The Korean Journal of Applied Statistics
2024, Vol. 37, No. 2, 157–176

DOI: https://doi.org/10.5351/KJAS.2024.37.2.157
Print ISSN1225-066X / Online ISSN 2383-5818

A comparison study of canonical methods: Application to -Omics data

Seungsoo Lee[a], Eun Jeong Min[1, a,b]

[a]Department of Biomedicine & Health Sciences, The Catholic University of Korea;
[b]Department of Medical Life Sciences, College of Medicine, The Catholic University of Korea

Abstract

Integrative analysis for better understanding of complex biological systems gains more attention. Observing subjects from various perspectives and conducting integrative analysis of those multiple datasets enables a deeper understanding of the subject. In this paper, we compared two methods that simultaneously consider two datasets gathered from the same objects, canonical correlation analysis (CCA) and co-inertia analysis (CIA). Since CCA cannot handle the case when the data exhibit high-dimensionality, two strategies were considered instead: Utilization of a ridge constant (CCA-ridge) and substitution of covariance matrices of each data to identity matrix and then applying penalized singular value decomposition (CCA-PMD). To illustrate CIA and CCA, both extensions of CCA and CIA were applied to NCI60 cell line data. It is shown that both methods yield biologically meaningful and significant results by identifying important genes that enhance our comprehension of the data. Their results shows some dissimilarities arisen from the different criteria used to measure the relationship between two sets of data in each method. Additionally, CIA exhibits variations dependent on the weight matrices employed.

Keywords: CCA, CIA, integrative analysis, omics, NCI60

1. 서론

기술의 발달로 인해 우리는 연구 대상을 다양한 시각에서 관찰하여 다양한 데이터들을 생산해낼 수 있게 되었다. 이는 연구 대상에 대한 좀 더 깊은 이해를 끌어내는 바탕이 되었는데, 의생명과학분야가 그 대표적인 예라 할 수 있다. 현재 우리는 한 명의 환자에게서 임상데이터뿐만 아니라 각종 검사로 얻는 이미지 데이터 및 임상검사 측정값 그리고 여러 종류의 유전체 데이터 등 다양한 관점에서 환자 상태를 관측한 데이터들을 얻을 수 있다. 따라서 기존의 각각의 관측데이터를 분석하여 환자의 건강 상태에 대해 이해하던 수준에서 벗어나, 다각도로 관측한 여러 데이터를 동시에 고려함으로써 환자의 몸에서 일어나고 있는 상황에 대한 복합적인 이해를 위한 분석방법론에 대한 필요 또한 점점 높아지고 있다.

이러한 요구에 맞추어 둘 혹은 그 이상의 데이터를 동시에 고려할 수 있는 다양한 통계적 방법론들이 제시되어 왔는데, 그중에서도 정준상관분석(canonical correlation analysis; CCA) (Hotelling, 1936)은 두 개의 데이터를 함께 고려하여 분석하는 모형 중 가장 대표적인 방법이다. 정준상관분석은 피어슨 상관계수를 통해 두 데이터 간의 관련성을 확인하는 방법으로 데이터에 계수(coefficient)를 곱한 값을 정준변량(canonical

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. NRF-2021R1F1A1058613).
[1]Corresponding author: Department of Medical Life Sciences, College of Medicine, The Catholic University of Korea, Banpo-daero 222, Seocho-gu, Seoul 06591, Korea. E-mail: ej.min@catholic.ac.kr

Published 30 April 2024 / journal homepage: http://kjas.or.kr

variate)이라 할 때 두 개의 정준변량 사이의 피어슨 상관계수가 최대가 되도록 하는 최적의 계수를 추정하는 방법이다. 분석 결과로 얻은 피어슨 상관계수의 크기를 통해 두 정준변량간의 관련도를 파악하며 추정된 계수를 통해 연관성에 대한 각 변수의 설명력을 가늠한다. 정준상관분석은 공분산행렬의 역행렬로 만들어진 행렬을 특이값 분해하여 그 분석 결과를 도출하는데 만일 각 데이터 표본의 개수보다 변수의 개수가 더 많은 경우 공분산행렬이 비정칙행렬(singular matrix)이되어 분석수행에 어려움이 있다. 이러한 문제점을 해결하기 위한 선행연구를 여럿 찾아볼 수 있는데 대표적인 방법으로 Vinod (1976)가 제시한 방법(CCA-ridge)이 있다. CCA-ridge는 기존의 공분산 행렬에 능형상수(Ridge constant)를 추가하여 역행렬 계산이 가능하도록 변환하는 방법이다. 또한 Witten 등 (2009)은 벌점화 행렬분해를 이용한 방법(CCA-penalised matrix decomposition; CCA-PMD)을 제안하였는데, 분석 과정에서 역행렬이 존재하지 않는 각 데이터의 공분산 행렬을 항등행렬로 치환하여 분석을 수행하도록 하였다.

공관성분석(Co-inertia analysis; CIA) (Dolédec와 Chessel, 1994)은 두 개의 데이터를 함께 고려하는 또다른 융합 분석 모형 중 하나이다. 공관성분석은 생태학 분야에서 처음으로 제시된 방법으로, 장소별 생물의 종의 개수와 환경적 요인의 관련성을 알아보기 위해 사용되었다 (Dolédec와 Chessel, 1994; Dray 등, 2003). 공관성분석은 정준상관분석과 달리 두 선형변환된 데이터 간의 관련성을 공관성 계수(Co-inertia; COI)로 측정하고 이를 최대화하는 최적의 선형조합을 찾는 방법으로, 여기에서 사용되는 공관성 계수는 두 데이터 간의 공분산행렬에 세 개의 가중치 행렬들을 곱한 후 선형변환한 두 변량간의 연관성을 나타내는 측도이다. 공관성 계수는 일종의 가중공분산으로도 이해할 수 있다. 공관성분석의 경우 데이터의 표본수가 변수의 수보다 작은 경우에도 문제없이 적용할 수 있는 장점이 있어, 최근들어 유전체 데이터 분석에 적용되기 시작하였다 (Culhane 등, 2003; Cao 등, 2009; Min 등, 2019).

정준상관분석과 공관성분석은 같은 표본으로부터 얻은 두 다변량 그룹을 선형변환한 조합의 관련성을 다룬다는 점에서 비슷한 분석방법이지만 두 변수간의 관련성을 정의하는데 쓰이는 측도가 다르기 때문에 두 방법간의 차이에 대한 이해를 바탕에 두고 처한 연구 상황에 따른 적절한 분석 방법의 선택이 중요하다고 할 수 있다. 본 연구에서는 정준상관분석(CCA-ridge, CCA-PMD)과 공관성분석을 NCI60 세포주 데이터에 적용하여 분석결과를 비교하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 다양한 정준상관분석 방법과 공관성분석을 소개한다. 3장에서는 NCI60 세포주 데이터 데이터를 소개하고 NCI60 세포주 데이터에 각 방법을 적용한 분석 결과를 비교한다. 마지막으로 4장에서 본 연구의 내용을 요약하고 결론을 제시한다.

2. 분석모형

2.1. 정준상관분석(CCA)

정준상관분석은 각 데이터의 선형조합으로 생성된 두 변량간의 피어슨 상관계수가 최대가 되도록 만드는 최적의 계수를 찾는 방법이다 (Hotelling, 1936). p개의 변수에 대해 관측된 데이터를 중심화한 자료를 $\boldsymbol{X}$, q개의 변수에 대해 관측된 데이터를 중심화한 자료를 $\boldsymbol{Y}$라고 할 때 정준상관분석의 목적함수는 다음과 같다.

$$\text{argmax}_{\boldsymbol{u},\boldsymbol{v}}\ \boldsymbol{u}^{\top}\boldsymbol{C}_{xy}\boldsymbol{v} \quad \text{s.t} \quad \boldsymbol{u}^{\top}\boldsymbol{C}_{xx}\boldsymbol{u} = \boldsymbol{v}^{\top}\boldsymbol{C}_{yy}\boldsymbol{v} = 1. \tag{2.1}$$

$\boldsymbol{C}_{xx}$, $\boldsymbol{C}_{yy}$, $\boldsymbol{C}_{xy}$는 각각 $\boldsymbol{X}$의 공분산 행렬, $\boldsymbol{Y}$의 공분산행렬, 그리고 $\boldsymbol{X},\boldsymbol{Y}$간의 공분산행렬을 나타낸다. 정준상관분석의 목적함수를 최대화하는 최적의 $\boldsymbol{u},\boldsymbol{v}$는 $\boldsymbol{C}_{xx}^{-1/2}\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1/2}$의 특이값분해를 통해 도출할 수 있다. 정준상관분석의 가장 큰 단점은 고차원데이터에 적용할 수 없다는 점으로, 이는 표본의 개수보다 변수의 개수가 많은 경우 두 역행렬 $\boldsymbol{C}_{xx}^{-1}$ 와 $\boldsymbol{C}_{yy}^{-1}$이 고유한 값을 가지지 못한다는 점에서 기인한다. 이 문제를 해결하여 고차원 데이터에도 정준상관분석을 이용할 수 있도록 하는 여러 방안이 제안된 바 있으며 (Wilms와 Croux, 2016), 그 중 두 가지 방법이 다음에 소개되어 있다.

2.2. 능형상수를 이용한 정준상관분석(CCA-ridge)

정준상관분석을 고차원 데이터에 적용 가능하도록 하는 방법 중 하나는 공분산행렬에 능형상수를 추가하여 역행렬의 계산이 가능하도록 만드는 방법(CCA-ridge)이다 (Vinod, 1976). CCA-ridge의 목적함수는 다음과 같다.

$$\text{argmax}_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^\top \boldsymbol{C}_{xy} \boldsymbol{v} \quad \text{s.t} \quad \boldsymbol{u}^\top (\boldsymbol{C}_{xx} + \boldsymbol{c}_1 \boldsymbol{I}) \boldsymbol{u} = \boldsymbol{v}^\top \left(\boldsymbol{C}_{yy} + \boldsymbol{c}_2 \boldsymbol{I}\right) \boldsymbol{v} = 1. \tag{2.2}$$

이때 $\boldsymbol{c}_1$과 $\boldsymbol{c}_2$는 임의의 작은 양수로 능형상수이다. 정준상관분석의 목적함수 (2.1)와 비교하였을때 기존의 공분산행렬 $\boldsymbol{C}_{xx}, \boldsymbol{C}_{yy}$에 데이터에 담긴 정보를 크게 왜곡시키지 않을 작은 오차를 더하여 역행렬의 계산이 가능하도록 했음을 알 수 있다. 위 목적함수 (2.2)의 최적해는 변수들간의 공선성(collinearity)이 있는 경우에도 안정적으로 추정된다는 장점이 있다 (Vinod, 1976).

2.3. 대각벌점화 정준상관분석(CCA-PMD)

정준상관분석의 고차원데이터에의 적용을 위한 또다른 연구로는 2009년에 Witten 등이 발표한 방법(CCA-PMD)이 있다. 해당 논문에서는 벌점화 특이값분해 모형(PMD)이 제안되고 그 방법을 정준상관분석에 적용한 결과를 제시하였는데, 공분산행렬을 대각행렬로 치환하여도 좋은 결과를 얻었다는 선행연구에 기반하여 (Dudoit 등, 2001; Tibshirani 등, 2003) 분석과정에서 공분산행렬을 항등행렬로 설정하여 역행렬을 사용하지 않는 방법을 제시했다. CCA-PMD의 목적함수는 다음과 같다.

$$\text{argmax}_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{u}^\top \boldsymbol{C}_{xy} \boldsymbol{v} \quad \text{s.t} \quad \|\boldsymbol{u}\|_2 \le 1,\ \|\boldsymbol{v}\|_2 \le 1,\ P_1(\boldsymbol{u}) \le \boldsymbol{c}_1,\ P_2(\boldsymbol{v}) \le \boldsymbol{c}_2. \tag{2.3}$$

위의 식 (2.3)에서 $P_1(\boldsymbol{u})$와 $P_2(\boldsymbol{v})$는 각 $\boldsymbol{u}$와 $\boldsymbol{v}$에 적용하는 벌점함수로 least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996)와 Fused LASSO (Tibshirani 등, 2005)등이 해당 연구에서 고려되었다. CCA-PMD방법은 고차원데이터를 정준상관분석을 적용할 수 있도록 CCA를 확장하였을 뿐 아니라 변수 선택이 모형적합과 동시에 진행되기 때문에 그 결과의 해석이 매우 용이하다는 장점이 있다. 해당 모형의 알고리즘은 R 패키지 `PMA`로 구현되어 있다.

2.4. 공관성분석(CIA)

두 데이터의 관련성을 측정하는 지표로 피어슨 상관계수 이외에도 다른 지표를 사용할 수 있다. 공관성 계수(COI)의 값은 두 데이터의 공분산과 가중치를 통해 두 데이터의 관련성을 나타내는 지표로, 그 값을 계산하는 식은 다음과 같다.

$$\text{COI} = \sum_{i=1}^{p} \sum_{j=1}^{q} \left(\boldsymbol{u}_i^\top \boldsymbol{Q}_x \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{Y} \boldsymbol{Q}_y \boldsymbol{v}_j\right)^2.$$

이때 $\boldsymbol{Q}_x, \boldsymbol{Q}_y, \boldsymbol{D}$ 는 각각 $\boldsymbol{X}$ 변수의 가중치 행렬, $\boldsymbol{Y}$ 변수의 가중치 행렬, 표본의 가중치 행렬로, 모든 가중치 행렬은 정칙행렬이다. 공관성분석은 이러한 공관성 계수가 최대가 되도록 만드는 선형결합을 찾아 그 결합된 데이터간의 관련성을 살펴보는 분석방법으로 Dolédec와 Chessel (1994)이 처음 제시하였다. 공관성분석의 목적함수는 다음과 같으며,

$$\text{argmax}_{\boldsymbol{u},\boldsymbol{v}} \left(\boldsymbol{u}^\top \boldsymbol{Q}_x \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{Y} \boldsymbol{Q}_y \boldsymbol{v}\right)^2 \quad \text{s.t} \quad \boldsymbol{u}^\top \boldsymbol{Q}_x \boldsymbol{u} = \boldsymbol{v}^\top \boldsymbol{Q}_y \boldsymbol{v} = 1. \tag{2.4}$$

위 식 (2.4)를 만족시키는 $\boldsymbol{u}$와 $\boldsymbol{v}$는 $\boldsymbol{Q}_x^{1/2} \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{Y} \boldsymbol{Q}_y \boldsymbol{Y}^\top \boldsymbol{D} \boldsymbol{X} \boldsymbol{Q}_x^{1/2}$의 최대 고유값을 도출해내는 고유벡터 쌍과 동일함임이 같은 논문에서 증명되었다.

Table 1: Weight matrices considered in the NCI60 data analysis

Case	$\boldsymbol{Q}_x$	$\boldsymbol{Q}_y$	$\boldsymbol{D}$
Case1	$\boldsymbol{I}_p$	$\boldsymbol{I}_q$	$\boldsymbol{I}_d$
Case2	$\boldsymbol{Q}_x^{(1)} = \text{diag}(\frac{\boldsymbol{x}_i^*}{\sum_{i=1}^p \boldsymbol{x}_i^*})$	$\boldsymbol{Q}_y^{(1)} = \text{diag}(\frac{\boldsymbol{y}_j^*}{\sum_{j=1}^q \boldsymbol{y}_j^*})$	$\boldsymbol{I}_d$
Case3	$\boldsymbol{Q}_x^{(1)}$	$\boldsymbol{Q}_y^{(1)}$	$\boldsymbol{D}^{(1)} = \text{diag}(\frac{\boldsymbol{r}_k^*}{\sum_{k=1}^n \boldsymbol{r}_k^*})$
Case4	$\boldsymbol{Q}_x^{(2)} = \text{diag}(\frac{\sum_{i=1}^p \boldsymbol{x}_i^*}{\boldsymbol{x}_i^*})$	$\boldsymbol{Q}_y^{(2)} = \text{diag}(\frac{\sum_{j=1}^q \boldsymbol{y}_j{}^*}{\boldsymbol{y}_j^*})$	$\boldsymbol{I}_d$
Case5	$\boldsymbol{Q}_x^{(2)}$	$\boldsymbol{Q}_y^{(2)}$	$\boldsymbol{D}^{(2)} = \text{diag}(\frac{\sum_{k=1}^n \boldsymbol{r}_k^*}{\boldsymbol{r}_k^*})$
Case6	$\boldsymbol{Q}_x^{(1)}$	$\boldsymbol{Q}_y^{(1)}$	$\boldsymbol{D}^{(2)}$
Case7	$\boldsymbol{Q}_x^{(2)}$	$\boldsymbol{Q}_y^{(2)}$	$\boldsymbol{D}^{(1)}$

Asterisk means that corresponding data is transformed data via `isDataFrame` function of the R package `MADE4`. r_k stands for the k^{th} row of the concatenated matrix of X and Y, and $\boldsymbol{I}_l$ is a $l \times l$ identity matrix.

공관성분석의 가장 큰 특징은 가중치 행렬을 연구자의 목적에 따라 다양하게 설정할 수 있다는 점이다. 처음 공관성분석이 제시되고 사용된 생태학 분야에서는 그 연구목적에 따라 다양한 가중치행렬이 제시된 바 있다(Table 1) (Dray 등, 2003). 다양한 형태의 정칙행렬이 가중치행렬로 사용될 수 있으나, 실제 해석의 용이함을 위해서는 대각행렬을 사용하는 것이 좋다. 가중치 행렬을 설정하는 방법 중 하나의 예로, 전체 변수 중 사전에 특정 변수가 중요하다는 사실이 알려져 있다면 연구자는 가중치 행렬에서 해당하는 변수의 가중치를 크게 설정하고 그 외의 대각원소를 작게 설정하여 사전정보를 이용하는 방식이 있을 수 있다. 또는 표본의 편향을 염려되는 경우, 그 영향을 줄이기 위해 각 표본 확률의 역확률을 대각원소로 하는 가중치행렬을 고려할 수 있다 (Min 등, 2019). 공관성분석은 R 패키지 `MADE4`에 그 방법이 구현되어 있다.

3. 데이터 분석

3.1. 데이터 및 분석과정

앞서 소개한 공관성분석과 두 가지의 확장버전의 정준상관분석을 비교하기 위해 NCI60 세포주 데이터(https://discover.nci.nih.gov/cellminer)를 사용하였다. NCI60 세포주 샘플은 미국 국립 암 연구소(national cancer institute; NCI)에서 제공받을 수 있으며 9가지 종류 암세포(유방암, 중추신경계암, 대장암, 백혈병, 폐암, 흑색종암, 난소암, 전립선암, 신장암)에서 얻은 60개의 세포주로 구성되어 있다. CellMiner라는 웹페이지에서 여러 연구자들이 NCI60 세포주 표본 분석을 통해 얻은 유전자 발현 데이터(gene expression data), 약물 데이터(drug data), 메틸화 데이터(methylation data) 등의 다양한 데이터가 공유되어 누구든지 내려받을 수 있다. 본 논문은 그 중에서 RNA 시퀀싱 데이터 (Liu 등, 2010)와 단백질 시퀀싱 데이터 (Nishizuka 등, 2003)를 이용하였다. RNA 시퀀싱 데이터는 마이크로어레이(Microarray) 방법을 통해 mRNA 양을 측정한 값으로 $\log_2$ 변환 후 데이터의 순위를 매겨 전체 데이터에서 75%에 해당하는 값을 뺀 값으로 사용했다 (Liu 등, 2010). 단백질 시퀀싱 데이터는 역상 단백질 리세이트 마이크로어레이(reverse phase protein microarrays)를 통해 측정한 단백질 양으로 $\log_2$ 변환 후 평균을 뺀 값을 사용했다 (Nishizuka 등, 2003). 결과 해석을 위해 유전자 이름이 확인된 probe 데이터만을 골라 사용하였고 계산 부담을 줄이기 위해 RNA 시퀀싱 데이터에서 관측값의 분산이 상위 5%에 해당하는 유전자를 선택하여 사용했다. 최종적으로 1,582개의 유전자에 대한 관측값을 가진 RNA 시퀀싱 데이터와 94개의 관측치를 가진 단백질 시퀀싱 데이터가 분석에 이용되었다.

먼저, 분석을 진행하기에 앞서 유전자별 동등한 비교를 위해 데이터별로 표준화(standardization)를 시행하여 사용하였다. 정준상관분석의 두가지 방법 중 CCA-ridge에서는 능형상수가 이용되는데, 본 논문에서 사용하는 능형상수 $\boldsymbol{c}_1$과 $\boldsymbol{c}_2$는 $\boldsymbol{N}(0, 0.001)$ 에서 2개의 값을 생성하여 그 절댓값으로 설정하였다. CCA-PMD

방법의 경우, (2.3)에서 P_1과 P_2로 LASSO 벌점함수를 사용하였으며, R 패키지 `PMA`를 사용하여 분석하였다. 공관성분석의 경우, 분석에 가중치행렬이 입력값으로 미리 정의되어 공급되어야 한다. 본 연구에서는 공관성분석에 대한 이전 연구 (Meng 등, 2014)의 방식을 따라서 R 패키지 `MADE4`의 `isDataFrame` 함수를 통해 양수로 변환한 데이터를 이용하여 가중치 행렬을 설정하였는데, 다양한 방식으로 총 일곱 가지의 가중치 행렬을 고려하였다. 첫 번째는 가중치를 사용하지 않는 상황으로 모든 가중치 행렬이 항등행렬인 경우이다. 나머지 여섯 가지의 가중치 행렬 조합은 각 데이터 내의 유전자별 관측치의 비율을 이용하여 변수 공간에 대한 가중치를 구성하고 표본 내 각 암종 구성 비율이나 항등행렬을 표본공간에 대한 가중치로 이용하여 구성한 경우들의 조합으로 이루어져 있다. 변수 공간의 가중치의 경우 유전자의 발현량이 많을수록 더 높은 가중치를 고려할 경우 $\boldsymbol{Q}_x^{(1)}$와 $\boldsymbol{Q}_y^{(1)}$를 고려하였고 반대로 더 낮은 가중치를 고려할 경우 $\boldsymbol{Q}_x^{(2)}$와 $\boldsymbol{Q}_y^{(2)}$가 이용되었으며, 표본공간에 대한 가중치의 경우 많은 표본을 가진 암종에 더 높은 가중치를 고려한다면 $\boldsymbol{D}^{(1)}$을 이용하고 그 반대의 가중치를 고려한다면 $\boldsymbol{D}^{(2)}$를 이용하였다. 모든 가중치 행렬은 해석의 편의를 위해 대각행렬로 정의되었다. 위에서 설명한 가중치 행렬에 대한 정확한 계산식 및 조합에 대한 내용은 Table 1에 정리되어있다.

3.2. 평가지표

정준상관분석과 공관성분석은 두 다변량 데이터간의 상관관계에 대해 분석하는 방법으로 회귀모형과 달리 예측오차를 계산할 수 없어 직접적인 결과 비교가 어렵다는 점이 선행연구에서 논의 된 바 있다 (Cao 등, 2009). 따라서 본 연구는 이전 논의에서 고려된 다양한 방식들을 따라서 여러 가지의 측도값과 그래프를 통해 분석 결과를 비교하였다.

먼저 정준상관분석과 공관성분석을 비교하기 위해 피어슨 상관계수, redundancy criterion (Rd) (Tenenhaus, 1998; Cao 등, 2009) 측도값을 사용하였다. Rd값은 데이터와 추정된 선형조합변량간의 관계를 설명하는 측도로 고려하는 조합에 따라 네가지의 값을 생각해 볼 수 있으며 사용하고자하는 추정 선형조합변량의 개수에 따라 값이 달라진다. p개의 변수를 가진 $\boldsymbol{X}$와 q개의 변수를 가진 $\boldsymbol{Y}$가 있을 때 H개의 추정 선형조합변량을 사용한 Rd값은 다음과 같다.

$$\mathrm{Rd}\left(\boldsymbol{X};\boldsymbol{Xu}_1,\ldots,\boldsymbol{Xu}_H\right)=\frac{1}{p}\sum_{h=1}^{H}\sum_{i=1}^{p}\mathrm{cor}^2\left(\boldsymbol{x}^i,\boldsymbol{Xu}_h\right),\quad \mathrm{Rd}\left(\boldsymbol{Y};\boldsymbol{Yv}_1,\ldots,\boldsymbol{Yv}_H\right)=\frac{1}{q}\sum_{h=1}^{H}\sum_{j=1}^{q}\mathrm{cor}^2\left(\boldsymbol{y}^j,\boldsymbol{Yv}_h\right). \tag{3.1}$$

$$\mathrm{Rd}\left(\boldsymbol{X};\boldsymbol{Yv}_1,\ldots,\boldsymbol{Yv}_H\right)=\frac{1}{p}\sum_{h=1}^{H}\sum_{i=1}^{p}\mathrm{cor}^2\left(\boldsymbol{x}^i,\boldsymbol{Yv}_h\right),\quad \mathrm{Rd}\left(\boldsymbol{Y};\boldsymbol{Xu}_1,\ldots,\boldsymbol{Xu}_H\right)=\frac{1}{q}\sum_{h=1}^{H}\sum_{j=1}^{q}\mathrm{cor}^2\left(\boldsymbol{y}^j,\boldsymbol{Xu}_h\right). \tag{3.2}$$

식(3.1)에서의 두개의 Rd값은 추정 선형결합 변량과 해당 변량 조합시 이용된 데이터간의 관련도를 나타내며 식(3.2)에서의 두개의 Rd값은 추정한 변량과 해당 변량의 조합에 사용되지 않은 다른 데이터간의 관련도를 나타낸다.

또한 공관성분석의 경우, 각기 다른 가중치행렬 사용에 따른 비교를 위해 RV-coefficient (Culhane 등, 2003), cumulated percentage of variability explained by estimated loadings (CumVar) (Culhane 등, 2003; Min 등, 2019)를 사용하였다. RV-coefficient값은 피어슨 상관계수에서 사용되는 상관계수를 공관성계수로 치환하여 확장한 형태로, 그 계산식은 다음과 같다.

$$\text{RV-coefficient}=\frac{\mathrm{COI}(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{\mathrm{COI}(\boldsymbol{X},\boldsymbol{X})}\sqrt{\mathrm{COI}(\boldsymbol{Y},\boldsymbol{Y}).}}$$

CumVar는 공관성분석의 목적함수를 최적화하여 얻는 고유값들의 총합 대비 가장 큰 첫 h개의 고유값의 합이 차지하는 비율로, 각 방법에서 추정한 선형조합에 의해 계산된 두 변량이 두 데이터간의 관계를 얼마나 설명할 수 있는지 확인할 수 있는 지표이다. 본 연구에서는 $h = 2$를 사용하였다.

Table 2: RV-coefficient, CumVar, Pearson correlation coefficients

Method	Cor($\tilde{X}a_1, \tilde{Y}b_1$)	Cor($\tilde{X}a_2, \tilde{Y}b_2$)	Cor($\tilde{X}a_3, \tilde{Y}b_3$)	RV-coefficient	CumVar
CIA-case1	0.9058	0.9284	0.8742	0.6370	0.5171
CIA-case2	0.9021	0.9336	0.8899	0.6469	0.5293
CIA-case3	0.9021	0.9331	0.8878	0.6461	0.5253
CIA-case4	0.8901	0.9066	0.8330	0.5688	0.4995
CIA-case5	0.8922	0.9086	0.8345	0.5699	0.5026
CIA-case6	0.9021	0.9342	0.8918	**0.6478**	**0.5331**
CIA-case7	0.8878	0.9046	0.8314	0.5679	0.4963
CCA-ridge	**1.0000**	**1.0000**	**1.0000**	-	-
CCA-PMD	0.9357	0.9139	0.9372	-	-

Table 3: Rd of associated data

Method	Rd($\tilde{X}; \tilde{X}a_1$)	Rd($\tilde{X}; \tilde{X}a_1, \tilde{X}a_2$)	Rd($\tilde{X}; \tilde{X}a_1, \tilde{X}a_2, \tilde{X}a_3$)	Rd($\tilde{Y}; \tilde{Y}b_1$)	Rd($\tilde{Y}; \tilde{Y}b_1, \tilde{Y}b_1$)	Rd($\tilde{Y}; \tilde{Y}b_1, \tilde{Y}b_2, \tilde{Y}b_3$)
CIA-case1	**0.0978**	0.1579	0.2393	0.1555	0.2587	0.3282
CIA-case2	0.0954	0.1553	0.2223	0.1547	0.2544	0.3323
CIA-case3	0.0952	0.1547	0.2230	0.1538	0.2533	0.3306
CIA-case4	0.0971	0.1598	0.2503	0.1558	0.2602	0.3250
CIA-case5	0.0970	**0.1600**	**0.2504**	**0.1568**	**0.2612**	0.3253
CIA-case6	0.0956	0.1560	0.2216	0.1557	0.2556	**0.3342**
CIA-case7	0.0973	0.1597	0.2503	0.1549	0.2593	0.3248
CCA-ridge	0.0768	0.1314	0.1773	0.1517	0.2487	0.3201
CCA-PMD	0.0860	0.1404	0.2074	0.1343	0.2382	0.3030

그래프는 총 3가지를 사용하였다. 첫 번째 그래프는 분석에 이용된 두 데이터간의 유사성을 볼 수 있는 그래프 (Culhane 등, 2003; Cao 등, 2009; Min 등, 2019)로, 같은 표본에 대한 각 데이터상의 관측치를 추정된 선형조합으로 변환하여 같은 공간위에에 사영시킨후 화살표로 두 점을 이어서 표현하였다. 화살표의 길이가 짧을수록 해당하는 표본공간으로 사영된 해당 표본의 각 관측치가 가까운 것이며 이는 해당표본의 두 변량간 관련성이 높은 것을 의미한다. 또한 암종별로 다른 색을 사용하여 같은 암종에 속한 표본들의 사영된 결과가 군집을 이루는지 확인하였다. 만일 같은 색으로 표현된 여러 표본들의 사영결과가 군집을 이루고 있다면, 데이터가 각 암종별로 비슷한 정보를 지니고 있다고 판단할 수 있다. 두 번째로 사용한 그래프는 추정된 계수의 크기를 내는 그래프이다 (Culhane 등, 2003; Min 등, 2019). 추정된 계수의 절대값의 크기가 클 수록 해당 변수가 두 데이터 간의 연관성을 설명하는 공관성 계수를 계산하는데 있어 그 영향력이 크므로, 그래프에서 멀리 떨어진 변수일 수록 그 설명력이 크다고 볼 수 있다. 마지막 그래프는 추정된 선형조합변량과 해당 선형조합에 연관된 데이터 간의 피어슨 상관계수의 크기를 확인하는 correlation circle (Cao 등, 2009)이다. 원점을 기준으로 멀리 있을수록 추정된 변량이 연관된 데이터를 설명하는 정도가 높은 것을 의미하며 반지름이 0.5인 원을 그려 주요변수의 크기를 가늠할 수 있도록 하였다.

마지막으로, 여러 측도를 계산함에 있어 공관성분석의 경우 가중치가 고려된 방벙이기 때문에 측도값과 그래프를 계산할 때 가중치를 반영된 값이며, 이때의 계산을 위한 계수와 변량은 다음과 같이 사용되었다.

$$\boldsymbol{a}_i = \boldsymbol{Q}_x^{1/2}\boldsymbol{u}_i, \quad \boldsymbol{b}_j = \boldsymbol{Q}_y^{1/2}\boldsymbol{v}_j, \quad \tilde{\boldsymbol{X}} = \boldsymbol{D}^{1/2}\boldsymbol{X}\boldsymbol{Q}_x^{1/2}, \quad \tilde{\boldsymbol{Y}} = \boldsymbol{D}^{1/2}\boldsymbol{Y}\boldsymbol{Q}_y^{1/2}.$$

정준상관분석의 경우 가중치가 없는 방법이므로 $\boldsymbol{Q}_x, \boldsymbol{Q}_y, \boldsymbol{D}$ 모두 항등행렬로 치환하여 계산하는 것과 동일하다.

Table 4: Rd of opposite data

Method	$\mathbf{Rd}(\tilde{X};\tilde{Y}b_1)$	$\mathbf{Rd}(\tilde{X};\tilde{Y}b_1,\tilde{Y}b_2)$	$\mathbf{Rd}(\tilde{X};\tilde{Y}b_1,\tilde{Y}b_2,\tilde{Y}b_3)$	$\mathbf{Rd}(\tilde{Y};\tilde{X}a_1)$	$\mathbf{Rd}(\tilde{Y};\tilde{X}a_1,\tilde{X}a_2)$	$\mathbf{Rd}(\tilde{Y};\tilde{X}a_1,\tilde{X}a_2,\tilde{X}a_3)$
CIA-case1	0.0753	0.1246	0.1777	0.1219	0.2073	0.2523
CIA-case2	0.0756	0.1265	0.1713	0.1203	0.2031	0.2574
CIA-case3	0.0754	0.1258	0.1711	0.1196	0.2020	0.2556
CIA-case4	0.0741	0.1213	0.1808	0.1193	0.2036	0.2401
CIA-case5	0.0743	0.1221	**0.1817**	0.1208	0.2056	0.2420
CIA-case6	0.0758	0.1272	0.1714	0.1211	0.2043	0.2592
CIA-case7	0.0740	0.1206	0.1799	0.1178	0.2016	0.2382
CCA-ridge	**0.0768**	**0.1314**	0.1773	**0.1517**	**0.2487**	**0.3201**
CCA-PMD	0.0754	0.1188	0.1757	0.1206	0.2081	0.2751

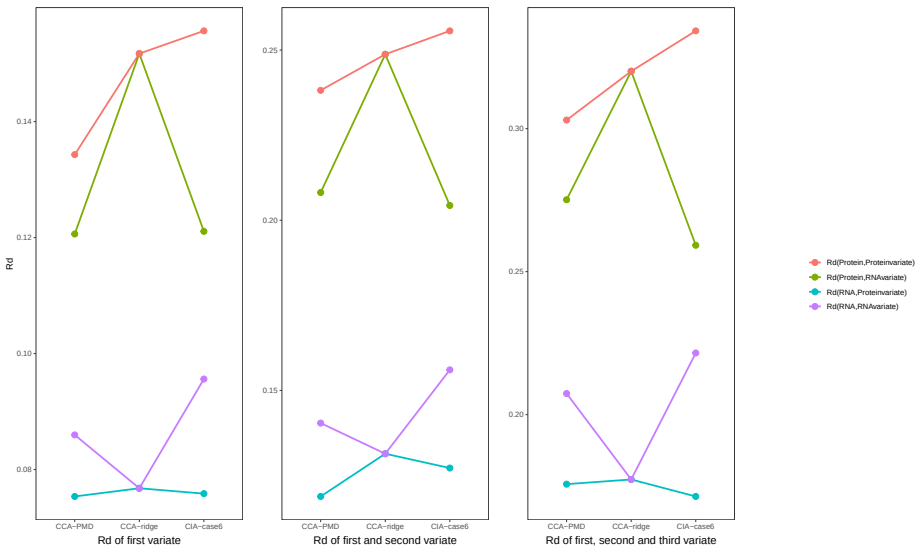

Figure 1: *Visualization of Table 3 and Table 4.*

3.3. 결과

Table 2는 각 방법의 분석결과에서 계산된 피어슨 상관계수, RV-coefficient, 그리고 CumVar값을 포함하고 있다. 피어슨 상관계수의 값을 볼 때 여러 가중치 행렬에서의 공관성분석과 정준상관분석 대부분 0.9 이상의 높은 값을 보였으며, 공관석분석에 비해 정준상관분석이 좀 더 높은 결과를 보였다. 두 방법 모두 상관관계가 높은 선형조합을 찾아내는데 좋은 성능을 보임을 알 수 있는데, 정준상관분석의 경우, 그 목적함수가 직접적으로 두 선형조합된 데이터간의 상관계수를 최대화 하는 계수를 찾는데 있으므로 그 값이 공분산분석에 비해 더 높은 값을 보이는 것이라 추측할 수 있다.

각 가중치 조합별 공관성분석결과를 비교했을 때, 각 데이터의 변수에 대한 가중치 행렬로 $\boldsymbol{Q}_x^{(1)}$과 $\boldsymbol{Q}_y^{(1)}$을 사용한 경우(CIA-case 2, 3, 6)가 가중치 행렬로 항등행렬을 설정한 경우(CIA-case1)나 $\boldsymbol{Q}_x^{(2)}$와 $\boldsymbol{Q}_y^{(2)}$를 사용한 경우(CIA-case 4, 5, 7)에 비해 높은 피어슨 상관계수값, RV-coefficient값과 CumVar값을 보였다. 이는 각 데이터에서 발현량이 높은 유전자에 더 높은 가중치를 사용하여 추정한 계수가 더 관련도가 높은 선형조합을

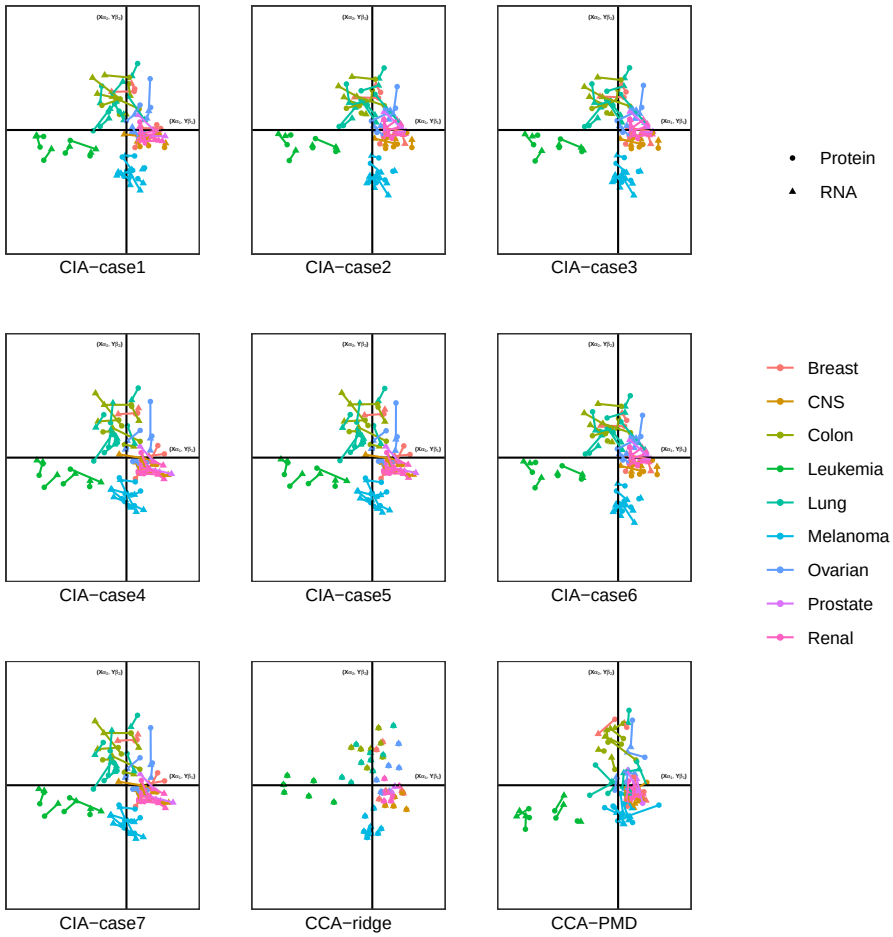

Figure 2: *Relationship between RNA sequencing data and Protein sequencing data.*

만들어 냄을 알 수 있다. 또한 표본공간에 대한 가중치를 비교해 보면 $\boldsymbol{D}^{(2)}$를 사용한 경우가 $\boldsymbol{D}^{(1)}$을 사용한 경우보다 더 좋은 성능결과를 보였다. 이는 표본에서 그 크기가 작은 그룹에 속한 표본이 높은 가중치를 할당할 때 추정된 공관성계수가 향상되는 것으로, NCI60 데이터의 경우 각 암종의 표본수 편향에 따른 영향을 줄였다고 볼 수 있다.

Table 3는 추청된 선형조합변량과 해당 변량을 계산하는데 쓰인 데이터와의 Rd값을 포함하고 있으며 Table 4는 추정된 선형조합변량과 해당 변량을 계산하는데 쓰이지 않은 데이터와의 Rd값을 포함하고 있다. Figure 1은 Table 3와 Table 4의 결과에서 CCA-PMD, CCA-ridge, 그리고 CIA-case 6의 결과를 시각화한것으로, 가로축에는 각 방법을, 세로축에는 각 방법에서의 Rd값을 표현하였으며 Rd값을 계산하는 선형조합변량과 데이터셋의 조합에 따라 각기 다른 색으로 그 값을 표시하였다. CIA의 경우 가독성을 위해 RV-coefficient와 CumVar에서 가장 좋은 결과를 보인 case 6의 경우만을 그래프로 나타내었다. Figure 1을 보면 단백질 시퀀싱 데이터를 중심으로 구했을 때 Rd값이 RNA 시퀀싱 데이터를 중심으로 Rd값을 구했을 때보다 높은데 이는 RNA 시퀀싱 데이터 변수의 개수가 단백질 시퀀싱 데이터 변수의 개수보다 많아 값이 줄어들었다고 볼 수 있

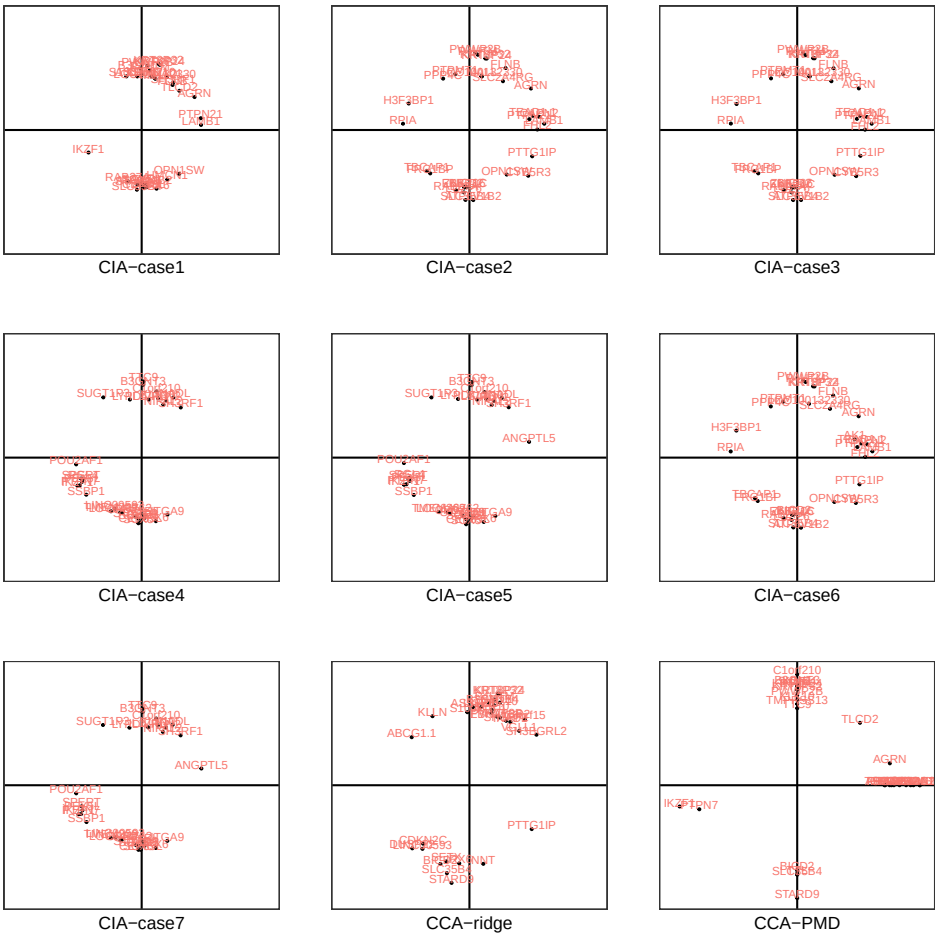

Figure 3: *Top 30 of Selected coefficients of RNA sequencing data.*

다. 분석별로 비교해보면 공관성분석은 대부분의 경우 선형조합 변량을 계산하는데 쓰인 데이터와의 Rd값이 높게 나타나는 반면, 정준상관분석은 대부분 변량을 계산하는데 쓰이지 않은 데이터와의 Rd값이 공관석분석보다 높은 경향을 보였는데, 이러한 결과는 이는 공관성분석의 목적함수가 두 데이터간의 공관성계수뿐만 아니라 각 데이터의 가중 분산도 고려하는 값으로 (Dray 등, 2003), 정준상관분석에 비해 변량을 계산하는데 쓰인 연관된 데이터와의 관련성이 좀 더 반영된 결과로 추측해 볼 수 있다. 정준상관분석 내에서 CCA-PMD와 CCA-ridge를 비교할 경우 CCA-PMD가 CCA-ridge에 비해 대부분의 Rd에서 낮은 값을 보였는데, 이는 CCA-PMD가 벌점함수 이용을 통한 변수 선택으로 인해 선택된 변수만 사용하게 됨에 따른 약간의 정보 손실이 원인이라고 추측해 볼 수 있다.

Figure 2는 각 데이터를 추정된 계수를 이용해 같은 표본공간에 사영시킨 그래프로, 하나의 표본에서 얻은 RNA 시퀀싱 데이터와 단백질 시퀀싱 데이터간의 각각 사영된 지점간의 거리가 가깝다면 두 변량간 관련성이 높다고 판단한다. 먼저 군집화된 경향을 볼때 전반적으로 모든 방법이 비슷한 결과를 보이는 것을 알 수 있다. 첫 번째 추정변량으로 결장암과 백혈병은 신장암, 중추신경계암과 구별할 수 있으며 두 번째 추정변량이

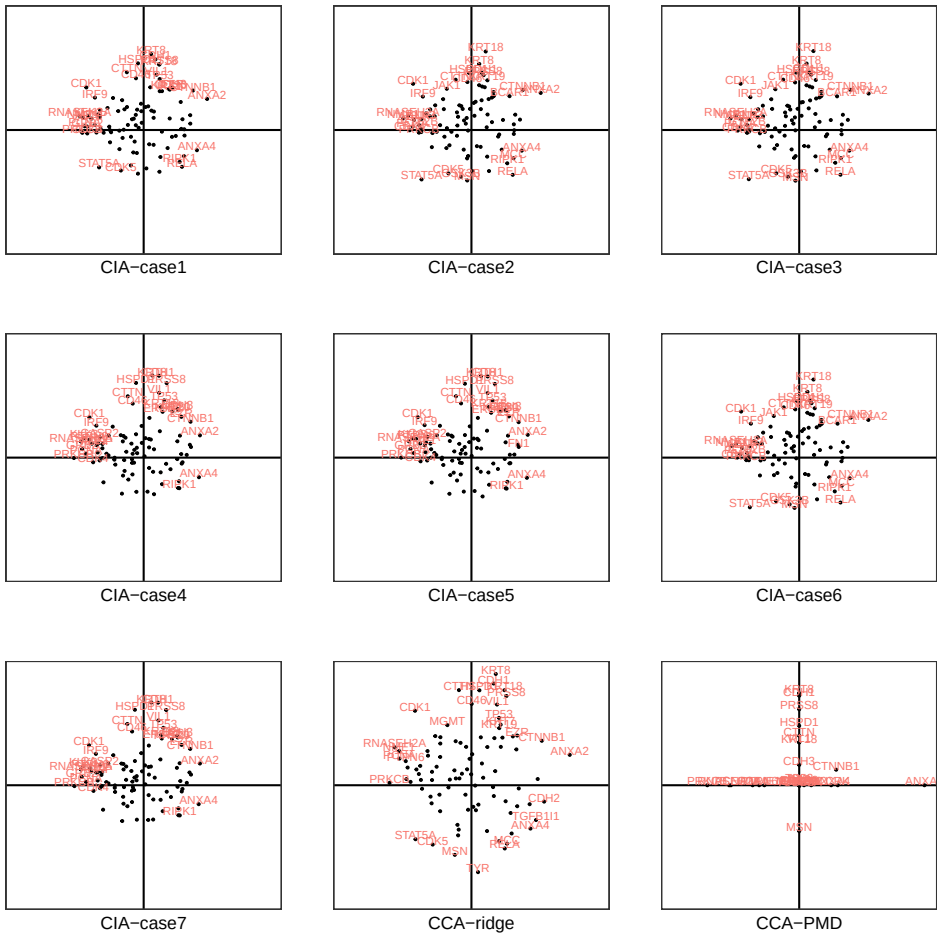

Figure 4: *Coefficients of protein sequencing data.*

상피 특성을 가진 세포에서 발생한 흑색종암을 다른 암과는 확실하게 구분해내는 양상을 보이는데 이 결과는 NCI60 자료를 이용한 이전의 연구 결과 (Culhane 등, 2003)에서도 확인된 결과이다. CCA가 CIA와 마찬가지로 큰 차이없이 비슷한 군집화 결과를 보이는데 반해, 두 데이터간의 관련성을 나타내는 화살표의 길이에서는 확연한 차이가 보이는데, CCA-ridge의 경우 화살표의 길이가 매우 짧아 거의 두 점이 겹치다시피 하는 반면 CCA-PMD의 경우 화살표의 길이가 상대적으로 길고 CIA의 결과에 비해서도 짧지 않은 편임을 알 수 있다. 이는 피어슨 상관계수에서 보았던 결과와 마찬가지의 이유에서 비롯된 결과로 추측해볼 수 있는데, CCA의 경우 그 목적함수가 두 데이터간의 상관계수를 최대화 하는 것이므로 추정된 두 선형조합간의 관련성이 최대화 되는 결과를 지향한다고 볼 수 있다. 따라서 전반적으로 CCA의 결과가 CIA에 비해 더 짧은 화살표의 길이를 보이리라 예상할 수 있는데, 다만 CCA-PMD의 경우 CCA-ridge와 달리 변수 선택을 통한 정보의 손실로 인해 두 점 사이의 거리가 상대적으로 멀어진 것으로 추측해 볼 수 있다.

Figure 3과 Figure 4는 각각 RNA 시퀀싱 데이터와 단백질 시퀀싱 데이터의 첫 두 개의 추정 선형조합계수를 각 축의 좌표로 하는 그래프로, 원점에서 머리가 가장 먼 30개의 유전자에 이름을 명시하였다. 각 데이터의

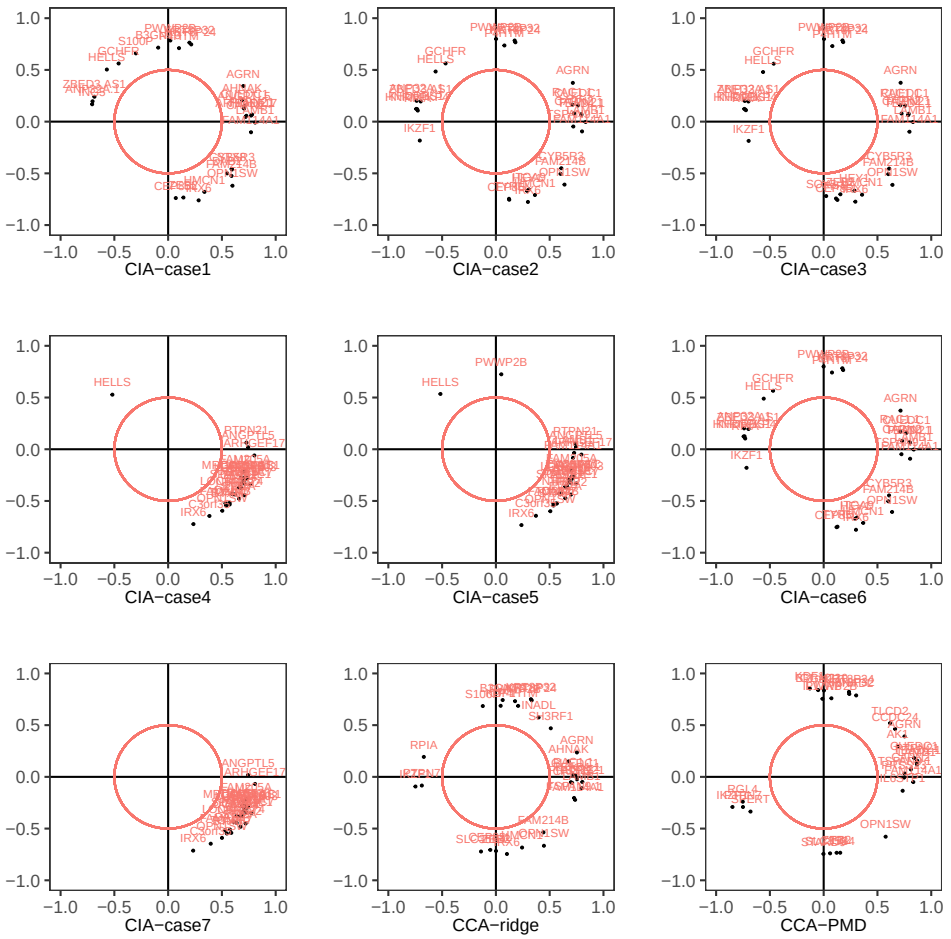

Figure 5: *Top 30 of selected correlation circle of RNA sequencing data.*

자세한 Top 30 유전자 리스트는 Appendix Table A.1과 A.2로 제공하였다. 이때 Figure 3의 경우 유전자의 개수가 많아 상위 30개의 유전자만 그 점을 표시하였고 Figure 4는 모든 데이터를 나타내었다. 단 CCA-PMD의 경우 단백질 시퀀싱 데이터에서 선택된 변수가 30개 미만인 25개로, Figure 4에서는 선택된 25개의 유전자만 표시되어있다. Figure 4를 살펴보면 대장암과 연관있는 KRT18 (Zhang 등, 2019), 삼중음성 유방암과 관련이 깊다는 연구결과가 발표된 CDK4 (Sheikh 등, 2021)등이 모든 결과에서 큰 값을 가지는 등, 두 방법 모두 의미 있는 변수를 찾아냄을 알 수 있다. 하지만 각 방법별로, 사용하는 가중치행렬의 조합별로 Top 30 유전자순위에 차이가 존재하는 것을 관찰할 수 있다. Table A.1과 Table A.2를 살펴보면 $\boldsymbol{Q}_x^{(2)}, \boldsymbol{Q}_y^{(2)}$를 사용한 공관성분석에서의 결과가 대체적으로 정준상관분석과 비슷함도 알 수 있는데, 이는 각 변수에 $\boldsymbol{Q}_x^{(2)}, \boldsymbol{Q}_y^{(2)}$의 가중치행렬을 사용할 때, 다시 말해 각 변수의 크기에 반비례하는 가중치를 이용한 공관성 분석의 결과가 정준상관분석의 결과와 비슷하다는 것으로 두 경우가 비슷하게 데이터를 표준화하여 이해하는 방식에 가깝기 때문이라고 생각해볼 수 있다. 또한, 공관성분석중 $\boldsymbol{Q}_x^{(1)}, \boldsymbol{Q}_y^{(1)}$을 사용한 경우 유방암과 연관이 있는 CYB5R3 (Lund 등, 2015)와 BCAR1 (Centonze 등, 2021)이 각각 RNA 시퀀싱 데이터와 단백질 시퀀싱 데이터에서 선

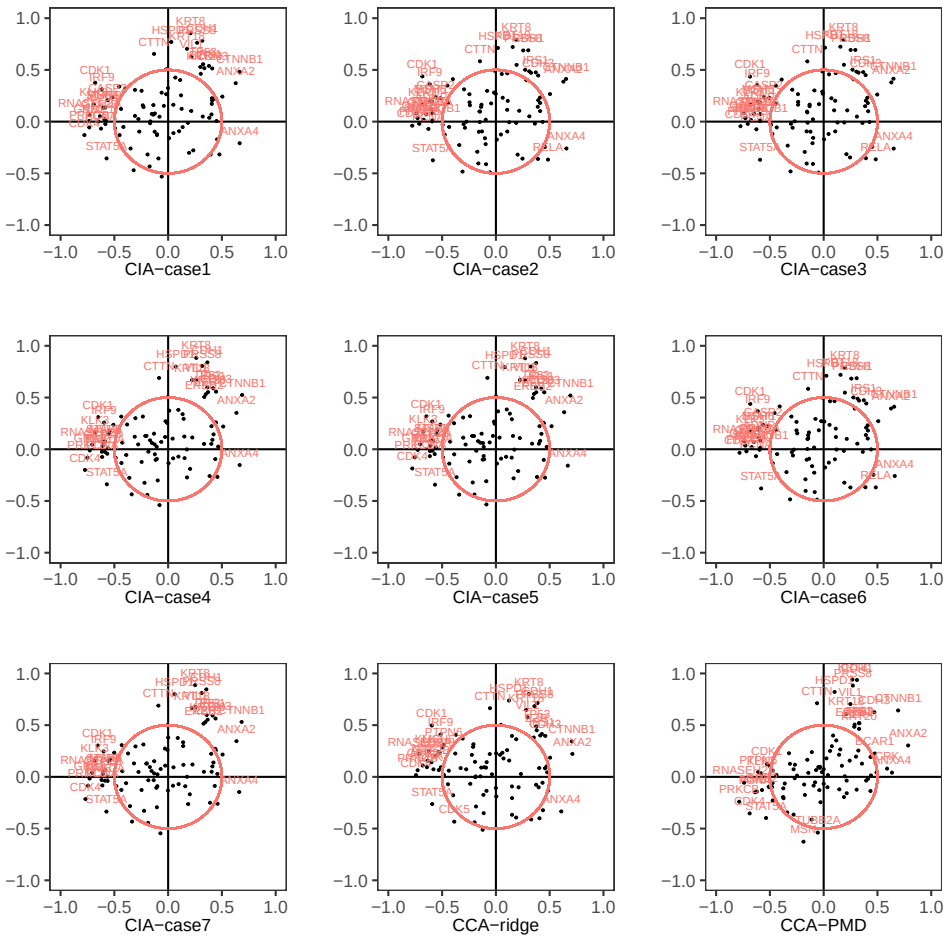

Figure 6: *Correlation circle of protein sequencing data.*

택되었으나 정준상관분석과 $\boldsymbol{Q}_x^{(2)}, \boldsymbol{Q}_y^{(2)}$를 사용한 공관성분석의 경우 선택되지 않았다. 이러한 결과는 유전자 발현량에 대한 정보를 반영하는 정도에 따라 RV-coefficient값과 CumVar값뿐만이 아니라 각 유전자의 설명력에 대한 추정치가 일부 달라질 수 있음을 의미한다.

Figure 5와 Figure 6는 RNA 시퀀싱 데이터와 단백질 시퀀싱 데이터에서의 correlation circle로 각 유전자의 좌표는 데이터와 첫 두 개의 추정변량간의 상관계수값이며 상관계수가 가장 큰 30의 유전자에 이름을 명시하였다. 각 데이터의 자세한 Top 30 리스트는 Appendix에 있는 Table A.3와 Table A.4에 나타나있다. Figure 3와 같은 이유로 Figure 5는 상위 30개의 유전자만 표시하였으며 Figure 6는 전부 표시했다. Figure 6를 보면 모든 분석에서 난소암과 연관있는 PRSS8 (Tamir 등, 2016), 대장암, 난소암, 유방암등의 여러 암종과 관련성이 연구된 KRT8 (Tan 등, 2017)이 모든 분석결과에서 큰 값을 가지는등의 결과를 확인할 수 있었다. 정준상관분석과 공관성분석의 RNA 시퀀싱 데이터 correlation circle을 비교했을 때 Figures 3, 4와는 다른 양상으로 $\boldsymbol{Q}_x^{(1)}, \boldsymbol{Q}_y^{(1)}$을 가중치행렬로 이용한 경우가 정준상관분석의 결과와 매우 비슷한 리스트를 보이고 있음을 확인할 수 있다.

4. 결론

본 연구는 두 데이터간의 연관성에 대한 정보를 얻는 방법인 정준상관분석과 공관성분석을 소개하고 실제 유전체 데이터인 NCI60 데이터에 적용하여 그 결과를 비교하였다. 정준상관분석의 경우 표본수보다 변수의 수가 더 많은 NCI60 데이터의 특성상, 분석방법을 그대로 적용하는 대신 고차원 데이터로의 적용이 가능하도록 확장된 방법인 CCA-ridge와 CCA-PMD를 사용하였으며, 공관성분석의 경우 가중치에 따른 차이를 살펴보기 위해 일곱가지의 가중치행렬 조합을 고려하였다.

정준상관분석과 공관성분석 모두 추정된 선형조합변량간의 관련성이 높게 나타났으며 데이터에 비추어 보았을떼 유의미한 선형계수조합을 추정해내었고 데이터 표본에 그룹정보가 있을경우 그 정보가 유의미하게 드러나는 결과를 추출할 수 있음을 알 수 있엇다. 그러다 두 분석방법간의 차이 또한 분명하게 드러났는데, 공관성분석은 두 데이터간의 상관관계의 정도와 더불어 각각의 데이터의 변동도 함께 고려하는 반면, 정준상관분석은 두 데이터간의 상관정도에 좀 더 중점을 두는 결과를 보였다. 정준상관분석의 경우 CCA-ridge 모형이 두 변량간의 관련성이 더 높은 결과를 보였는데, 이는 CCA-PMD의 경우 벌점함수를 통한 변수선택의 결과로 일부 중요한 변수만을 사용하여 결과를 도출했기 때문으로 보이며 약간의 정보의 손실에 의한 추정정준상관계수의 손실을 대가로 하여 해석의 용이함을 얻는 것이라 볼 수 있다. 공관성분석의 경우 정준상관분석과 달리 별다른 조치 없이도 고차원데이터에서도 사용할 수 있다는 장점이 있어 그대로 데이터에 적용할 수 있었다. 또한 연구목적에 따라 적절한 가중치를 설정할 수 있음에따라 일곱가지의 다양한 경우로 가중치행렬을 설정, 적용하였다. 공관성분석의 데이터 분석결과, 목적함수가 상관계수 뿐만 아니라 각 데이터의 가중분산까지 함께 고려함에도 불구하고 두 추정선형결합간의 관련성이 충분히 높은 결과를 얻을 수 있었다. 또한 정준상관분석과 비교하여볼 때, 각 추정된 선형결합계수가 상대 데이터와의 관련성은 정준상관분석에 비해서는 낮지만 해당하는 데이터와의 관련성은 충분히 높게 반영된 결과로 추정되었다. 이는 공관성분석에서 사용하는 목표함수인 공관성계수가 상관계수뿐만이 아니라 두 데이터의 가중분산을 고려하는 값에서 기인하는 것으로 생각된다. 또한 공관성분석은 그 사용하는 가중치행렬의 차이에 따라 결과에서도 차이를 보였는데, NCI60 데이터에의 적용결과를 보면 유전자의 발현량이 많을수록 더 높은 가중치를 고려한 경우가 더 낮은 가중치를 고려할 때, 그리고 적은 표본을 가진 암종에 더 높은 가중치를 고려한 경우 두 추정선형조합간의 관련도가 높게 나타났다. 이러한 결과는 각 변수의 변동성을 고려하여, 변동성이 큰 경우에 가중치를 크게 줌으로써 그 변동성에 대해 더 집중하고, 표본크기의 편향은 그 크기의 역가중치로 보정함으로써 더 개선된 결과를 이끌어낸 것으로 보인다.

본 연구에서는 정준상관분석과 공관성분석은 같은 표본으로부터 얻은 두 다변량 그룹간의 관련성을 다룬다는 점에서 유사한 분석방법으로 보이나 두 변수 간의 관련성을 정의하는 측도의 선택에 차이가 있으며 그로 인해 결과에 분명한 차이가 있음을 보였다. 본 연구에서 공관성분석의 경우 변수선택에 대한 가능성을 고려하지 않았으나, 추후 변수선택이 가능한 공관성분석과의 비교연구가 흥미로운 연구주제가 될 것으로 보이며, 본 연구의 결과는 향후 연구자들이 연구 목적에 따른 적절한 분석방법을 선택하는 데 도움이 될 수 있을것이다.

Appendix A: Table

Table A.1: Ranking of RNA sequencing data coefficient size

CIA-case1	CIA-case2	CIA-case3	CIA-case4	CIA-case5	CIA-case6	CIA-case7	CCA-ridge	CCA-PMD
KRT8P32	PWWP2B	PWWP2B	TTC9	TTC9	PWWP2B	TTC9	STARD9	STARD9
KRT8P24	P4HTM	P4HTM	B3GNT3	B3GNT3	P4HTM	B3GNT3	KRT8P24	C1orf210
PWWP2B	KRT8P24	KRT8P24	SUGT1P3	SUGT1P3	KRT8P24	SUGT1P3	KRT8P32	PTPN21
B3GNT3	KRT8P32	KRT8P32	C1orf210	C1orf210	KRT8P32	C1orf210	SLC35B4	LAMB1
P4HTM	ATP6V1B2	CYB5R3	INADL	SOX5	ATP6V1B2	INADL	P4HTM	IKZF1
IRX6	SLC35B4	SLC35B4	SOX5	INADL	FLNB	SOX5	SYNE4	TEAD1.1
SLC35B4	CYB5R3	ATP6V1B2	IRX6	IRX6	SLC35B4	CEP85L	B3GNT3	CUEDC1
AGRN	FLNB	FLNB	CEP85L	CEP85L	CYB5R3	IRX6	C1orf210	B3GNT3
INADL	AGRN	AGRN	IKZF1	IKZF1	AGRN	IKZF1	BICD2	KDF1
CEP85L	LAMB1	LAMB1	SSBP1	ITGA9	LAMB1	SSBP1	C6orf15	SYNE4
ZEB2	CAPN2	CAPN2	PTPN7	PTPN7	CAPN2	RLN2	DUSP22	KRT8P24
KDF1	LUZP6	LUZP6	ITGA9	SSBP1	LUZP6	PTPN7	NNT	FAM114A1
OPN1SW	PTTG1IP	PTTG1IP	SH3RF1	SOX5.1	PTTG1IP	KLK10	SETX	INADL
S100P	RAB27A	RAB27A	RLN2	SH3RF1	RAB27A	SH3RF1	SH3BGRL2	KRT8P32
SH3RF1	FHL2	FHL2	SOX5.1	TMEM209	FHL2	SOX5.1	IRX6	CAPN2
STARD9	H3F3BP1	H3F3BP1	KLK10	RLN2	H3F3BP1	ZEB2	LYPD5	PTPN7
HMCN1	RPIA	RPIA	ZEB2	KLK10	RPIA	ITGA9	ASS1P8	PWWP2B
PTPN21	TEAD1.1	TEAD1.1	TMEM209	ZEB2	TEAD1.1	STARD9	KLK10	SLC35B4
C1orf210	FKBP1C	FKBP1C	POU2AF1	LYPD4	ZNF106	TMEM209	KLLN	TTL
ATP6V1B2	ZNF106	SLC2A4RG	STARD9	POU2AF1	FKBP1C	POU2AF1	LINC00593	RAC1.1
LAMB1	SLC2A4RG	ZNF106	LYPD4	SPERT	PTPMT1	PEX5L	KDF1	AGRN
ASS1P8	PTPMT1	PTPMT1	PEX5L	TYRL	SLC2A4RG	LYPD4	INADL	KLK10
TYRL	BICD2	BICD2	SPERT	PEX5L	BICD2	LOC646762	S100P	BICD2
LOC100132330	PPP4C	PPP4C	IL17RB	CPB2	PPP4C	IL17RB	EBNA1BP2	TSPAN9.1
FLNB	TBCAP1	TBCAP1	LOC646762	IL17RB	FRG1BP	SPERT	PWWP2B	TLCD2
IKZF1	FRG1BP	OPN1SW	CPB2	STARD9	TBCAP1	CPB2	SH3RF1	PRSS23
RAB27A	OPN1SW	FRG1BP	TYRL	LOC646762	OPN1SW	LINC00593	CDKN2C	TMPRSS13
MALL	PTPN21	PTPN21	RGL4	RGL4	PTPN21	NIPAL2	PTTG1IP	AHNAK
SOX5.1	LOC100132330	LOC100132330	LINC00593	NIPAL2	AK1	TYRL	ABCG1.1	FHL2
TLCD2	MALL	MALL	NIPAL2	ANGPTL5	LOC100132330	ANGPTL5	VGLL1	TTC9

Table A.2: Ranking of protein sequencing data coefficient size

CIA-case1	CIA-case2	CIA-case3	CIA-case4	CIA-case5	CIA-case6	CIA-case7	CCA-ridge	CCA-PMD
KRT8	KRT18	KRT18	CDH1	CDH1	KRT18	CDH1	KRT8	ANXA2
CDH1	ANXA2	ANXA2	KRT8	KRT8	ANXA2	KRT8	CDH1	KRT8
PRSS8	CDK1	CDK1	PRSS8	PRSS8	CDK1	PRSS8	KRT18	CDH1
KRT18	STAT5A	STAT5A	HSPD1	HSPD1	STAT5A	HSPD1	CTTN	PRKCB
CDK1	KRT8	KRT8	VIL1	VIL1	KRT8	VIL1	HSPD1	PRSS8
HSPD1	NME1	NME1	CDK1	PRKCB	NME1	CDK1	PRSS8	RNASEH2A
ANXA2	CTNNB1	CTNNB1	CTTN	CDK1	CTNNB1	CTTN	ANXA2	PCNA
CTTN	RNASEH2A	RNASEH2A	PRKCB	CTTN	RNASEH2A	PRKCB	CDK1	HSPD1
CTNNB1	RELA	CDH1	TP53	TP53	RELA	TP53	TYR	CTTN
RNASEH2A	CDH1	RELA	RNASEH2A	RNASEH2A	CDH1	RNASEH2A	VIL1	MSN
VIL1	HSPD1	HSPD1	CDH3	CDH3	HSPD1	CDH3	CD46	VIL1
NME1	PRSS8	PRSS8	KLK3	KLK3	CDK4	KLK3	RNASEH2A	KRT18
PRKCB	CDK4	CDK4	IRS1	IRS1	PRSS8	IRS1	CTNNB1	CDK4
STAT5A	IRF9	IRF9	GRB2	ANXA2	IRF9	GRB2	STAT5A	CTNNB1
IRF9	PCNA	ADNP	CTNNB1	GRB2	PCNA	CTNNB1	PRKCB	PTPN6
PCNA	ADNP	PCNA	ANXA2	CTNNB1	ADNP	ANXA2	NME1	CRK
TP53	CTTN	CTTN	CD46	CD46	CTTN	CD46	TP53	NME1
CD46	KRT19	KRT19	ANXA4	ANXA4	KRT19	IRF9	MSN	ANXA4
ANXA4	MSN	MSN	IRF9	EZR	MSN	ANXA4	PCNA	CDH3
KLK3	ANXA4	ANXA4	KRT20	PTPN6	ANXA4	KRT20	RELA	CDH2
GRB2	BCAR1	CDK5	EZR	KRT20	BCAR1	EZR	CDK5	MVP
CDK4	CDK5	BCAR1	KRT18	KRT18	CD46	KRT18	ANXA4	TP53
RELA	CD46	GSK3B	PTPN6	IRF9	CDK5	PCNA	TGFB1I1	EZR
EZR	GSK3B	CD46	PCNA	PCNA	PRKCB	PTPN6	CDH2	ERBB2
CDH3	PRKCB	GTF2B	ERBB2	ERBB2	MCC	ERBB2	MCC	TUBB2A
ADNP	MCC	PRKCB	NME1	NME1	GSK3B	NME1	KRT7	-
KRT20	GTF2B	MCC	CDK4	CDK4	JAK1	CDK4	EZR	-
IRS1	JAK1	KLK3	CASP2	CASP2	GTF2B	ADNP	PTPN6	-
RIPK1	RIPK1	RIPK1	RIPK1	RIPK1	RIPK1	RIPK1	MGMT	-
CDK5	KLK3	JAK1	ADNP	FN1	KLK3	CASP2	KRT19	-

Table A.3: Ranking the correlation coefficients between the RNA sequencing data and the first two estimated variates

CIA-case1	CIA-case2	CIA-case3	CIA-case4	CIA-case5	CIA-case6	CIA-case7	CCA-ridge	CCA-PMD
OPN1SW	OPN1SW	OPN1SW	FSIP2	FSIP2	OPN1SW	FSIP2	KRT8P32	IKZF1
IRX6	LAMB1	LAMB1	SDR9C7	SDR9C7	LAMB1	SDR9C7	KRT8P24	PTPN21
LAMB1	IRX6	IRX6	EVI2A	EVI2A	IRX6	EVI2A	PTPN21	TEAD1.1
FAM214B	AGRN	AGRN	ARHGEF17	ARHGEF17	FAM114A1	ARHGEF17	LAMB1	LAMB1
KRT8P32	FAM114A1	KRT8P32	PAX1	OPN1SW	AGRN	PAX1	OPN1SW	KDF1
PWWP2B	KRT8P32	FAM114A1	TOMM7	DAB2	TEAD1.1	TOMM7	AGRN	CUEDC1
PTPN21	PWWP2B	PWWP2B	CACNA1C	IRX6	KRT8P32	CACNA1C	FAM114A1	KRT8P24
AGRN	PTPN21	PTPN21	FABP5	CACNA1C	HMCN1	FABP5	TSPAN9.1	AGRN
FAM114A1	TEAD1.1	HMCN1	OPN1SW	PAX1	PTPN21	PLVAP	TEAD1.1	INADL
KRT8P24	HMCN1	KRT8P24	PLVAP	FABP5	PWWP2B	LOC339803	IKZF1	B3GNT3
TEAD1.1	KRT8P24	TEAD1.1	LOC339803	TOMM7	KRT8P24	NTRK2	PWWP2B	KRT8P32
HELLS	FAM214B	FAM214B	DAB2	STX8	CUEDC1	SPATA31E1	IRX6	C1orf210
HMCN1	CUEDC1	CUEDC1	STX8	LOC339803	FAM214B	NT5C1A	B3GNT3	FAM114A1
CYB5R3	CEP85L	CEP85L	SPATA31E1	PLVAP	ANP32A.1	OPN1SW	CUEDC1	OPN1SW
ZEB2	ANP32A.1	ANP32A.1	NT5C1A	TM4SF20	CEP85L	STX8	SLC35B4	CAPN2
STX8	CYB5R3	CYB5R3	TM4SF20	FAM214B	TYRL	OR1F1	HMCN1	CCDC24
DAB2	TYRL	TYRL	IRX6	SPATA31E1	CYB5R3	FAM71B	RAC1.1	TLCD2
CEP85L	KHDRBS1	KHDRBS1	NTRK2	NT5C1A	KHDRBS1	DAB2	P4HTM	PTPN7
ARHGEF17	HNRNPKP4	HELLS	OR1F1	ANGPTL5	HNRNPKP4	TM4SF20	ZEB2	RGL4
ANP32A.1	HELLS	HNRNPKP4	FAM71B	C3orf35	P4HTM	FAM205A	CAPN2	PWWP2B
GCHFR	P4HTM	P4HTM	C3orf35	HELLS	HELLS	C3orf35	CEP85L	SPERT
ZBED3.AS1	RPIA	RPIA	FAM205A	PTPN21	IKZF1	HR	PRSS23	IL17RB
ING5	GCHFR	RAC1.1	FAM214B	NTRK2	RPIA	IRX6	INADL	TSPAN9.1
CUEDC1	CAPN2	CAPN2	ANGPTL5	OR1F1	GCHFR	LOC642574	RPIA	TTL
S100P	RAC1.1	GCHFR	HR	FAM71B	CAPN2	RAC1	S100P	AK1
CAPN2	IKZF1	HEY1	HELLS	FAM205A	HEY1	MRGPRG.AS1	FAM214B	PRSS23
B3GNT3	HEY1	ZBED3.AS1	LOC642574	HR	RAC1.1	ANGPTL5	SH3RF1	IL6STP1
AHNAK	ZBED3.AS1	IKZF1	PTPN21	LAMB1	ITGA9	DCANP1	PTPN7	STARD9
P4HTM	ITGA9	SOX5.1	MRGPRG.AS1	PWWP2B	ZBED3.AS1	FAM214B	AHNAK	ZEB2
ANGPTL5	TSPAN9.1	ZEB2	LIF	FAM114A1	TSPAN9.1	LIF	KDF1	SLC35B4

Table A.4: Ranking the correlation coefficients between the protein sequencing data and the first two estimated variates

CIA-case1	CIA-case2	CIA-case3	CIA-case4	CIA-case5	CIA-case6	CIA-case7	CCA-ridge	CCA-PMD
KRT8	KRT8	KRT8	KRT8	KRT8	KRT8	KRT8	KRT8	CDH1
CDH1	CDK1	CDK1	CDH1	CDH1	CDK1	CDH1	CDH1	KRT8
CTNNB1	CDK4	CDK4	PRSS8	CTNNB1	CDK4	PRSS8	CTNNB1	CTNNB1
PRSS8	CTNNB1	CTNNB1	CTNNB1	PRSS8	CTNNB1	CTNNB1	CDK1	PRSS8
CDK4	CDH1	CDH1	HSPD1	CDK4	CDH1	HSPD1	PRSS8	ANXA2
CDK1	RNASEH2A	RNASEH2A	CDK4	HSPD1	RNASEH2A	CDK4	HSPD1	HSPD1
HSPD1	KRT18	KRT18	PRKCB	PRKCB	ANXA2	PRKCB	RNASEH2A	PRKCB
PRKCB	ANXA2	PRSS8	GRB2	ANXA2	KRT18	GRB2	ANXA2	CDH3
GRB2	PRSS8	ANXA2	ANXA2	GRB2	PRSS8	ANXA2	CDK4	CDK4
RNASEH2A	GRB2	GRB2	VIL1	VIL1	GRB2	IRS1	PRKCB	VIL1
ANXA2	IRF9	HSPD1	IRS1	IRS1	PRKCB	VIL1	KRT18	RNASEH2A
KRT18	HSPD1	IRF9	CDK1	CDK1	IRF9	CDK1	ANXA4	CTTN
KLK3	PRKCB	PRKCB	CDH3	CDH3	HSPD1	CDH3	GRB2	STAT5A
ANXA4	ANXA4	KLK3	RNASEH2A	RNASEH2A	ANXA4	RNASEH2A	KLK3	PCNA
IRF9	KLK3	ANXA4	KRT18	KRT18	KLK3	KRT18	CTTN	MSN
STAT5A	STAT5A	STAT5A	TP53	TP53	STAT5A	KLK3	IRF9	KRT18
CTTN	NME1	NME1	KLK3	KLK3	NME1	TP53	NME1	NME1
VIL1	GTF2B	GTF2B	CTTN	CTTN	GTF2B	CTTN	PCNA	ANXA4
IRS1	MSH6	MSH6	ANXA4	ANXA4	MSH6	ANXA4	STAT5A	TP53
CDH3	PCNA	PCNA	KRT20	KRT20	PCNA	STAT5A	VIL1	IRS1
NME1	STAT1	STAT1	STAT5A	STAT5A	STAT1	KRT20	GTF2B	PTPN6
TP53	ADNP	ADNP	IRF9	IRF9	CASP2	IRF9	TP53	CRK
GTF2B	CASP2	RELA	EZR	EZR	ADNP	EZR	CDH3	GRB2
PCNA	RELA	IRS1	PCNA	PCNA	RELA	NME1	ADNP	KLK3
STAT1	IRS1	CASP2	NME1	NME1	IRS1	PCNA	IRS1	ERBB2
MSH6	CTTN	CTTN	STAT1	GTF2B	CTTN	ERBB2	EZR	KRT20
KRT20	PRKCA	CDH3	GTF2B	STAT1	PRKCA	STAT1	MSH6	EZR
EZR	CDH3	PRKCA	ERBB2	ERBB2	CDK7	GTF2B	CDK7	CDK1
ADNP	CDK7	SMARCB1	MSH6	MSH6	CDH3	PRKCA	CDK5	TUBB2A
CASP2	SMARCB1	CDK7	PRKCA	PRKCA	SMARCB1	MSH6	PTPN6	BCAR1

References

Centonze G, Natalini D, Salemme V, Costamagna A, Cabodi S, and Defilippi P (2021). p130Cas/BCAR1 and p140Cap/SRCIN1 adaptors: The Yin Yang in breast cancer?, *Frontiers in Cell and Developmental Biology*, **9**, 729093.

Culhane AC, Perrière G, and Higgins DG (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis, *BMC Bioinformatics*, **4**, 59.

Vinod HD (1976). Canonical ridge and econometrics of joint production, *Journal of Econometrics*, **4**, 147–1663.

Hotelling H (1936). Relations between two sets of variates, *Biometrika*, **28**, 321–377.

Wilms I and Croux C (2016). Robust sparse canonical correlation analysis, *BMC Systems Biology*, **10**, 72.

Lê Cao KA, Martin PG, Robert-Granié C, and Besse P (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study, *BMC Bioinformatics*, **10**, 34.

Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, Weinstein JN, Pommier Y, and Reinhold WC (2010). mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities, *Molecular Cancer Therapeutics*, **9**, 1080–1091.

Lund RR, Leth-Larsen R, Caterino TD, Terp MG, Nissen J, Lænkholm AV, Jensen ON, and Ditzel HJ (2015). NADH-Cytochrome b5 reductase 3 promotes colonization and metastasis formation and is a prognostic marker of disease-free and overall survival in estrogen receptor-negative breast cancer, *Molecular & Cellular Proteomics (MCP)*, **14**, 2988–2999.

Meng C, Kuster B, Culhane AC, and Gholami AM (2014). A multivariate approach to the integration of multi-omics datasets, *BMC Bioinformatics*, **15**, 162.

Tenenhaus M (1998). *La régression PLS: théorie et Pratique*, Editions Technip.

Min EJ, Safo SE, and Long Q (2019). Penalized co-inertia analysis with applications to -omics data, *Bioinformatics (Oxford, England)*, **35**, 1018–1025.

Nishizuka S, Charboneau L, Young L *et al.* (2003). Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 14229–14234.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Tibshirani R, Saunders M, Rosset S, Zhu J, and Knight K (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 91–108.

Tibshirani R, Hastie T, Narasimhan B, and Chu G (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**, 104–117.

Dudoit S, Fridly J, and Speed TP (2001). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, **96**, 1151–1160.

Sheikh MS and Satti SA (2021). The emerging CDK4/6 inhibitor for breast cancer treatment, *Molecular and Cellular Pharmacology*, **13**, 9.

Dray S, Chessel D, and Thioulouse J (2003). Co-inertia analysis and the linking of ecological data tables, *Ecology*, **84**, 3078–3089.

Dolédec S and Chessel D (1994). Co-inertia analysis: An alternative method for studying species–environment relationships, *Freshwater Biology*, **31**, 277–294.

Tamir A, Gangadharan A, Balwani S *et al.* (2016). The serine protease prostasin (PRSS8) is a potential biomarker

for early detection of ovarian cancer, *Journal of Ovarian Research*, **9**, 20.

Tan HS, Jiang WH, He Y *et al.* (2017). KRT8 upregulation promotes tumor metastasis and is predictive of a poor prognosis in clear cell renal cell carcinoma, *Oncotarget*, **8**, 76189–76203.

Witten DM, Tibshirani R, and Hastie T (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics (Oxford, England)*, **10**, 515–534.

Zhang J, Hu S, and Li Y (2019). KRT18 is correlated with the malignant status and acts as an oncogene in colorectal cancer, *Bioscience Reports*, **39**, 8.

Received October 23, 2023; Revised October 30, 2023; Accepted October 31, 2023

오믹스 자료를 이용한 정준방법 비교

이승수[a], 민은정[1, a,b]

[a]가톨릭대학교 의생명 · 건강과학과; [b]가톨릭대학교 의과대학 의생명과학교실

요 약

생명현상의 복잡한 시스템에 대한 이해를 위한 융합분석의 중요성이 점점 커지고 있다. 하나의 연구대상을 다양한 관점에서 관찰하여 얻게 되는 여러 데이터의 융합분석은 통해 좀 더 대상에 대한 깊은 이해를 가능하게 한다. 본 연구에서는 그중에서도 특히 하나의 샘플에서 두개의 고차원 데이터가 생성된 경우 다룰 수 있는 분석인 공관성분석과 정준상관분석을 비교하였다. 정준상관분석의 경우 고차원 데이터를 다룰 수 없는 단점이 있기에, 해당 문제를 극복하기 위하여 능형상수를 이용하는 방법(CCA-ridge)과 각 데이터의 공분산행렬을 항등행렬로 가정하여 벌점화 특이값분해를 이용한 방법(CCA-PMD) 두 가지를 고려하였으며 각 방법을 NCI60 세포주 패널에서 얻은 RNA 시퀀싱 데이터와 단백질 시퀀싱 데이터 분석에 적용하였다. 그 결과 정준상관분석의 경우 두 정준변수간의 상관관계에 좀 더 집중하는 반면 공관성분석은 각 데이터의 선형조합간의 상관관계뿐 아니라 각 선형조합의 변동성을 함께 고려함을 확인할 수 있었다. 또한 공관성분석의 경우 여러가지의 가중치행렬을 고려하여 그 결과값을 비교하고 중요 시사점을 도출하였다.

주요용어: 정준상관분석, 공관성분석, 융합분석, 오믹스, NCI60

이 논문은 2021년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2021R1F1A1058613).

[1]교신저자: (06591) 서울특별시 서초구 반포대로 222, 가톨릭대학교 의과대학. E-mail: ej.min@catholic.ac.kr