다출처자료인자모형의 일반화와 통계이론
(Generalization of multi-source integrative factor models and statistical inference)

연구의 필요성

다출처 자료란 서로 다른 특성을 가진 여러 개의 자료 또는 데이터셋을 통합적으로 이른다. 기술의 발전으로 정보의 대규모 수집이 가능해지고 있어, 한 관측유닛에 대한 여러 개의 자료 또는 한 현상에 대한 여러개의 실험·관찰자료가 폭발적으로 늘어나고 있다. 이러한 다출처자료를 연계하여 기저의 시스템을 추출하고 미관측된 현상을 예측하는 방법론의 개발이 현재 의료영상분석, 생물정보학, 기계학습 등의 분야에서 활발하게 이루어지고 있다. 기계학습의 대표적 응용분야인 컴퓨터비전에서는 실시간으로 현상(scene)을 라이다, 레이다, 초음파센서와 카메라로부터 읽어들여, 그 다출처 자료를 이용하여 주변환경의 지도를 만든다. 생물정보학에서는 한 환자에 대해 대량의 omics 자료와 혈청학적 자료 (gene expression, copy number variation, methylation 등)를 이용하여 병의 유무와 관련된 예측모형을 만드는 것이 주요 목표 중 하나이다. 다출처 자료 연계의 방법론은 지난 수년 동안 급격히 발전해 왔으나, 개발된 방법론의 이론적 근거가 부족하고, 성능평가의 이론적 지표가 개발되지 않아 휴리스틱 또는 실증적 평가에 기대고 있는 실정이다. 대규모의 자료가 확보 가능할 때는 많은 데이터에 기반한 실증적 성능평가가 가능하지만, 생물정보학 또는 의료영상 분야와 같이 환자수 ($n \simeq$ 몇백)보다 측정된 특성의 개수 ($p \simeq$ 몇천~몇십만)가 많은 상황에서는 실증적인 성능평가가 매우 비정확하다. 데이터의 차원 p가 데이터의 개수 n 보다 월등히 높을 때, 차원의 저주를 피하기 위한 엄밀한 이론과 방법론의 개발이 절실하지만, 다출처 자료 연계와 관련된 이론 연구는 아직 초기 단계이다.

본 연구에서는 그동안의 다출처자료연계방법론에서 경시되었던 통계적 이론의 정립을 최종목표로 삼는다. 여기서, 통계적 이론이란 자료가 관측된 기저시스템을 몇 개의 특성과 오차로 모형화하여 시스템에 대한 추론을 엄격하게 하는 방법론을 의미한다. 특히, 고차원-저표본 ($p \gg n$) 상황에서 차원의 저주를 피하기 위해 고안된 스파이크 모형-희소적 신호모형과 그에 따른 모형적합방법을 개발하고, 모형과 추정값의 평가를 위해 HDLSS, Random Matrix Theory (RMT), Tail bound 등 최근 개발된 고차원통계이론을 적극 활용한다. 본 연구사업의 완성으로 국제적인 다출처자료연계이론 개발을 선도하고, 데이터의 형태다변화 등 미래의 통계학적 문제 해결에 지표가 될 것으로 기대한다.

연구내용

연구의 내용를 크게 고차원-저표본 다출처자료 연계 분석에서의 (1) 확률이론 연구와 그에 따르는 (2) 방법론 개발, (3) 방법론의 응용 및 (4) 데이터형태다변화에 대응하는 연구주제 발굴의 네 가지 소주제로 나누어 설명한다.

1. 확률이론: 다출처자료 분석의 기본이 될 선형인자모형을 먼저 통계적으로 분석한다. 선형인자모형은 다변량자료를 몇 개의 인자(factor)를 이용하여 설명하는 모형으로 기저시스템을 소량의 인자만으로 단순화하여 설명하는 통계방법론의 하나이며, 다출처자료로의 확장도 많이 연구가 되었다. 다출처자료에서는 한 인자가 모든 출처의 자료를 설명하거나 또는 일부의 자료를 설명할 수도 있기 때문에 모형 identification의 문제가 발생한다. 예를 들어, 카메라에는 보이지 않는 사람이 초음파센서에는 감지될 수 있는데, 이를 카메라와 초음파센서 모두에 있다고 가정하면 잘못 identified된 모형이다. 이를 수학적으로 엄밀하게 풀어낼 착안점은 모형

identification의 문제를 예측된 인자가 span하는 subspace (인자공간) 간의 거리 (canonical angle)에 관한 문제로 바꿀 수 있다는 것이다. 따라서, 고차원-저표본 다출처 자료에서 계산될 Canonical angle의 일반적인 확률분포를 도출하는 것이 첫 번째 핵심적인 연구목표이다. 인자공간은 n차원 공간 내에서 p개의 변량들의 (선형)결합으로 이루어지므로, 모형파라미터의 적절한 선택으로 그 변동을 제어할 수 있을 것으로 예측되며, 인자공간 간 거리의 확률분포가 degenerate하지 않는 파라미터의 집합을 찾는 것이 주요 도전과제가 되겠다. 본 연구에서는 이와 같은 접근방법이 선형인자모형에서 유효한지를 먼저 살펴보고, 이를 바탕으로 일반화 선형인자모형, Tensor 분해 모형, 다출처자료를 이용한 회귀(예측)모형으로 확장할 것이다.

2. 방법론 개발: 다출처 선형인자모형에서 canonical angle의 분포가 이론적으로 도출되면, 먼저 모형 identification의 문제를 통계적가설검정의 관점으로 평가할 수 있다. 특히, 기존의 추정방법과 새로 개발될 방법이 model identification consistency (모형선택일치성)을 가지고 있는지판단할 수 있게 된다. 고전적인 통계학에서 '일치성'이란 자료의 수(n)이 증가할 때 추정된 모형이 실제 모형에 매우 가깝게 되는 성질을 말하지만, 고차원-저표본 ($p \gg n$) 상황에서는 일치성을 보이기가 매우 까다롭다. 본 연구자는 출처가 하나인 상황에서 다양한 다변량분석법이 일치성을 가질 조건을 RMT, maximal tail bound등을 이용하여 도출한 바 있기 때문에, 같은 수학적도구를 다출처자료분석에도 시도할 것이다. 또한, canonical angle의 분포를 이용한 overlapping hierarchical clustering를 이용하여 모형을 추정하는 새로운 관점으로 방법론을 개발한다. 이때 새로 개발될 방법이 모형선택일치성을 가지도록 하는 것이 주안점이다.

3. 응용문제 해결: Alzheimer's Disease Neuroimaging Initiative (ADNI), Genotype-Tissue Expression (GTEx) project, the Cancer Genome Atlas (TCGA) project, Human Microbiome Project 등의 다양한 고차원-저표본 다출처자료에 위에서 개발된 방법론을 적용하여 분석할 예정이다. 실제 자료에 방법론을 적용할 때는 대용량자료, 대량결측 (massive group-wise missing values) 또는 compositional restriction (microbiome 자료의 경우) 등의 새로운 문제가 수반되므로, 이러한 문제를 해결하기 위해 효율적인 계산알고리즘개발, 공통인자를 이용한 결측치 imputation 등의 연구를 진행한다.

4. 데이터 형태 다변화 대처: 현대 통계학의 주요 과제 중 하나는 형태가 다변화된 데이터로부터 기저의 시스템을 추론하는 것이다. 일반적인 숫자들로 이루어진 데이터가 아닌 비유클리드 공간의 값을 가지는 데이터(함수, 구성비, 3차원모형, 형상, 확산텐서 등)를 연계할 필요성이 앞으로 더 커질 것으로 예측된다. 따라서, 형태가 다양한 데이터셋들간의 연계를 위한 통합된 방법론과 이론이 필요하다. 본 연구에서는 이 큰 주제의 사전연구로서, 데이터가 범주형 (yes/no 또는 Type I/II/III/IV 등)인 경우의 연계 일반화선형인자모형, 데이터가 리만 매니폴드 (manifold)의 값을 가질 때의 연계 일반화선형인자모형, 구성비데이터(compositional data)일 때의 연계선형모형, 구성비데이터의 행과 열이 모두 같은 유닛인 경우의 구성비 텐서 데이터일 때의 연계선형모형 등을 제시한다. 각각의 경우에 해당하는 모형을 따로 개발한 뒤, 모든 경우를 아우를 수 있는 통합적인 모형화 방법, 추정 방법이 있는지를 살펴볼 것이다. 이 연구의 결과는 형태가 다양한 데이터 간의 공통 인자를 찾는 일반론의 개발에 지대한 영향을 미칠 것으로 기대한다.

연구인력	연구기간	총 연구비
총 0명 (교수 0명/연구원 0명)	'00년 00 ~ '00년 00 (00개월)	백만원

Deep generative model의 통계적 성질 연구
(Statistical perspective of deep generative models)

연구의 필요성

현재 4차 산업혁명과 인공지능 혁신을 이끄는 핵심 기술은 딥러닝이라고 해도 과언이 아니다. 딥러닝은 과학 및 산업 전반에 걸쳐 거대한 영향을 미치고 있지만, 다른 방법에 비해 딥러닝이 왜 뛰어난 성능을 보이는지에 대한 이론적 연구는 많은 부분이 미지의 영역으로 남아 있다. 최근 2-3년 사이에 딥러닝의 근사 이론에서 괄목할만한 성과가 나왔는데, 그 중 핵심은 DNN(deep neural network)이 적은 수의 모수로도 임의의 부드러운 함수를 잘 근사할 수 있다는 것이다. 본 연구에서는 이러한 근사 이론을 활용하여 자료의 분포를 추정하는 DGM(deep generative model)의 통계적 성질, 특히 딥러닝이 어떠한 분포를 효율적으로 추정할 수 있는지를 연구하고자 한다. 딥러닝의 분포 추정 이론에 관한 연구는 최근 활발히 이루어지고 있지만, 대부분이 고전적인 비모수 추론 관점에 머물러 있다. 하지만 실제 딥러닝의 적용 분야는 이미지나 음성 등의 초고차원 자료이기 때문에, 이러한 이론으로는 초고차원 자료에서 뛰어난 성능을 보이는 딥러닝을 설명하기에 한계가 있다. 본 연구의 핵심 주제는 다양체(manifold) 이론을 활용하여 딥러닝이 어떻게 효과적으로 차원의 저주(curse of dimensionality)를 피하는지를 설명하려는 것이다. 이를 통해 딥러닝에 대한 우리의 이해도를 크게 증진시킬뿐만 아니라 인공지능 분야의 연구 방향에 대한 새로운 이정표를 제시할 수 있을 것으로 기대한다.

연구내용

다양한 통계적 방법의 성능을 이론적으로 평가하는 좋은 방법 중 하나는 추정량의 수렴속도(convergence rate)를 비교하는 것이다. 서로 독립이고 같은 분포를 갖는 D차원 자료 $X_1, \dots, X_n$가 주어졌을 때 X_i의 분포를 P_0, Lebesgue 밀도함수를 p_0라고 하자. 고전적인 비모수 방법인 커널이나 웨이블릿 등을 활용하면 p_0가 β-Hölder 공간에 속하는 경우 $\| \hat{p} - p_0 \| \leq n^{-\beta/(2\beta+D)}$를 만족하는 추정치 $\hat{p}$을 찾을 수 있다. 최근 연구에 따르면 GAN(generative adversarial network)과 같은 DGM 기반의 방법을 쓰더라도 유사한 수렴속도를 얻을 수 있다. 이러한 관점에서 보면 딥러닝과 고전적 비모수 방법론의 성능은 비슷하지만, 위 수렴속도는 자료의 차원인 D가 커지면 급격히 떨어지기 때문에 고차원 자료에 대한 이론으로 적절하지 않다. 따라서 밀도함수 공간에 Hölder 연속성 이외의 추가적인 구조를 부여하지 않고서는 딥러닝의 뛰어난 성능을 설명할 수 없다. 그렇다면 딥러닝이 잘 추정하는 분포는 어떤 구조를 갖고 있을까?

DGM 기반의 방법은 자료의 분포를 $f(Z)$의 분포라 가정한 뒤 DNN 모형을 통해 함수 $f: \mathbb{R}^d \to \mathbb{R}^D$를 추정한다. 여기서 Z는 균일분포나 정규분포처럼 간단한 분포를 따르는 d차원 확률벡터이다. 대부분의 경우 d는 D보다 훨씬 작기 때문에 $f(Z)$는 $\mathbb{R}^D$ 상의 저차원 다양체에 분포되어 있다. 따라서 이러한 모형이 잘 추정할 수 있는 분포, 즉 P_0에 대한 적절한 가정은 분포가 $\mathbb{R}^D$ 상의 저차원 다양체 주변에 집중되어 있다는 것이다. 본 연구에서는 이러한 가정을 수학적으로 $X_i = Y_i + \epsilon_i$로 표현할텐데, 여기서 Y_i와 ϵ_i는 독립인 확률벡터, Y_i의 지지집합(support)은 저차원 다양체 $\mathcal{M}$, $\epsilon_i \sim N(0, \sigma^2 I_D)$ (I_D는 D차원 항등행렬), 그리고 오차분산 σ^2는 매우 작다고 가정할 것이다. 그림 1은 2차원 공간 상의 1차원 다양체인 직선에 집중된 자료의 예시를 보여준다. 본 연구에서는 자료의 분포가 이러한 저차원 구조를 가질 때 딥러닝이 다른 방법에 비해 뛰어난 성능을 보인다는 것을 이론적으로 보이고자 한다. 구체적으로는 minimax

optimal 수렴속도를 규명하고 딥러닝 기반의 추정량이 이 속도로 수렴하지만 다른 비모수 방법은 훨씬 느린 속도로 수렴한다는 것을 증명할 것이다. 특히 minimax optimal 수렴속도가 주변(ambient) 차원 D가 아닌 다양체 차원 d에 의존한다는 것이 핵심이다. 본 연구에서 파생되는 다양한 문제 중 중요한 것들을 추리면 다음과 같다.

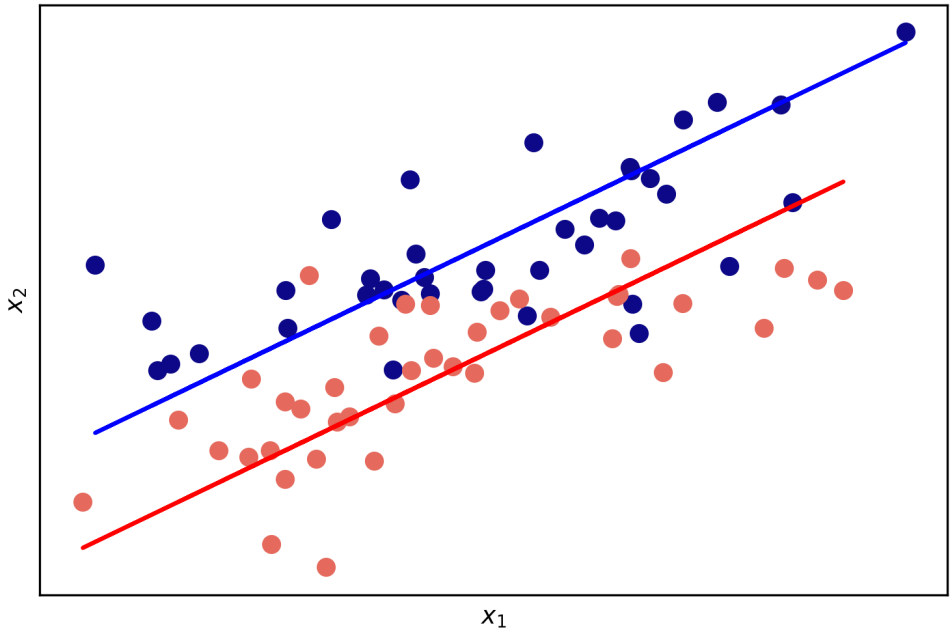
첫째는 밀도함수와 모형의 정규성(regularity)에 관한 것이다. 일반적으로 밀도함수의 정규성은 Hölder 연속성 등을 통해 표현하지만 P_0가 저차원 다양체 $\mathcal{M}$ 주변에 집중되어 있기 때문에 그림 2처럼 p_0는 $\mathcal{M}$ 주변에서 매우 뾰족한 형태의 함수이다. 따라서, p_0를 $\mathbb{R}^D$ 상에서의 부드러운 함수로 보기에는 무리가 있다. 한편, q_0를 Y_i의 분포인 Q_0의 밀도함수라고 했을 때 q_0가 부드러운 함수라는 가정은 매우 자연스럽다. 단, q_0는 Lebesgue 측도가 아닌 Hausdorff 측도에 대한 밀도함수라는 것이 중요하다. 본 연구에서는 자료 분포의 정규성을 p_0가 아닌 q_0의 부드러운 정도로 표현할 것이며 이는 기존 연구와의 가장 큰 차이점이다. 한편 DGM에서는 Q_0를 $\hat{f}(Z)$의 분포로 추정하는데 $\hat{f}$이 DNN 형식을 취할 때 실제로 Q_0를 잘 근사할 수 있는지 여부가 DGM의 성능을 결정하는데 큰 역할을 할 것이다. 이를 보이기 위해 최근 정립된 DNN의 근사 이론과 Caffarelli의 optimal transport 정규성 이론을 활용하고자 한다.

둘째는 자료의 분포 P_0 또는 밀도함수 p_0의 추정에 관한 것이다. 분산 σ^2가 작은 경우 그림 2와 같이 실제로는 $f(Z)$의 분포가 P_0에 가깝다 할지라도, $f(Z)$의 밀도함수는 p_0와 거리가 멀 수도 있다. 따라서 통상적으로 쓰이는 L^1이나 Hellinger 거리를 사용하게 되면 밀도함수 추정량이 p_0로 수렴하지 않을 수조차 있다. 본 연구에서는 우선 σ가 어느 정도로 작을 때까지 밀도함수가 L^1-거리 관점에서 일치성(consistency)을 갖도록 추정 가능한지 조사하고자 한다. 만약 σ가 그 경계보다 작은 경우, L^1-거리는 적절하지 못한 측도이기 때문에 위상적으로 더 약한(weaker) Wasserstein 거리를 사용하여 추정량의 성능을 측정할 것이다. 본 연구에서 사용하는 Wasserstein 거리와 자료 분포에 대한 저차원 기하학적 가정은 고전적인 비모수 밀도함수추정 연구에서 다룬 적이 없기 때문에, 새로운 통계 이론 개발이 필요할 것으로 보인다.

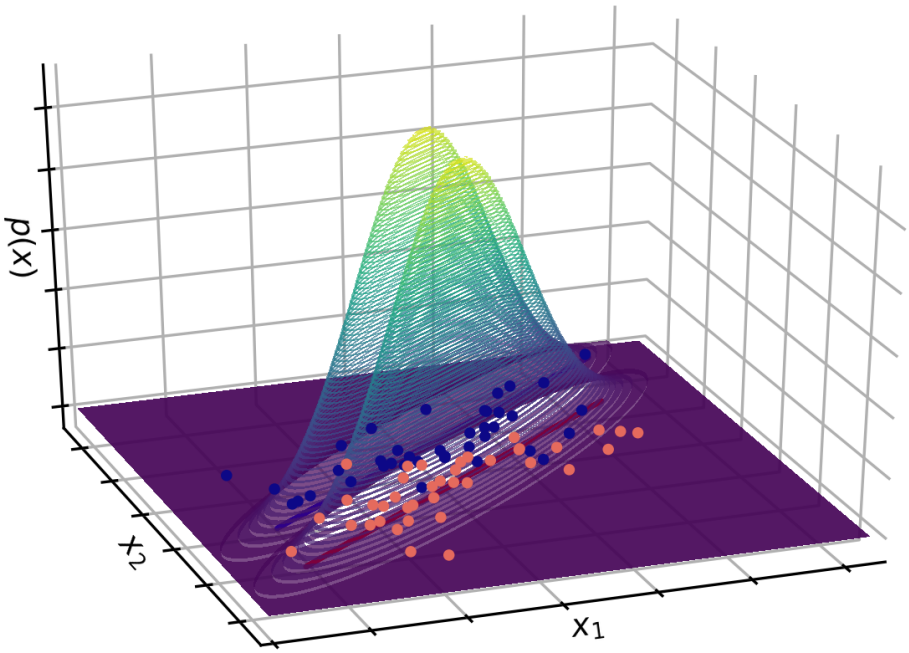
셋째는 Y_i의 분포인 Q_0를 추정하는 문제이다. 이는 자료의 분포 P_0에서 오차 ϵ_i를 제거하는 일종의 deconvolution 문제로 볼 수 있다. 일반적인 정규분포 deconvolution 문제에서의 minimax 수렴속도는 $(\log n)^{-1}$로 매우 느리지만 본 연구에서 고려하는 문제는 σ가 매우 작기 때문에 훨씬 빠른 수렴속도를 얻을 수 있을 것으로 예상된다. 여기서 가장 중요한 점은 Q_0를 추정하기 위해 추가적인 identifiability 조건이 반드시 필요하다는 것이다. 본 연구에서는 곡선의 곡률과 유사한 개념인 다양체의 reach를 통해 이 조건을 수학적으로 표현하고자 한다.

넷째는 다양체 $\mathcal{M}$의 추정에 관한 것이다. 실제 자료 분석에서 다양체 자체를 추정하는 것이 중요한 문제이지만 이에 대한 이론적 연구는 많지 않다. 본 연구에서는 DGM을 활용하여 다양체를 추정하고, 다양체 사이의 Hausdorff 거리를 통해 그 성능을 평가할 것이다. 이를 위해 적절한 정규성 가정 하에서 Hausdorff 거리에 대한 minimax 수렴속도를 규명하는 것이 필요하고, 또한 DGM 기반의 방법이 최적의 속도로 수렴하는지 여부를 조사하는 것이 필요하다.

연구인력	연구기간	총 연구비
총 0명 (교수 0명/연구원 0명)	'00년 00 ~ '00년 00 (00개월)	백만원



[그림 1] 2 차원 공간 상의 1 차원 다양체 주위에 집중되어 있는 두 자료. 두 직선 사이의 거리는 가깝기 때문에 두 자료의 분포는 비슷하다.



[그림 2] 위 자료의 밀도함수. 두 분포의 Wasserstein 거리는 가깝지만, 밀도함수의 L^1-거리로 보면 두 분포는 완전히 떨어져 있다.