 차원을 축소하고, 상기 차원이 축소된 제1 변수 집합과 제2 변수 집합에 표준 CCA를 적용할 수 있다.

[0018] 본 발명의 일실시예에 따른 SeedCCA 처리 방법을 수행하는 컴퓨팅 장치는, 프로세서를 포함하고, 상기 프로세서는, SeedCCA 함수의 type값을 식별하고, type값에 따라 표준 CCA, 보통 최소 제곱, 부분 최소 제곱, SeedCCA 중 어느 하나로 제1 변수 집합과 제2 변수 집합을 처리하고, 상기 처리된 결과를 통해 서브 클래스를 리턴할 수 있다.

발명의 효과

[0019] 본 발명의 일실시예에 따르면, 표준 상관 분석과 부분 최소 제곱을 함께 고려할 수 있는 SeedCCA를 처리할 수 있다.

도면의 간단한 설명

[0020] 도 1은 본 발명의 일실시예에 따른 SeedCCA를 처리하는 컴퓨팅 장치를 도시한 도면이다.

도 2는 본 발명의 일실시예에 따른 변수 집합에 기초한 상관 관계를 도출하는 CCA 과정을 설명하는 도면이다.

도 3은 본 발명의 일실시예에 따른 펄프 데이터에 대한 누적 CCA 그래프를 도시한 도면이다.

도 4는 본 발명의 일실시예에 따른 SeedCCA로부터 생성된 u에 대한 그래프를 도시한 도면이다.

도 5는 본 발명의 일실시예에 따른 고유벡터의 집합을 선택하기 위한 그래프를 도시한 도면이다.

도 6은 본 발명의 일실시예에 따른 SeedCCA로부터 생성된 ux 및 uy에 대한 그래프를 도시한 도면이다.

도 7은 본 발명의 일실시예에 따른 정규화 CCA와 SeedCCA로부터 canonical variates의 제1 쌍을 도시한 도면이

다.

도 8은 본 발명의 일실시예에 따른 정규화 CCA와 SeedCCA로부터 canonical variates의 제2 쌍을 도시한 도면이다.

도 9는 본 발명의 일실시예에 따른 서로 다른 조건에서 SeedCCA로부터 생성된 u에 대한 그래프를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0021] 이하, 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0022] 아래 설명하는 실시예들에는 다양한 변경이 가해질 수 있다. 아래 설명하는 실시예들은 실시 형태에 대해 한정하려는 것이 아니며, 이들에 대한 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0023] 제1 또는 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 이해되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0024] 실시예에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 실시예를 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수 개의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0025] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0026] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0027] 이하, 본 발명의 실시예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0028] 도 1은 본 발명의 일실시예에 따른 SeedCCA를 처리하는 컴퓨팅 장치를 도시한 도면이다.

[0029] seedCCA는 새로운 R 패키지로 표준 및 시드된 CCA 뿐만 아니라 부분 최소 제곱도 수행 가능하다. 새로운 R 패키지는 CCA를 large-p 데이터와 small-n 데이터에 맞출 수 있다. 시드된 CCA는 기존의 R 패키지에 있는 정규화된 CCA와 비교될 수 있다.

[0030] $\max(p,r) > n$ 인 경우를 처리하기 위해 시드된 CCA인 SeedCCA가 제안되었다. seed CCA는 2가지 과정이 진행된다. 초기 단계에서, n보다 큰 변수 집합은 반복 예측(iterative projection)에 기초하여 감소될 수 있다. 그리고, 초기 과정을 통해 획득한 두 개의 변수 집합에 표준적인 CCA를 적용함으로써 데이터의 CCA가 완성된다. SeedCCA의 과정이 large-p 데이터와 small-n 데이터에 대해 널리 사용되는 통계 방법 중 하나인 부분 최소 제곱과 밀접한 관련성이 있다. 그래서, SeedCCA는 PLS 추정치를 산출할 수 있다. 본 발명의 일실시예에 따른 컴퓨팅 장치는 seedCCA를 처리하는 과정을 설명한다.

[0031] 도 2는 본 발명의 일실시예에 따른 변수 집합에 기초한 상관 관계를 도출하는 CCA 과정을 설명하는 도면이다.

[0032] <CCA 과정>

[0033] 본 발명에서는 다음과 같은 표현을 사용한다. p차원 확률 변수 X는 $\mathbf{X} \in \mathbb{R}^p$ 로 표시된다. 따라서 $\mathbf{X} \in \mathbb{R}^p$ 는 확률변수를 의미한다. $\mathbf{X} \in \mathbb{R}^p$ 및 $\mathbf{Y} \in \mathbb{R}^r$ 에 대해, $\mathrm{cov}(\mathbf{X}) = \boldsymbol{\Sigma}_x$, $\mathrm{cov}(\mathbf{Y}) = \boldsymbol{\Sigma}_y$ $\mathrm{cov}(\mathbf{X},\mathbf{Y}) = \boldsymbol{\Sigma}_{xy}$ 및

$\mathrm{cov}(\mathbf{Y},\mathbf{X}) = \mathbf{\Sigma}_{yx}$ 가 정의된다. 또한, $\mathbf{\Sigma}_x$ 및 $\mathbf{\Sigma}_y$ 는 양의 정부호라고 가정한다.

[0034] 도 2에 의하면, 표준 상관 분석(Canonical correlation analysis: CCA)이 수행된다. 변수 $\mathbf{X} \in \mathbb{R}^p$ 및 $\mathbf{Y} \in \mathbb{R}^r$ 의 두 집합을 가정하고 $U = \mathbf{a}^{\mathrm{T}}\mathbf{X}$ 및 $V = \mathbf{b}^{\mathrm{T}}\mathbf{Y}$ 의 선형 조합이 고려된다. 그러면 $\mathrm{var}(U) = \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{a}$, $\mathrm{var}(V) = \mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_y\mathbf{b}$, $\mathrm{cov}(U,V) = \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_{xy}\mathbf{b}$ 가 됩니다. 여기서 $\mathbf{a} \in \mathbb{R}^{p\times 1}$ 및 $\mathbf{b} \in \mathbb{R}^{r\times 1}$ 이다. U와 V사이의 상관 관계인 Pearson-correlation은 수학식 1과 같다.

수학식 1

[0035]
$$\mathrm{cor}(U,V) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_{xy}\mathbf{b}}{\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{a}}\sqrt{\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}_y\mathbf{b}}}$$

[0036] CCA 과정은 $\mathrm{cor}(U,V)$ 를 최대화하기 위한 a 및 b가 도출된다.

[0037] (1) 컴퓨팅 장치는 수학식 1의 상관 관계를 최대함으로써 제1 정규 변량 쌍(canonical variate pair) $(U_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{X}, V_1 = \mathbf{b}_1^{\mathrm{T}}\mathbf{Y})$ 을 도출할 수 있다.

[0038] (2) 컴퓨팅 장치는 $\mathrm{var}(U_2) = \mathrm{var}(V_2) = 1$ 이고, (U_1,V_1) 및 (U_2,V_2) 가 서로 상관되지 않는다는 제한 조건에서 제2 정규 변량 쌍 $(U_2 = \mathbf{a}_2^{\mathrm{T}}\mathbf{X}, V_2 = \mathbf{b}_2^{\mathrm{T}}\mathbf{Y})$ 를 도출할 수 있다.

[0039] (3) 컴퓨팅 장치는 k단계에서 $\mathrm{var}(U_k) = \mathrm{var}(V_k) = 1$ 이고, K번째 정규 변량쌍인 (U_k, V_k) 는 K-1번째 정규 변량쌍과 상관되지 않는다는 제한 조건에서 K 번째 정규 변량쌍 $(U_k = \mathbf{a}_k^{\mathrm{T}}\mathbf{X}, V_k = \mathbf{b}_k^{\mathrm{T}}\mathbf{Y})$ 을 도출할 수 있다.

[0040] (4) 컴퓨팅 장치는 K가 $q\,(=\min(p,r))$ 가 될 때까지 (1) 내지 (3)의 과정을 반복할 수 있다.

[0041] (5) 컴퓨팅 장치는 X와 Y의 상관 관계를 나타내기 위해 (U_k,V_k) 의 첫번째 d 쌍을 선택할 수 있다.

[0042] 위에서 설명한 과정을 통해 쌍 $(\mathbf{a}_i,\mathbf{b}_i)$ 는 $i = 1,\ldots,q$ 에 대해 $\mathbf{a}_i = \mathbf{\Sigma}_x^{-1/2}\psi_i$ 과 $\mathbf{b}_i = \mathbf{\Sigma}_y^{-1/2}\phi_i$ 로 결정된다. 이 때, $(\psi_1,\ldots,\psi_q)$ 및 $(\phi_1,\ldots,\phi_q)$ 는 $\rho_1^{*2} \geq \cdots \geq \rho_q^{*2} \geq 0$ 의 대응 공통 차수 고유값에서 각각 $\mathbf{\Sigma}_x^{-1/2}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_y^{-1/2}$ 및 $\mathbf{\Sigma}_y^{-1/2}\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}\mathbf{\Sigma}_x^{-1/2}$ 의 q 고유 벡터이다.

[0043] 그러면, 행렬 $\mathbf{M}_x = (\mathbf{a}_1,\ldots,\mathbf{a}_d)$ 및 $\mathbf{M}_y = (\mathbf{b}_1,\ldots,\mathbf{b}_d)$ 는 $d = 1,\ldots,q$ 에 대한 표준 계수 행렬(canonical coefficient matrices)로 정의된다. 그리고, $\mathbf{M}_x^{\mathrm{T}}\mathbf{X}$ 및 canonical coefficient matrices는 표준 변수(canonical variate)로 정의된다.

[0044] <SeedCCA 과정>

[0045] 표준 CCA는 $\hat{\mathbf{\Sigma}}_x$ 와 $\hat{\mathbf{\Sigma}}_y$ 의 역전(inversion)을 필요하기 때문에 $\max(p,r) > n$ 인 고차원 데이터에는 적절하지 않다. $\max(p,r) > n$ 인 고차원 데이터에 적용하기 위해 본 발명의 일실시예에 따른 SeedCCA가 제안되었다.

[0046] SeedCCA는 아래와 같은 2단계를 거쳐서 진행된다.

[0047] (1) 초기 단계:

[0048] 두 개의 변수 집합은 CCA에 대한 정보 손실 없이 m 차원의 쌍으로 축소된다. 초기 단계에서, $m << n$을 강제하는 것이 필요하다.

[0049] (2) 최종 단계:

[0050] 표준 CCA는 리페어 및 직교 정규성을 위해 초기 단계에서 차원이 축소된 쌍에 적용된다.

[0051] $\mathcal{S}(\mathbf{M})$는 $\mathbf{M} \in \mathbb{R}^{p \times r}$의 열에 걸쳐 있는 부분 공간(subspace)으로 정의된다. $\mathcal{S}(\mathbf{M})$는 수학식 2에 따라 정의된다.

수학식 2

[0052] $$\mathcal{S}(\mathbf{M}_x) = \mathcal{S}(\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}) \text{ and } \mathcal{S}(\mathbf{M}_y) = \mathcal{S}(\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx})$$

[0053] 수학식 2에 의하면, $\mathbf{M}_x$와 $\mathbf{M}_y$가 $\mathcal{S}(\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy})$와 $\mathcal{S}(\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx})$의 기저 행렬(basis matrix)을 형성하고, $\mathbf{M}_x$와 $\mathbf{M}_y$가 $\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}$와 $\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx}$에서 복원될 수 있다.

[0054] 행렬 $\mathbf{R}_{x,u_1}$와 $\mathbf{R}_{y,u_2}$는 수학식 3에 따라 정의된다.

수학식 3

[0055] $$\begin{aligned} \mathbf{R}_{x,u_1} \in \mathbb{R}^{p \times ru_1} &= (\mathbf{\Sigma}_{xy}, \mathbf{\Sigma}_x\mathbf{\Sigma}_{xy}, \ldots, \mathbf{\Sigma}_x^{u_1-1}\mathbf{\Sigma}_{xy}) \\ \mathbf{R}_{y,u_2} \in \mathbb{R}^{r \times pu_2} &= (\mathbf{\Sigma}_{yx}, \mathbf{\Sigma}_y\mathbf{\Sigma}_{yx}, \ldots, \mathbf{\Sigma}_y^{u_2-1}\mathbf{\Sigma}_{yx}) \end{aligned}$$

[0056] $\mathbf{R}_{x,u_1}$ 및 $\mathbf{R}_{y,u_2}$에서, u_1 및 u_2의 숫자를 종료 인덱스(termination index)라고 정의한다. u_1 및 u_2의 숫자들은 각각 $\mathbf{\Sigma}_x$와 $\mathbf{\Sigma}_y$에 대한 $\mathbf{\Sigma}_{xy}$와 $\mathbf{\Sigma}_{yx}$의 투영 수를 결정한다.

[0057] $\mathbf{M}_{x,u_1}^0$와 $\mathbf{M}_{y,u_2}^0$는 수학식 4에 따라 정의된다.

수학식 4

[0058] $$\begin{aligned} \mathbf{M}_{x,u_1}^0 \in \mathbb{R}^{p \times r} &= \mathbf{R}_{x,u_1}(\mathbf{R}_{x,u_1}^{\mathrm{T}}\mathbf{\Sigma}_x\mathbf{R}_{x,u_1})^{-1}\mathbf{R}_{x,u_1}^{\mathrm{T}}\mathbf{\Sigma}_{xy} \\ \mathbf{M}_{y,u_2}^0 \in \mathbb{R}^{r \times p} &= \mathbf{R}_{y,u_2}(\mathbf{R}_{y,u_2}^{\mathrm{T}}\mathbf{\Sigma}_y\mathbf{R}_{y,u_2})^{-1}\mathbf{R}_{y,u_2}^{\mathrm{T}}\mathbf{\Sigma}_{yx} \end{aligned}$$

[0059] 수학식 4에서, $\mathcal{S}(\mathbf{M}_{x,u_1}^0) = \mathcal{S}(\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy})$, $\mathcal{S}(\mathbf{M}_{y,u_2}^0) = \mathcal{S}(\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_{yx})$로 정의된다. 그래서, $\mathbf{M}_{x,u_1}^0$ 및 $\mathbf{M}_{y,u_2}^0$는 각각 $\mathbf{M}_x$ 및 $\mathbf{M}_y$를 추론하는 데 사용된다. $\mathbf{M}_{x,u_1}^0$ 및 $\mathbf{M}_{y,u_2}^0$를 사용하는 장점은 $\mathbf{\Sigma}_x$ 및 $\mathbf{\Sigma}_y$를 반전(inversion)할 필요가 없다는 것이다. 실제로, $\Delta_{x,u_1} = \mathbf{M}_{x,u_1+1}^0 - \mathbf{M}_{x,u_1}^0$ 및 $\Delta_{y,u_2} = \mathbf{M}_{y,u_2+1}^0 - \mathbf{M}_{y,u_2}^0$가 정의되므로, 종료 인덱스인 u_1 및 u_2에 대해 적절한 값을 선택하는 것이 중요하다.

[0060] 또한, u_1 및 u_2의 증가에 대한 측정값이 수학식 5와 같이 정의된다.

수학식 5

[0061] $$nF_{x,u_1} = n\text{trace}(\boldsymbol{\Delta}_{x,u_1}^{\mathrm{T}}\boldsymbol{\Sigma}_x\boldsymbol{\Delta}_{x,u_1})$$

[0062] $$nF_{y,u_2} = n\text{trace}(\boldsymbol{\Delta}_{y,u_2}^{\mathrm{T}}\boldsymbol{\Sigma}_y\boldsymbol{\Delta}_{y,u_2})$$

[0063] nF_{x,u_1} 및 nF_{x,u_1+1} 및 nF_{y,u_2} 및 nF_{y,u_2+1} 의 변화가 거의 없도록, u의 적절한 값이 선택될 필요가 있다. $\mathbf{M}_{x,u_1}^0$ 및 $\mathbf{M}_{y,u_2}^0$ 에 대해 선택된 u_1 및 u_2 가 공통일 필요는 없다.

[0064] $\mathbf{X}$ 와 $\mathbf{Y}$ 의 원래 두 변수 셋트는 $\mathbf{M}_{x,u_1}^{0\,\mathrm{T}}\mathbf{x} \in \mathbb{R}^r$ 및 $\mathbf{M}_{y,u_2}^{0\,\mathrm{T}}\mathbf{Y} \in \mathbb{R}^p$ 로 대체된다. $\mathbf{X}$와 $\mathbf{Y}$ 의 감소는 CCA에 대한 정보 손실이 발생되지 않는다. $\mathcal{S}(\mathbf{M}_{x,u_1}^0) = \mathcal{S}(\mathbf{M}_x)$ 및 $\mathcal{S}(\mathbf{M}_{y,u_2}^0) = \mathcal{S}(\mathbf{M}_y)$ 이며, 초기화 CCA라고 합니다. 초기화 CCA는 아래와 같은 두 가지 경우가 있다.

[0065] 사례 1: $\min(p,r) = r << n$ 이라고 가정한다. 이 때, 원래 X만 $\mathbf{M}_{x,u_1}^{0\,\mathrm{T}}\mathbf{X}$ 로 대체되고, 원래 Y는 유지된다.

[0066] 사례 2: $\min(p,r) = r$ 이 n보다 상당히 작지 않으면, $\mathbf{R}_{x,u_1}$, $\mathbf{R}_{y,u_2}$, $\mathbf{M}_{x,u_1}^0$ 및 $\mathbf{M}_{y,u_2}^0$ 구성에서 $\boldsymbol{\Sigma}_{xy}$ 및 $\boldsymbol{\Sigma}_{yx}$ 가 m개의 가장 큰 고유 벡터로 대체된다.

[0067] M을 적절한 값으로 결정하기 위해, 다음 두 가지 방법이 이용된다. 한가지 방법은 $\boldsymbol{\Sigma}_{xy}$의 고유값에 대한 스크리 플롯(scree plot)에 의한 그래픽적인 결정이다. 나머지 방법은 고유값의 수가 $\boldsymbol{\Sigma}_{xy}$의 전체 변동의 90% 이상을 포함하도록 하는 것이다. 초기화 과정의 주요 목적은 CCA에 대한 정보 손실 없이 X 및 Y를 n 미만으로 줄이는 것이다. 사례 1의 경우, X와 Y는 r차원 변수로 축소되지만, 사례 2의 경우, X와 Y는 m차원의 변수 집합으로 대체됩니다. 초기화 과정 이후에, r과 m은 n보다 상당히 작다.

[0068] 초기화 과정 이후의 다음 단계는 수정(repairing) 및 직교성을 위해 $\mathbf{M}_{x,u_1}^{0\,\mathrm{T}}\mathbf{X}$ 및 $\mathbf{M}_{y,u_2}^{0\,\mathrm{T}}\mathbf{Y}$에 대한 표준 CCA를 수행하는 것이다. 이와 같은 CCA 어플리케이션이 최종 CCA이다. 마지막으로 CCA에 대한 2단계를 시드 CCA라고 정의한다.

[0069] <부분 최소 제곱 (Partial least squares)>

[0070] 표준 CCA 및 시드 CCA의 주요 목표는 $\mathbf{Y}|\mathbf{X}$의 조건부 관계가 아닌 X와 Y의 결합 관계를 기반으로 차원 축소하는 것이다. 단순화를 위해, 하위 섹션에서 r = 1인 Y를 가정한다. Y는 $\mathbf{Y}|\mathbf{X}$ 의 회귀에서 응답 변수로 사용된다.

[0071] 컴퓨팅 장치는 수학식 (3)에서 $\mathbf{R}_{x,u_1}$ 를 호출하고, 수학식 (4)에서 $\mathbf{M}_{x,u_1}^0$를 호출한다. 이 때, $\mathbf{R}_{x,u_1}$와 $\mathbf{M}_{x,u_1}^0$ 는 수학식 6과 같이 결정된다.

수학식 6

[0072] $$\mathbf{R}_{x,u_1} = (\boldsymbol{\Sigma}_{xy}, \boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_{xy}, \boldsymbol{\Sigma}_x^2\boldsymbol{\Sigma}_{xy}, \ldots, \boldsymbol{\Sigma}_x^{u_1-1}\boldsymbol{\Sigma}_{xy})$$

[0073] $$\mathbf{M}^{0}_{x,u_1} = \mathbf{R}_{x,u_1}(\mathbf{R}^{\mathrm{T}}_{x,u_1}\boldsymbol{\Sigma}_x\mathbf{R}_{x,u_1})^{-1}\mathbf{R}^{\mathrm{T}}_{x,u_1}\boldsymbol{\Sigma}_{xy}$$

[0074] $\mathbf{Y}|\mathbf{X}$의 회귀에 대한 u 성분이 있는 모집단 부분 최소 제곱(Partial Least Sqaure: PLS)은 수학식 7과 같다.

수학식 7

[0075] $$\boldsymbol{\beta}_{u_1,\mathrm{PLS}} = \mathbf{M}^{0}_{x,u_1}$$

[0076] 수학식 7에서, PLS 표현은 SeedCCA를 통한 X에 대한 표준 행렬과 동일하다.

[0077] SeedCCA 패키지의 아웃 라인이 이하에서 설명된다. SeedCCA의 메인 함수를 통해 위에서 설명한 과정들이 수행된다. SeedCCA의 메인 함수는 다음과 같다.

[0078] seedCCA(X, Y, type="seed2", ux=NULL, uy=NULL, u=10, eps=0.01, cut=0.9, d=NULL, AS=TRUE, scale=FALSE)

[0079] SeedCCA의 메인 함수는 type 값에 따라 "SeedCCA" 클래스와 3가지 서브 클래스를 리턴한다. Type값과 서브 클래스는 아래와 같이 정의된다.

[0080] type="cca": standard CCA (max(p, r) < n and min(p, r) > 1) / "finalCCA" subclass

[0081] type="cca": ordinary least squares (max(p, r) < n and min(p, r) = 1) / "seedols" subclass

[0082] type="seed1": seeded CCA with case1 (max(p, r) ≥ n) / "finalCCA" subclass

[0083] type="seed2": seeded CCA with case2 (max(p, r) ≥ n) / "finalCCA" subclass

[0084] type="pls": partial least squares (p ≥ n and r < n) / "seedpls" subclass

[0085] 함수 SeedCCA는 모든 서브 클래스에 대해 추정된 정규 계수 행렬을 출력하고, "finalCCA" 서브 클래스에 대해 표준 상관을 추가로 수행한다.

[0086] 표준 CCA는 펄프 데이터를 사용하였다. SeedCCA의 경우 정규화된 CCA 및 부분 최소 제곱과의 비교와 함께, 시드CCA 패키지에서 쿠키(cookie) 및 뉴트리무진(nutrimouse)이 설명된다.

[0087] 도 3은 본 발명의 일실시예에 따른 펄프 데이터에 대한 누적 CCA 그래프를 도시한 도면이다.

[0088] 펄프 데이터는 펄프 섬유와 펄프 섬유로 만든 종이의 특성을 측정한 것이다. 펄프 데이터는 62개의 샘플 크기를 가진 두 개의 변수 집합이 포함된다. 가 포함됩니다. 첫 번째 변수 집합인 Y는 산술 섬유 길이(arithmetic fiber length), 장섬유 분율(long fiber fraction), 미세 섬유 분율(fine fiber fraction) 및 제로 스팬 인장(zero spans tensile)과 같은 펄프 섬유의 특성에 대한 것입니다. 그리고, 두번째 변수 집합인 X는 파단 길이(breaking length), 탄성 모듈러스(elastic modulus), 파손 시 스트레스(stress at failure) 및 버스트 세기(burst strength)와 같은 종이의 속성에 관한 것이다. 표준 CCA 응용 어플리케이션을 수행하기 위해, type="cca"인 seedCCA 함수가 사용될 필요가 있다.

[0089] 이 경우, SeedCCA 함수는 "finalCCA" 서브 클래스를 결과로 출력한다. SeedCCA함수는 두 가지 행렬 유형을 요구하며, cor, xcoef, ycoef, Xscores 및 Yscores의 다섯 가지 구성 요소를 반환한다. 첫 번째 구성 요소는 샘플 표준 상관 관계(coris)이다. 그리고, 다음 두 가지 구성 요소인 xcoef 및 ycoef는 X 및 Y에 대한 추정된 표준 행렬(canonical matrices)이다. 마지막 두 가지 구성 요소인 Xscoresand Yscores는 X와 Y에 대한 추정된 표준 변량(canonical variates)이다. 명령어 plot(object)은 표준 쌍(canonical pair)의 수에 대한 누적 상관 관계의 플롯을 구성한다. 명령어 plot(object)는 기본적으로 90% 기준선을 제공하며, plot(object,ref=percent)으로 기준선이 변경될 수 있다.

[0090] ## loading pulp data

[0091] > pulp <- read.table("http://home.ewha.ac.kr/~yjkstat/pulp.txt", header=TRUE)

[0092] > Y <- as.matrix(pulp[,1:4])

[0093] > X <- as.matrix(pulp[,5:8])

[0094] ## standard CCA for X and Y

[0095] > fit.cca <- seedCCA(X, Y, type="cca")

[0096] NOTE: The standard CCA is fitted for the two sets.

[0097] > names(fit.cca)

[0098] [1] "cor" "xcoef" "ycoef" "Xscores" "Yscores"

[0099] ## plotting cumulative canonical correlation

[0100] > par(mfrow=c(1, 2))

[0101] > plot(fit.cca, ref=80)

[0102] > plot(fit.cca)

[0103] ## first two canonical pairs

[0104] > X.cc <- fit.cca$Xscores[,1:2]

[0105] > Y.cc <- fit.cca$Yscores[,1:2]

[0106] 80% 누적 정규 상관 관계가 있는 도 3의 왼쪽 그래프와 오른쪽 그래프에 의하면, 2개의 표준 쌍으로 충분하지만 90% 누적 정규 상관 관계에서는 3개의 정규 쌍이 양호해야 한다.

[0107] <보통최소제곱: 펄프 데이터>

[0108] X 또는 Y의 차원이 1인 경우, 표준 CCA에서 추정된 표준 계수 행렬은 일반 최소 제곱에서와 동일하다. 이 경우, 명령어 seedCCA(X,Y[,1],type="cca")는 "seedols" 서브 클래스인 일반적인 최소 제곱 추정값을 생성한다. "seedols"의 출력은 추정된 계수와 두 개의 변수 집합인 세 가지 구성요소가 있다.

[0109] 예를 들어, 주어진 X에서 Y의 첫 번째 열인 산술 섬유 길이의 회귀 연구가 특정 관심 대상이라고 가정한다. seedCCA(X,Y[,1],type="cca") 및 seedCCA(Y[,1],X,type="cca")의 순서는 중요하지 않으며, 어느 것이나 같은 결과를 생성한다. 또한 명령어 coef(object) 및 fit(object)은 각각 일반 최소 제곱에서 추정된 계수와 적합한 값을 반환한다.

[0110] ## extracting arithmatic fiber from Y

[0111] > fit.ols <- seedCCA(X, Y[, 1], type="cca")

[0112] NOTE: One of the two sets are 1-dimensional, so a linear regression via ordinary least

[0113] square is fitted.

[0114] > names(fit.ols)

[0115] [1] "coef" "X" "Y"

[0116] > coef(fit.ols)

[0117] > fitted(fit.ols)

[0118] 도 4는 본 발명의 일실시예에 따른 SeedCCA로부터 생성된 u에 대한 그래프를 도시한 도면이다.

[0119] <Seed CCA (case 1): cookie data>

[0120] seedCCA에서, 쿠키라고 하는 비스킷의 반죽 데이터 세트는 NIR 분광법으로 비스킷의 구성을 분석하는 실험에서 도출된 것이다. 72개의 비스킷에서 컴퓨팅 장치는 두 개의 변수 집합을 획득한다.

[0121] 첫번째 변수 집합은 분광법으로 측정한 파장이다. 원래의 데이터 세트에서 1100~2798nm(나노미터) 사이에 2nm 간격으로 700개의 서로 다른 지점에서의 파장이 측정된다. 다만, 1380nm에서 2400nm까지는 4nm 간격으로 파장

이 측정되었다.

[0122] 두번째 변수 집합은 지방, 당분, 마른 밀가루, 물을 포함하는 4가지 비스킷 재료이다. p = 256 > n = 72이므로, 컴퓨팅 장치는 표준 CCA를 수행할 수 없으며, n = 72 > r = 4를 고려하여 seedCCA의 경우 1을 맞춰야 한다. 기본 명령은 seedCCA(X,Y,type="seed1")이며, 결과적으로 "finalCCA" 서브 클래스가 생성된다. X와 Y의 순서에 관계없이 초기 단계에서 하위 차원 집합만 축소된다.

[0123] 따라서 seedCCA(X,Y,type="seed1")와 seedCCA(Y,X,type="seed1")는 기본적으로 동일하게 생성된다. type="seed1"의 경우, ux, uy, u, eps 및 AS 옵션 값이 SeedCCA의 수행 과정에 영향을 미치며, 옵션들 각각의 기본값은 각각 NULL, NULL, 10, 0.01 및 TRUE으로 설정된다.

[0124] 옵션 u는 ux와 uya가 모두 지정되지 않는 한 최대 투영 수를 제어한다. ux=k 옵션은 첫 번째 변수 집합 X의 차원이 두 번째 변수 집합 Y의 차원보다 클 때만 작동한다. 그러면 최대 투영 수는 ux=k에 주어진 값이 된다. uy 옵션은 ux와 반대 방식으로 작동합니다. AS=TRUE 및 eps 옵션은 u, ux 또는 uy에 지정된 최대값에 도달하기 전에 투영의 자동 종료를 제어한다. 증가가 eps에 지정된 값보다 작아지면 투영이 종료된다.

[0125] 그럼 다음, 중지 기준을 만족하는 첫 번째 후보 값이 투영의 예측값으로 제안된다. ux, uy, uy 중 하나라도 자동 정지를 보장하기에 충분하지 않게 지정되는 경우, 컴퓨팅 장치는 이들을 증가시키라는 알림을 제공한다. seedCCA(X,Y,type="seed1")이 실행되면, 컴퓨팅 장치는 적절한 u 를 선택하기 위한 그래프를 자동으로 구성한다. 도 2의 그래프에서 파란색 세로 막대가 제안된 u 값을 나타낸다.

[0126] ## loading cookie data

[0127] > data(cookie)

[0128] > myseq<-seq(141, 651, by=2)

[0129] > A <- as.matrix(cookie[-c(23, 61), myseq])

[0130] > B <- as.matrix(cookie[-c(23, 61), 701:704])

[0131] ## seedec CCA with case 1

[0132] > fit.seed1.ab <- seedCCA(A, B, type="seed1") ## the first set A has been initial-CCAed.

[0133] NOTE: Seeded CCA with case 1 is fitted. The set with larger dimension is initially reduced.

[0134] The first and second sets are denoted as X and Y, respectively.

[0135] > fit.seed1.ba <- seedCCA(B, A, type="seed1") ## the second set A has been initial-CCAed.

[0136] NOTE: Seeded CCA with case 1 is fitted. The set with larger dimension is initially reduced.

[0137] The first and second sets are denoted as X and Y, respectively.

[0138] > names(fit.seed1.ab)

[0139] [1] "cor" "xcoef" "ycoef" "proper.u" "initialMX0" "newX" "Y" "Xscores" "Yscores"

[0140] > names(fit.seed1.ba)

[0141] [1] "cor" "xcoef" "ycoef" "proper.u" "X" "initialMY0" "newY" "Xscores" "Yscores"

[0142] > fit.seed1.ab$xcoef[, 3] <- -fit.seed1.ab$xcoef[, 3] ## changing the sign

[0143] > fit.seed1.ab$xcoef[, 4] <- -fit.seed1.ab$xcoef[, 4] ## changing the sign

[0144] > all(round(fit.seed1.ab$cor, 5)== round(fit.seed1.ba$cor, 5))

[0145] [1] TRUE

[0146] > fit.seed1.ab$proper.u

[0147] [1] 3

[0148] > fit.seed1.ba$proper.u

[0149] [1] 3

[0150] > all(round(fit.seed1.ab$xcoef, 5) == round(fit.seed1.ba$ycoef, 5))

[0151] [1] TRUE

[0152] > fit.seed1.ab.ux <- seedCCA(A, B, type="seed1", ux=2)

[0153] The maximum number of iterations is reached. So, users must choose u bigger than 2.

[0154] > fit.seed1.ab.ux$proper.u

[0155] [1] 2

[0156] fit.seed1.ab의 경우 첫 번째 변수 집합 A는 초기 단계에서 축소된다. 출력 컴포넌트인 initialMX0는 $\mathbf{M}_{x,u}^{0}$의 추정치이고 newX는 $\hat{\mathbf{M}}_{x,u_1}^{0\,\mathrm{T}}\mathbf{X}$입니다. 반대로 fit.seed1.ba의 경우 두 번째 변수 집합 A는 초기에 감소하므로 initialMY0 및 newY가 생성됩니다. 따라서, fit.seed1.ab$xcoef 및 fit.seed1.ba$ycoef는 같다고 말할 필요가 없이, fit.seed1.ab 및 fit.seed1.ba로부터 표준 상관과 u의 제안된 값은 같은 것으로 관측된다. 도 4에 도시된 바와 같이, u를 선택하기 위한 그래프에서 3가지 투영이 제안된다. ux는 seedCCA(A, B, type="seed1", ux=2)에서 충분히 크지 않으므로 아래와 같은 경고가 표시된다.

[0157] "The maximum number of iterations is reached. So, users must choose u bigger than 2."

[0158] 다음으로, 컴퓨팅 장치는 ux, uy, AS 및 eps의 값을 변경한다. type="seed1"에 대한 이들 옵션의 사용은 type="seed2" 및 type="pls"에 대한 옵션의 사용과 동일하다.

[0159] > seedCCA(A, B, type="seed1", ux=5)$proper.u

[0160] [1] 3

[0161] > seedCCA(B, A, type="seed1", eps=0.000001)$proper.u

[0162] [1] 4

[0163] > library(tictoc)

[0164] > tic()

[0165] > seedCCA(B, A, type="seed1", u=30)$proper.u

[0166] > toc()

[0167] 0.03 sec elapsed

[0168] > tic()

[0169] > seedCCA(B, A, type="seed1", u=30, AS=FALSE)$proper.u

[0170] > toc()

[0171] 0.29 sec elapsed

[0172] 옵션 AS의 사용에 주의할 필요가 있다. u 및 AS=FALSE를 선택이 많아질수록, 함수의 실행 시간이 길어진다.

[0173] 도 5는 본 발명의 일실시예에 따른 고유벡터의 집합을 선택하기 위한 그래프를 도시한 도면이다.

[0174] <Seeded CCA (case 2) versus Regularized CCA: nutrimouse data>

[0175] nutrimouse 데이터는 40마리의 생쥐(n = 40)에 대한 영양 연구에서 수집된다. 두 변수 집합 중 하나는 마이크로어레이 기술로 간 세포에서 측정된 120개 유전자의 발현이고, 다른 변수 집합은 가스크로마토그래피로 측정한 21종의 간지방산(FA) 농도이다.

[0176] 또한, 40마리의 마우스는 유전자형과 식단의 두 가지 요인에 따라 교차 분류된다. 야생형(WT) 및 PPARα 결핍(PPARα) 마우스의 2가지 유전자형과 5가지 식단이 존재한다. 5가지 식단은 옥수수 및 콜자 오일(50/50 REF), 포화 FA 식단을 위한 경화 코코넛 오일(COC), 오메가6 지방산이 풍부한 식단을 위한 해바라기유(SUN), 오메가3

가 풍부한 식단을 위한 아마인유(LIN) 및 옥수수/콜자/풍부한 어유(42.5/42.5/15, 물고기)를 포함한다.

[0177] 뉴트리마우스 데이터는 seedCCA 패키지에 포함된다. 뉴트리마우스 데이터에서 n = 40에 비해 min(120, 21)이 상대적으로 크기 때문에, SeedCCA의 사례 2가 사용될 필요가 있다. 그런 다음, SeedCCA의 사례 2는 얼마나 많은 $\tilde{\Sigma}_{xy}$ 의 고유 벡터가 그것을 대체하기에 충분한 것인지를 선택해야 한다. 이것은 투영의 수에 따른 SeedCCA의 사례 2를 위한 조정 파라미터이다.

[0178] SeedCCA에서 옵션인 cut은 $\tilde{\Sigma}_{xy}$ 의 고유 벡터의 수를 자동으로 선택하는 것을 제어한다. 옵션 cut=α는 고유값의 누적 비율이 α보다 크거나 같은 고유 벡터 세트를 결정합니다. 보수적으로 고유 벡터들의 집합을 선택하기 위해, 옵션 cut은 0.9로 설정될 수 있다. 또한 사용자는 d를 사용하여 고유 벡터의 수를 직접 제공할 수 있다. d가 NULL이 아니라면 옵션 cut은 버려진다. 즉, 옵션 cut은 d가 NULL인 경우에만 설정된다는 것을 의미한다.

[0179] 사용자가 d를 사용하기 위해, covplots 함수가 먼저 실행되어야 한다. Covplot 함수는 누적 백분율을 표시하기 위해 고유값의 수를 설정하는 옵션 mind가 있다. 옵션 mine의 기본값은 NULL이고, min(p, r)이 됩니다.

[0180] covplots 함수는 고유값의 scree 그래프와 함께 함께 총 변동의 60%, 70%, 80% 및 90%를 설명하는 고유값, 누적 백분율 및 고유 벡터 수를 반환한다. SeedCCA와 정규 CCA의 결과가 비교된다. 정규 CCA는 두 개의 파라미터의 적절한 값을 선택하기 위해 필요하므로, tictoc 패키지를 통해 정규 CCA와 SeedCCA를 위한 자동 검색의 실행 시간이 비교될 수 있다. seedCCA의에 대해 옵션 cut의 기본값이 사용된다.

[0181] > library(CCA)

[0182] > library(tictoc)

[0183] ## loading nutrimouse data

[0184] > data(nutrimouse)

[0185] > X <- scale(as.matrix(nutrimouse$gene))

[0186] > Y <- scale(as.matrix(nutrimouse$lipid))

[0187] ## determining the number of the eigenvectors of cov(X,Y) with cut=0.9

[0188] > tic("SdCCA")

[0189] > fit.seed2 <- seedCCA(X, Y)

[0190] > toc()

[0191] SdCCA: 0.13 sec elapsed

[0192] ## finding the optimal values of lambda1 and lambda2 for RCCA

[0193] > tic("Regularized CCA")

[0194] > res.regul <- estim.regul(X, Y, plt=TRUE, grid1=seq(0.0001, 0.2, l=51), grid2=seq(0, 0.2, l=51))

[0195] > toc()

[0196] Regularized CCA 819.58 sec elapsed

[0197] ## scree plot of cov(X, Y)

[0198] > names(covplot(X, Y, mind=10))

[0199] [1] "eigenvalue" "cum.percent" "num.evecs"

[0200] > names(fit.seed2)

[0201] [1] "cor" "xcoef" "ycoef" "proper.ux" "proper.uy" "d" "initialMX0" "initialMY0"

[0202] [9] "newX" "newY" "Xscores" "Yscores"

[0203] > fit.seed2$d

[0204] [1] 3

[0205] type="seed2"는 초기화된 CCA 단계에서 X 및 Y의 차원을 줄이기 때문에 initialMX0, initialMY0, newX, newY 및 d의 출력 컴퓨넌트가 보고된다. covplot(X,Y,mind=10)에서 생성된 그래프가 도 5에 도시된다. 도 5에 의하면, 처음 2개, 3개 및 4개의 고유값은 각각 $\hat{\Sigma}_{xy}$에 대한 전체 변동의 79.6%, 91.8% 및 95.9%를 나타낸다.설명합니다. 90% 보수적 지침을 사용하여, 처음 세 개의 가장 큰 고유 벡터가 $\hat{\Sigma}_{xy}$를 충분히 잘 대체한다고 결정된다.

[0206] 도 6은 본 발명의 일실시예에 따른 SeedCCA로부터 생성된 ux 및 uy에 대한 그래프를 도시한 도면이다.

[0207] 도 6을 참고하면, ux와 uy를 선택하기 위한 그래프가 도시된다. 도 6에서 ux와 uy가 각각 7과 6으로 도시된다. 정규 CCA와 SeedCCA의 파라미터의 선택 시간이 비교될 수 있다. SeedCCA는 ux와 uy를 선택하기 위해 필요한 시간이 표준 CCA보다 작다.

[0208] 도 7은 본 발명의 일실시예에 따른 정규화 CCA와 SeedCCA로부터 canonical variates의 제1 쌍을 도시한 도면이다. 도 8은 본 발명의 일실시예에 따른 정규화 CCA와 SeedCCA로부터 canonical variates의 제2 쌍을 도시한 도면이다.

[0209] 도 7 및 도 8에서 설명된 표준 변량 (canonical variates)은 다음과 같다.

[0210] ## Extracting the first two pairs of canonical variates

[0211] > sx1 <- fit.seed2$Xscores[, 1]

[0212] > sx2 <- fit.seed2$Xscores[, 2]

[0213] > sy1 <- fit.seed2$Yscores[, 1]

[0214] > sy2 <- fit.seed2$Yscores[, 2]

[0215] ## fitting the regularized CCA

[0216] > res.rcc <- rcc(X, Y, 0.168016, 0.004)

[0217] > RCCA.X <- X%*%res.rcc$xcoef

[0218] > RCCA.Y <- Y%*%res.rcc$ycoef

[0219] > rx1 <- RCCA.X[,1]

[0220] > rx2 <- RCCA.X[,2]

[0221] > ry1 <- RCCA.Y[,1]

[0222] > ry2 <- RCCA.Y[,2]

[0223] par(mfrow=c(1,2))

[0224] > with(plot(rx1, ry1, col=c(2,4)[genotype], pch=c(1,2)[genotype],

[0225] + main="1st pair from RCCA", xlab="rx1", ylab="ry1"), data=nutrimouse)

[0226] > with(legend(-1.4, 1.4, legend=levels(genotype), col=c(2,4), pch=c(1,2), cex=1.5),

[0227] + data=nutrimouse)

[0228] > with(plot(-sx1, -sy1, col=c(2,4)[genotype], pch=c(1,2)[genotype],

[0229] + main="1st pair from seedCCA", xlab="sx1", ylab="sy1"), data=nutrimouse)

[0230] > with(legend(-1.5, 1.6, legend=levels(genotype), col=c(2,4), pch=c(1,2), cex=1.5),

[0231] + data=nutrimouse)

[0232] > par(mfrow=c(1,2))

[0233] > with(plot(rx2, ry2, col=c(1:4,6)[diet], pch=c(15,16,17,18,20)[diet], cex=1.5,

[0234] + main="2nd pair from RCCA", xlab="rx2", ylab="ry2"), data=nutrimouse)

[0235] > with(legend(-2.3, 1.9, legend=levels(diet), col=c(1:4,6), pch=c(15:18,20)),

[0236] + data=nutrimouse)

[0237] > with(plot(sx2, sy2, col=c(1:4,6)[diet], pch=c(15,16,17,18,20)[diet], cex=1.5,

[0238] + main="2nd pair from seedCCA", xlab="sx2", ylab="sy2"), data=nutrimouse)

[0239] with(legend(-2.5, 1.9, legend=levels(diet), col=c(1:4,6), pch=c(15:18,20)), + data=nutrimouse)

[0240] 도 7 및 도 8을 참고하면, 두 CCA의 표준 변량의 제1 쌍은 유전자형을 매우 잘 구별하지만, 식단으로 표시된 제2 쌍은 매우 복잡하다. 식단에 대한 제2 쌍에 대한 더 많은 통찰력을 얻기 위해, 분산의 다변량 분석이 적합하다.

[0241] > library(lsmeans)

[0242] > fit2r <- manova(cbind(rx2, ry2)~diet, data=nutrimouse)

[0243] > fit3sd <- manova(cbind(sx2, sy2)~diet, data=nutrimouse)

[0244] > test(contrast(lsmeans(fit2r, "diet"), "pairwise"), side = "=", adjust = "fdr")

[0245] contrast estimate SE df t.ratio p.value

[0246] coc - fish -2.3842686 0.2684019 35 -8.883 <.0001

[0247] coc - lin -2.1749708 0.2684019 35 -8.103 <.0001

[0248] coc - ref -1.4881111 0.2684019 35 -5.544 <.0001

[0249] coc - sun -1.6582635 0.2684019 35 -6.178 <.0001

[0250] fish - lin 0.2092978 0.2684019 35 0.780 0.4897

[0251] fish - ref 0.8961575 0.2684019 35 3.339 0.0040

[0252] fish - sun 0.7260051 0.2684019 35 2.705 0.0175

[0253] lin - ref 0.6868597 0.2684019 35 2.559 0.0214

[0254] lin - sun 0.5167073 0.2684019 35 1.925 0.0780

[0255] ref - sun -0.1701524 0.2684019 35 -0.634 0.5302

[0256] Results are averaged over the levels of: rep.meas

[0257] P value adjustment: fdr method for 10 tests

[0258] > test(contrast (lsmeans(fit3sd, "diet"), "pairwise"), side = "=", adjust = "fdr")

[0259] contrast estimate SE df t.ratio p.value

[0260] contrast estimate SE df t.ratio p.value

[0261] coc - fish -2.14838660 0.3163067 35 -6.792 <.0001

[0262] coc - lin -1.79396325 0.3163067 35 -5.672 <.0001

[0263] coc - ref -1.07697196 0.3163067 35 -3.405 0.0035

[0264] coc - sun -1.03303440 0.3163067 35 -3.266 0.0041

[0265] fish - lin 0.35442334 0.3163067 35 1.121 0.3001

[0266] fish - ref 1.07141463 0.3163067 35 3.387 0.0035

[0267] fish - sun 1.11535219 0.3163067 35 3.526 0.0035

[0268] lin - ref 0.71699129 0.3163067 35 2.267 0.0371

[0269] lin - sun 0.76092885 0.3163067 35 2.406 0.0308

[0270] ref - sun 0.04393756 0.3163067 35 0.139 0.8903

[0271] Results are averaged over the levels of: rep.meas

[0272] P value adjustment: fdr method for 10 tests

[0273] 정규 CCA의 경우, "coc" 식단은 다른 것과 다릅니다. 또한 "fish"는 "sun"과 다릅니다. 그러나 다른 쌍별 비교는 상당히 혼합되어 있다. "fish-lin", "lin-sun" 및 "ref-sun" 사이에는 큰 차이가 없는 것으로 확인되었다.

[0274] 반대로, 합리적인 쌍별 비교 결과는 SeedCCA에서 도출된다. 다른 사람들과 마찬가지로 "coc" 식단은는 다른 사람들과 다르다. 또한, "fish-lin"은 크게 다르지 않으며, "ref-sun"은 유사한 것으로 결론지어진다. 어유에는 ω3가 포함되어 있는 것으로 알려져 있으며 아마인유가 이를 위해 설계되었다.

[0275] 또한 참조 오일 식단은 ω6을 포함하는 것으로 알려진 옥수수와 콜자 오일로 구성된다. 해바라기 기름은 실제로 ω6가 풍부한 식단에 적합하므로 이 결과도 합리적입니다. 이와 관련하여, SeedCCA 결과는 정규 CCA보다 바람직할 것입니다.

[0276] 도 9는 본 발명의 일실시예에 따른 서로 다른 조건에서 SeedCCA로부터 생성된 u에 대한 그래프를 도시한 도면이다.

[0277] <Partial least square application with nutrimouse data>

[0278] 뉴트리마우스 데이터를 사용하여 간 세포에서 측정된 120개 유전자의 발현이 주어지는 경우, 21개 간 지방산 농도에서 첫 번째 것인 "C14.0"의 회귀를 고려할 필요가 있다. 이 경우 부분 최소 제곱이 최선의 선택이다. 그런 다음, seedCCA에서 부분 최소 제곱의 추정기를 얻으려면 seedCCA(X,Y,type="pls")를 구현할 필요가 있다. 그러면, "seedpls" 서브 클래스가 생성된다.

[0279] 부분 최소 제곱에서 중요한 문제는 변수의 첫 번째 집합이 예측 변수일 필요가 있다. 응답 변수는 단변량 또는 다변량일 수 있습니다. 추정된 계수가 u에 주어진 값까지 보고되기 때문에 옵션 u는 합리적으로 작게 설정되는 것이 필요하다. scale=TRUE인 경우 제로 샘플 평균과 샘플 상관 행렬을 가지도록 예측자(predictor)가 표준화된다. 추정된 계수와 부분 최소 제곱에 의한 적합값은 함수 coef(object,u=NULL) 및 fitted(object,u=NULL)을 통해 획득될 수 있다. Coef 및 fitt에서 u의 기본값은 NULL이다. 위의 두 함수 모두에서 u의 사용은 동일합니다.

[0280] 만약 u=k인 경우, k개 투영으로부터 계산된 예측 계수와 적합값이 보고된다. 만약 u=NULL인 경우, 모든 예측 계수와 적합값은 u까지 보고된다. type="pls"의 경우 적절한 투영값을 제안하는 자동 절차가 수행되지 않는다. "seedpls" 서브 클래스의 경우, 함수 plot(object)은 다른 출력 구성 요소와 함께 적절한 투영 값을 제안한다. u의 값에 도달하기 전에 종료 조건이 충족되지 않으면 plot(object)은 u의 값을 증가시키는 주의를 제공한다.

[0281] > data(nutrimouse)

[0282] > Y <- as.matrix(nutrimouse$lipid)

[0283] > X <- as.matrix(nutrimouse$gene)

[0284] > Y1 <- as.matrix(Y[, 1]) ## univariate response

[0285] > Y12 <- as.matrix(Y[, 1:2]) ## multivariate response

[0286] ## fitting partial least square and obtaining the estimated coefficient vector

[0287] > fit.pls1.10 <- seedCCA(X, Y1, u=10, type="pls")

[0288] > fit.pls1.3 <- seedCCA(X, Y1, u=3, type="pls", scale=TRUE)

[0289] > names(fit.pls1.10)

[0290] [1] "coef" "u" "X" "Y" "scale"

[0291] > names(fit.pls1.10$coef)

[0292] [1] u=1" "u=2" "u=3" "u=4" "u=5" "u=6" "u=7" "u=8" "u=9" "u=10"

[0293] > names(fit.pls1.3$coef)

[0294] [1] "u=1" "u=2" "u=3"

[0295] > fit.pls1.3$scale

[0296] [1] TRUE

[0297] > par(mfrow=c(1,2))

[0298] > plot(fit.pls1.10)

[0299] $proper.u

[0300] [1] 6

[0301] $nFu

[0302] [1] 6.344725e+00 2.383108e+00 1.681329e+00 2.669394e+00 1.853061e+00 3.217472e-04 5.296046e-05

[0303] [8] 6.017641e-06 4.895905e-07 3.117371e-08

[0304] $u

[0305] [1] 10

[0306] $eps

[0307] [1] 0.01

[0308] > title("fit.pls1.10")

[0309] > plot(fit.pls1.3)

[0310] Caution: The terminating condition is NOT satisfied. The number of projections should be bigger than 3.

[0311] $proper.u

[0312] [1] 3

[0313] $nFu

[0314] [1] 6.344725 2.383108 1.681329

[0315] $u

[0316] [1] 3

[0317] $eps

[0318] [1] 0.01

[0319] > title("fit.pls1.3")

[0320] > names(fitted(fit.pls1.10))

[0321] [1] "u=1" "u=2" "u=3" "u=4" "u=5" "u=6" "u=7" "u=8" "u=9" "u=10"

[0322] > fitted(fit.pls1.10, u=6)

[0323] 1 2 3 4 5 6 7 8 9 10 11 12 13 14

[0324] 0.137 0.368 0.317 0.346 0.492 1.620 0.722 0.003 0.065 1.212 0.458 0.640 0.272 0.397

[0325] 15 16 17 18 19 20 21 22 23 24 25 26 27 28

[0326] -0.103 0.426 1.448 0.287 1.264 0.517 2.803 0.914 0.043 0.028 0.234 0.598 0.875 0.434

[0327] 29 30 31 32 33 34 35 36 37 38 39 40

[0328] 0.694 0.666 2.958 2.350 0.620 0.958 0.495 2.790 0.701 0.168 0.767 0.535

[0329] > fit.pls.m <- seedCCA(X, Y12, u=5, type="pls")

[0330] > dim(fit.pls.m$coef$'u=1')

[0331] [1] 120 2

[0332] seedCCA(X,Y1,u=10,type="pls") 및 seedCCA(X,Y1,u=3,type="pls",scale=TRUE)에 의한 두 개의 부분 최소제곱에 대한 투영 선택은 도 9에 도시된다. 도 9에 의하면, fit.pls1.10object에 대해서는 적절한 투영값이 6으로 제시되고, fit.pls1.3object에 대해서는 종료 조건을 만족하지 못하므로 주의가 제시된다.

[0333] 한편, 본 발명에 따른 방법은 컴퓨터에서 실행될 수 있는 프로그램으로 작성되어 마그네틱 저장매체, 광학적 판독매체, 디지털 저장매체 등 다양한 기록 매체로도 구현될 수 있다.

[0334] 본 명세서에 설명된 각종 기술들의 구현들은 디지털 전자 회로조직으로, 또는 컴퓨터 하드웨어, 펌웨어, 소프트웨어로, 또는 그들의 조합들로 구현될 수 있다. 구현들은 데이터 처리 장치, 예를 들어 프로그램가능 프로세서, 컴퓨터, 또는 다수의 컴퓨터들의 동작에 의한 처리를 위해, 또는 이 동작을 제어하기 위해, 컴퓨터 프로그램 제품, 즉 정보 캐리어, 예를 들어 기계 판독가능 저장 장치(컴퓨터 판독가능 매체) 또는 전파 신호에서 유형적으로 구체화된 컴퓨터 프로그램으로서 구현될 수 있다. 상술한 컴퓨터 프로그램(들)과 같은 컴퓨터 프로그램은 컴파일된 또는 인터프리트된 언어들을 포함하는 임의의 형태의 프로그래밍 언어로 기록될 수 있고, 독립형 프로그램으로서 또는 모듈, 구성요소, 서브루틴, 또는 컴퓨팅 환경에서의 사용에 적절한 다른 유닛으로서 포함하는 임의의 형태로 전개될 수 있다. 컴퓨터 프로그램은 하나의 사이트에서 하나의 컴퓨터 또는 다수의 컴퓨터들 상에서 처리되도록 또는 다수의 사이트들에 걸쳐 분배되고 통신 네트워크에 의해 상호 연결되도록 전개될 수 있다.

[0335] 컴퓨터 프로그램의 처리에 적절한 프로세서들은 예로서, 범용 및 특수 목적 마이크로프로세서들 둘 다, 및 임의의 종류의 디지털 컴퓨터의 임의의 하나 이상의 프로세서들을 포함한다. 일반적으로, 프로세서는 판독 전용 메모리 또는 랜덤 액세스 메모리 또는 둘 다로부터 명령어들 및 데이터를 수신할 것이다. 컴퓨터의 요소들은 명령어들을 실행하는 적어도 하나의 프로세서 및 명령어들 및 데이터를 저장하는 하나 이상의 메모리 장치들을 포함할 수 있다. 일반적으로, 컴퓨터는 데이터를 저장하는 하나 이상의 대량 저장 장치들, 예를 들어 자기, 자기-광 디스크들, 또는 광 디스크들을 포함할 수 있거나, 이것들로부터 데이터를 수신하거나 이것들에 데이터를 송신하거나 또는 양쪽으로 되도록 결합될 수도 있다. 컴퓨터 프로그램 명령어들 및 데이터를 구체화하는데 적절한 정보 캐리어들은 예로서 반도체 메모리 장치들, 예를 들어, 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM(Compact Disk Read Only Memory), DVD(Digital Video Disk)와 같은 광 기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-Optical Media), 롬(ROM, Read Only Memory), 램(RAM, Random Access Memory), 플래시 메모리, EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM) 등을 포함한다. 프로세서 및 메모리는 특수 목적 논리 회로조직에 의해 보충되거나, 이에 포함될 수 있다.

[0336] 또한, 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용매체일 수 있고, 컴퓨터 저장매체 및 전송매체를 모두 포함할 수 있다.

[0337] 본 명세서는 다수의 특정한 구현물의 세부사항들을 포함하지만, 이들은 어떠한 발명이나 청구 가능한 것의 범위에 대해서도 제한적인 것으로서 이해되어서는 안되며, 오히려 특정한 발명의 특정한 실시형태에 특유할 수 있는 특징들에 대한 설명으로서 이해되어야 한다. 개별적인 실시형태의 문맥에서 본 명세서에 기술된 특정한 특징들은 단일 실시형태에서 조합하여 구현될 수도 있다. 반대로, 단일 실시형태의 문맥에서 기술한 다양한 특징들 역시 개별적으로 혹은 어떠한 적절한 하위 조합으로도 복수의 실시형태에서 구현 가능하다. 나아가, 특징들이 특정한 조합으로 동작하고 초기에 그와 같이 청구된 바와 같이 묘사될 수 있지만, 청구된 조합으로부터의 하나 이상의 특징들은 일부 경우에 그 조합으로부터 배제될 수 있으며, 그 청구된 조합은 하위 조합이나 하위 조합의 변형물로 변경될 수 있다.

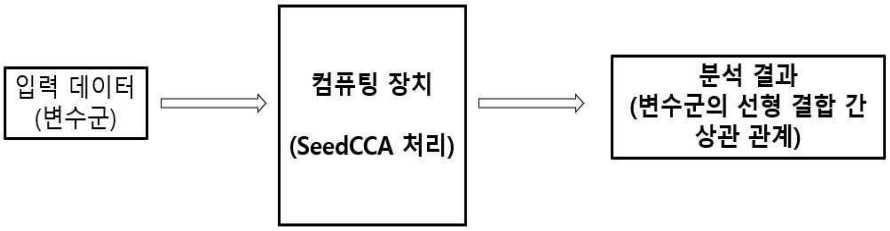
[0338] 마찬가지로, 특정한 순서로 도면에서 동작들을 묘사하고 있지만, 이는 바람직한 결과를 얻기 위하여 도시된 그

특정한 순서나 순차적인 순서대로 그러한 동작들을 수행하여야 한다거나 모든 도시된 동작들이 수행되어야 하는 것으로 이해되어서는 안 된다. 특정한 경우, 멀티태스킹과 병렬 프로세싱이 유리할 수 있다. 또한, 상술한 실시형태의 다양한 장치 컴포넌트의 분리는 그러한 분리를 모든 실시형태에서 요구하는 것으로 이해되어서는 안되며, 설명한 프로그램 컴포넌트와 장치들은 일반적으로 단일의 소프트웨어 제품으로 함께 통합되거나 다중 소프트웨어 제품에 패키징 될 수 있다는 점을 이해하여야 한다.

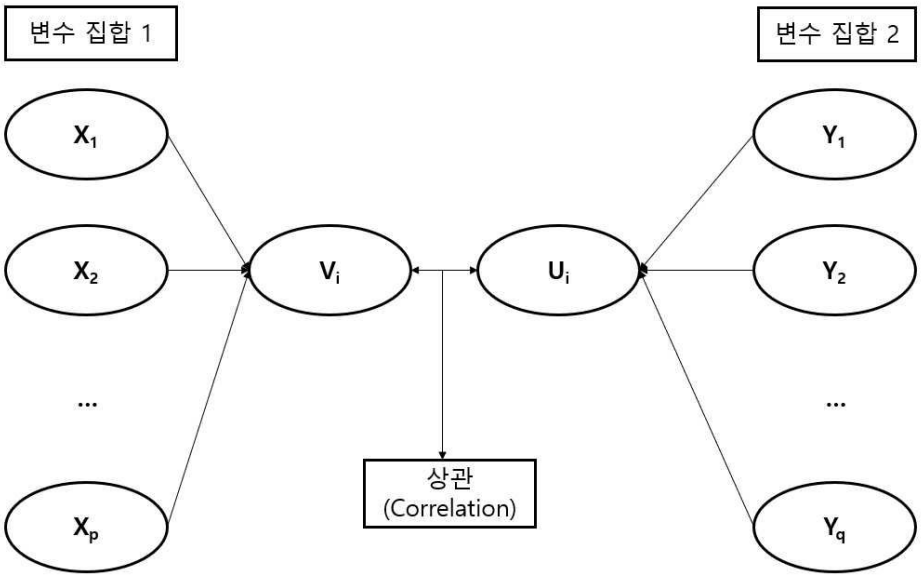
[0339] 한편, 본 명세서와 도면에 개시된 본 발명의 실시 예들은 이해를 돕기 위해 특정 예를 제시한 것에 지나지 않으며, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 개시된 실시 예들 이외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 자명한 것이다.

도면

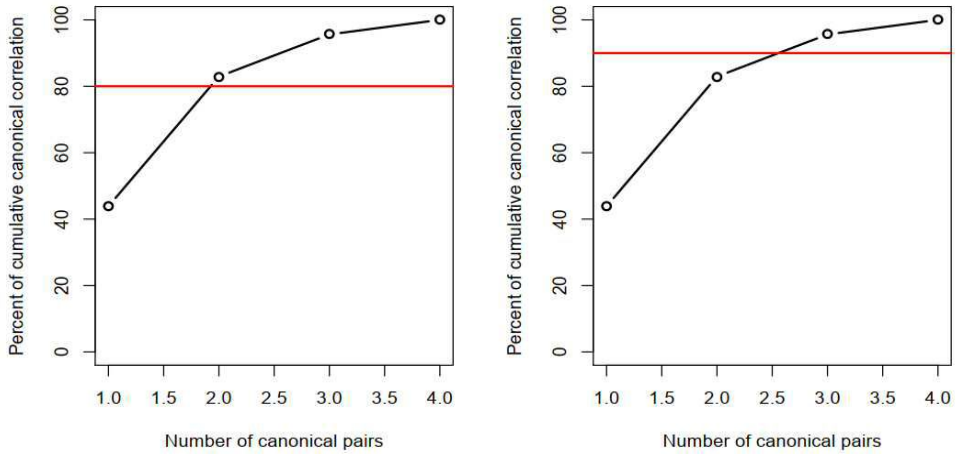
도면1



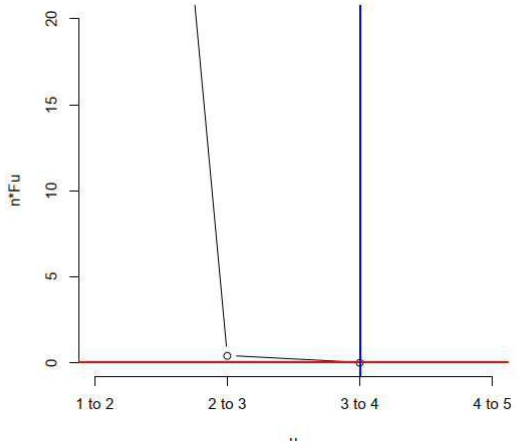
도면2



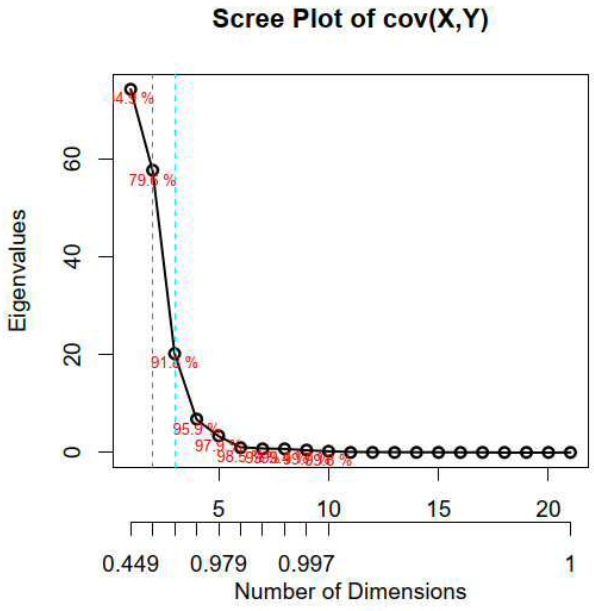
도면3


Percent of cumulative canonical correlation
0
20
40
60
80
100
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Number of canonical pairs
Percent of cumulative canonical correlation
0
20
40
60
80
100
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Number of canonical pairs

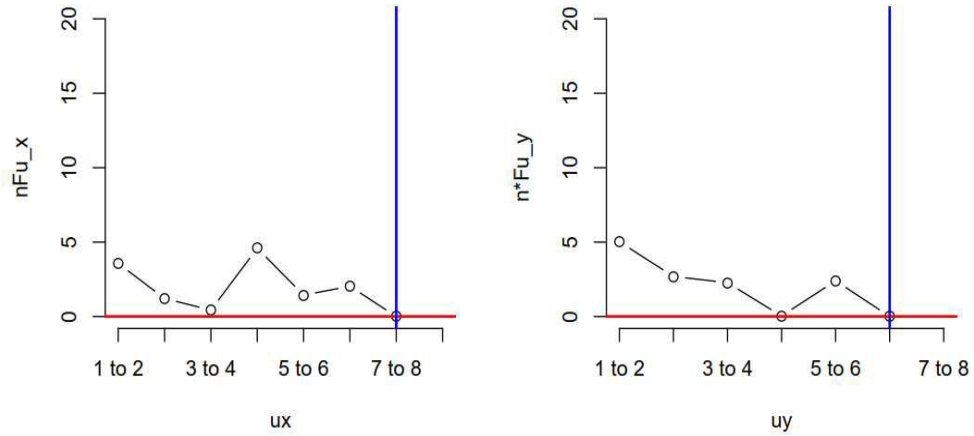
도면4


n*Fu
0
5
10
15
20
1 to 2
2 to 3
3 to 4
4 to 5

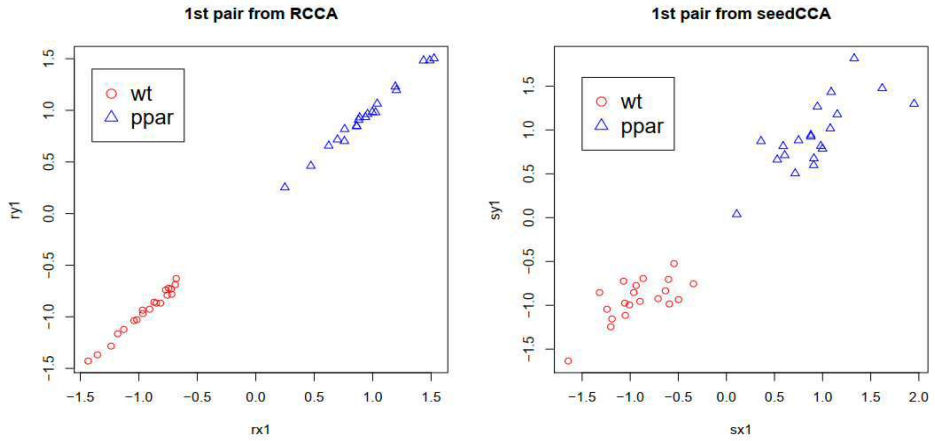
도면5


Scree Plot of cov(X,Y)
Eigenvalues
0
20
40
60
79.1 %
91.8 %
5
10
15
20
0.449
0.979
0.997
1
Number of Dimensions

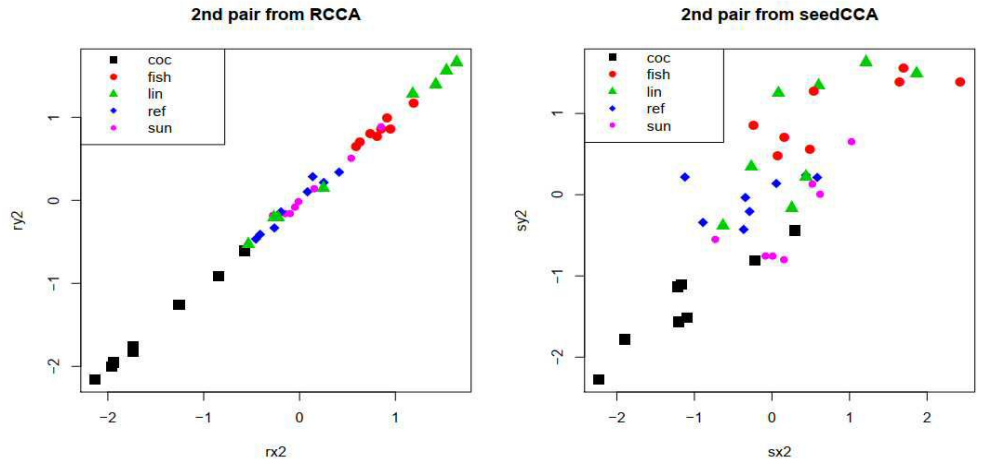
도면6


nFu_x
0
5
10
15
20
1 to 2
3 to 4
5 to 6
7 to 8
ux
n*Fu_y
0
5
10
15
20
1 to 2
3 to 4
5 to 6
7 to 8
uy

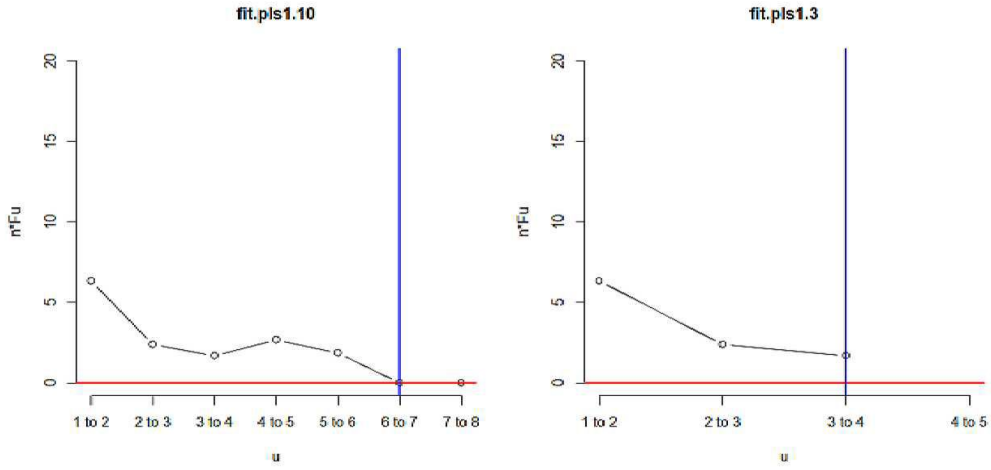
도면7


1st pair from RCCA
wt
ppar
ry1
rx1
1st pair from seedCCA
wt
ppar
sy1
sx1

도면8


2nd pair from RCCA
coc
fish
lin
ref
sun
ry2
rx2
2nd pair from seedCCA
coc
fish
lin
ref
sun
sy2
sx2

도면9


fit.pls1.10
n*Fu
0
5
10
15
20
1 to 2
2 to 3
3 to 4
4 to 5
5 to 6
6 to 7
7 to 8
u
fit.pls1.3
n*Fu
0
5
10
15
20
1 to 2
2 to 3
3 to 4
4 to 5
u