Hole or Grain? A Section Pursuit Index for Finding Hidden Structure in Multiple Dimensions 요약

2022.04.13. 예지혜

[Abstract]
다변량 데이터를 이해하기 위해 주로 선형 투영이 많이 사용되지만, 선형 투영은 분포의 중심 근처에 있는 패턴들을 잡아내지 못하기도 한다. 이러한 문제를 해결하기 위해 Section (또는 Slice)가 도움이 된다. 이 논문은 PP 방법론을 이용한 section pursuit 방법을 통해 데이터의 흥미로운 슬라이스들을 찾아낸다. 섹션의 내부와 외부의 분포를 비교하는 인덱스를 이용해 holes와 grains를 찾아낼 것이다. 이러한 방법은 데이터 분포가 유니폼, 정규분포가 아닐 때 유용하다.

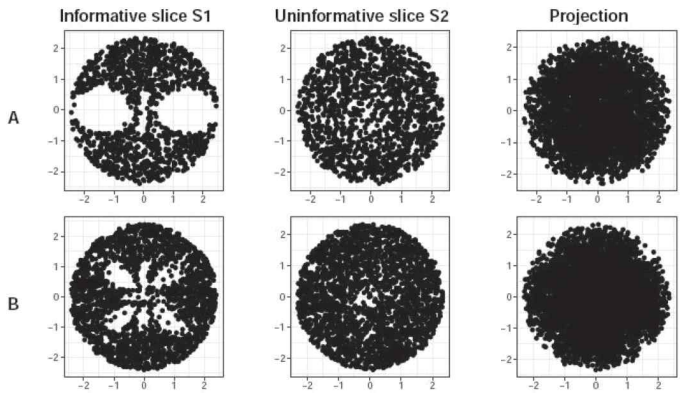
1. Introduction

프로젝션은 데이터의 중심에 위치한 패턴을 못 찾아내기 쉽고, 이를 해결하기 위해 비선형 매핑이나 다차원 스케일링이 사용되나 해석하기가 어렵다. Slice tour는 투영 대신 다차원의 section을 통해 숨겨진 구조를 찾을 수 있다. Slice는 슬라이스 내부, 외부의 데이터 분포 비교를 가능하게 하며, 여기서는 데이터의 가장 흥미로운 슬라이스를 위한 projection space를 찾아내는 section pursuit을 제안할 것이다. 여기서 흥미롭다는 것은 슬라이스 내부, 외부의 비유사성이 가장 큰 것을 의미한다.

2. Background

2.1. Grand Tour and Slice Tour

grand tour란 계속 움직이면서 랜덤하게 선택된 프로젝션의 변화를 확인하는 것이다. 연속적으로 보여지는 저차원의 관점을 통해 고차원을 추론할 수 있게 한다. 이를 통해 얻어진 정보들은 sectioning을 통해 보완될 수 있는데, 그 중 하나로 슬라이싱을 이용한다. 각 데이터 포인트로부터 projetion plane 까지의 직선 거리가 cutoff h보다 짧으면 슬라이스에 들어가는 포인트로 간주한다. 일반적으로 projection plane은 데이터의 중심을 지나야 하며, 거리 척도로는 유클리디안 거리를 이용한다.



다음 그림은 4차원 데이터의 산점도인데 informative slice를 통해 holes를 찾을 수 있다. 다만 프로젝션에서는 이러한 패턴들이 다 가려진 것을 확인할 수 있다.

2.2 Projection Pursuit and Guided Tour

PP와 Grand tour의 개념을 결합한 것이 Guided Tour이다. PP index를 최적화하며 projection plane을 선택하는 것이라 볼 수 있다. 또한 PP index의 변화를 이해하기 위해 tour 방법을 사용할 수도 있다.

3. A New Index for Finding Interesting Sections

슬라이스 내부, 외부의 분포를 비교하는 새로운 인덱스 함수를 정의할 것이다.

3.1. Taking a Slice and Binning

$Y = XA$, X는 n by p, A는 p by d로, p차원의 X를 A가 d차원의 Y로 축소시킨다고 이해할 수 있다. 이때 A를 2차원으로 두고 2차원 슬라이스를 만든다고 보자.

각 데이터 포인트로부터 A 평면까지의 유클리디안 거리는 다음과 같다.

$$h_i = ||\mathbf{x}_i - (\mathbf{x}_i \cdot \mathbf{a}_1)\mathbf{a}_1 - (\mathbf{x}_i \cdot \mathbf{a}_2)\mathbf{a}_2||.$$

이 거리가 cutoff h보다 짧으면 슬라이스 S에 들어가고, 그렇지 않으면 슬라이스 C에 들어간다고 정의한다. (C는 complement of S) 추가로 K개의 bin을 정의하여 투영된 데이터 Y는 k번째 bin에 속한다고 할 것이다. 그러면 데이터 포인트는 다음 중 하나에 들어간다.

$$S_k = \sum_i I(Y_i \in b_k)\, I(h_i < h), \;\; C_k = \sum_i I(Y_i \in b_k)\, I(h_i \geq h)$$

즉, binning과 slicing 두 과정을 거친다고 볼 수 있다.

3.2. Index Definition

슬라이스 내부와 외부의 분포를 비교하는 인덱스를 다음과 같이 정의하자.

$$I_A^{\text{low}} = \sum_k [(c_k - s_k)]_{>\varepsilon}, \;\; I_A^{\text{up}} = \sum_k [(s_k - c_k)]_{>\varepsilon}$$

s_k는 슬라이스 내부, c_k는 슬라이스 외부의 관측치의 상대적 개수이다. I_A^{low}가 크다는 것은 외부의 데이터가 많다는 뜻으로 hole 패턴을 잡아내고, I_A^{up}이 크다는 것은 내부의 데이터가 많은 것이므로 grain 패턴을 잡아낸다. ϵ은 노이즈를 피하고 bin 개수에 의한 인위적인 영향을 피할 수 있게 한다.

3.2.1. Generalized Index

지금까지 정의된 인덱스는 특정 bin을 강조하거나 민감도를 조정하기 위해 다음과 같이 일반화될 수 있다.

$$I_A^{\text{low}} = \sum_k w_k \left(\left[c_k^{1/q} - s_k^{1/q} \right]_{>\varepsilon} \right)^q \qquad I_A^{\text{up}} = \sum_k w_k \left(\left[s_k^{1/q} - c_k^{1/q} \right]_{>\varepsilon} \right)^q.$$

w는 특정 bin을 강조하거나 무시하기 위한 장치이고, q는 얼마나 민감하게 잡아낼지 조절하는 역할을 한다. q가 작으면 작은 차이도 잡아내고, q가 커지면 큰 패턴을 더 잘 잡아낸다.

4. Practical Issues <<전반적으로 내용이 어려움>>

4.1. Rotation Invariance

PP 인덱스를 비롯해, section pursuit에서 인덱스는 회전에 불변해야 한다. Regardless of the basis in the d-D plane defining the projection, the index value should be identical.
따라서, 슬라이싱에서 분포 비교 기준으로 spherically symmetric 분포를 이용한다. (여기서는 유니폼 분포를 사용하며, 다변량 정규분포도 가능하다.) 보통 데이터는 하이퍼큐브로 관측되나, 주 관심사가 데이터의 중심에 있으므로 가장자리는 shaving off 해서 사용한다. 또한, 일반적으로 프로젝션된 형태가 타원형 또는 구 형태를 띄기 때문에 큰 문제되지 않는다.

4.2. Polar Binning and Reweighting

Polar coordinates에 binning하는 것이 더 좋은 접근이며 radial, directional 2가지 요소로 분해할 수 있다. 여기서는 2차원이기 때문에 모든 θ에 대해 유니폼 hypersphere 분포이므로 K_θ equidistant angular bins를 이용할 수 있다.
radial binning은 p가 증가할수록 프로젝션이 중앙에 몰리기 때문에 좀 더 어렵다.
However, the expected radial distribution will also differ between the projected data in a slice and a projection of the full data, so the effect is accounted for instead by reweighting bin counts. The aim is to adjust for the expected difference in distribution: after the reweighting, the expected counts in all radial bins is the same, and this needs to be adjusted separately for the data in the slice and the projected data.
reweight을 통해 조절하는데, 프로젝션 데이터의 경우에는 cdf를 이용하고, 슬라이스 데이터는 h<<R에 대해 거의 유니폼 분포이다...

4.3. Sufficient Sample Size

차원이 증가할수록 데이터의 얇은 슬라이스로 피쳐를 해결하기 위해 필요한 데이터의 수가 엄청나게 증가한다. 필요한 충분한 데이터 크기에 대한 내용

4.4. Estimating the Magnitude of Noise, ϵ

bin에 따라 적절한 ϵ를 통해 변동성을 잡아주면 bin의 개수에 상관없이 일치하는 인덱스 값을 확인할 수 있다.

4.4.1 Dependence on q

q 조절을 통해 작은 차이에 대한 민감성을 조절할 수 있는데, ϵ를 결정할 때도 함께 고려될 수 있다.

4.5. Index Behavior

reparameterization?
why empty bins inside the slice?

4.6. Visualising the Index

1차원 PP 인덱스를 시각화해주는 Huber plot 아이디어를 이용해 2차원 PP 인덱스도 시각화할 수 있다. (topotrace plot)

1) starting plane A_0를 고른다.

2) A_0로부터 충분히 큰 m개의 방향을 랜덤하게 발생시킨다. $A_i, i = 1,...,m$

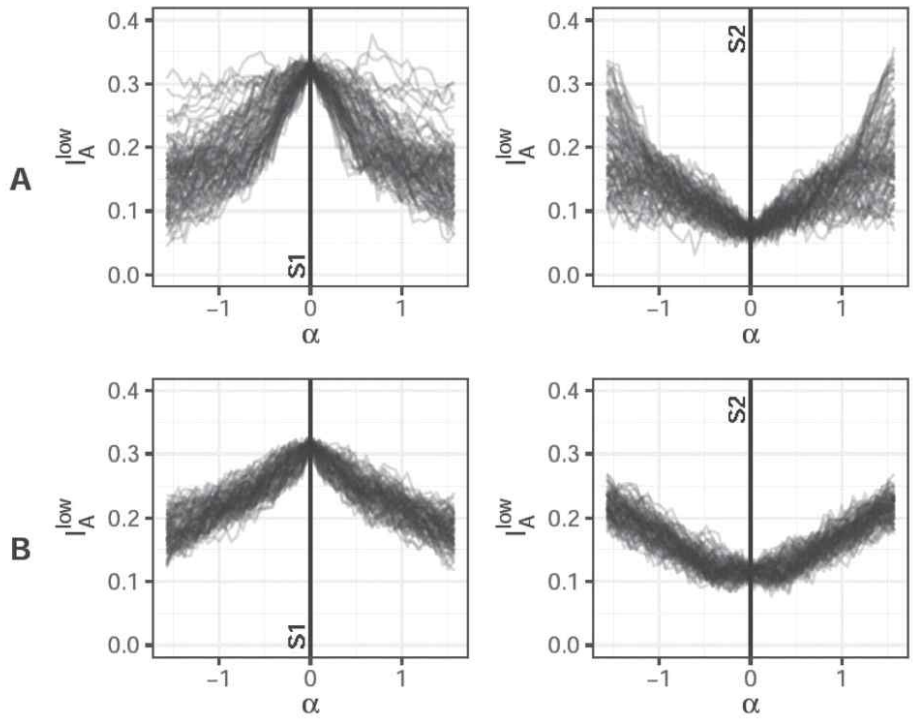
3) 각 방향에서 고정된 길이 또는 각도로 geodesic interpolation을 발생시킨다.

$A_{ij}, i = 1,...,m, j = -\alpha,...,0,...,\alpha$

4) 각 A_{ij}의 인덱스를 계산한다.

5) j에 대한 인덱스 $I_{A_{ij}}$를 개별 i에 대해 시각화한다.

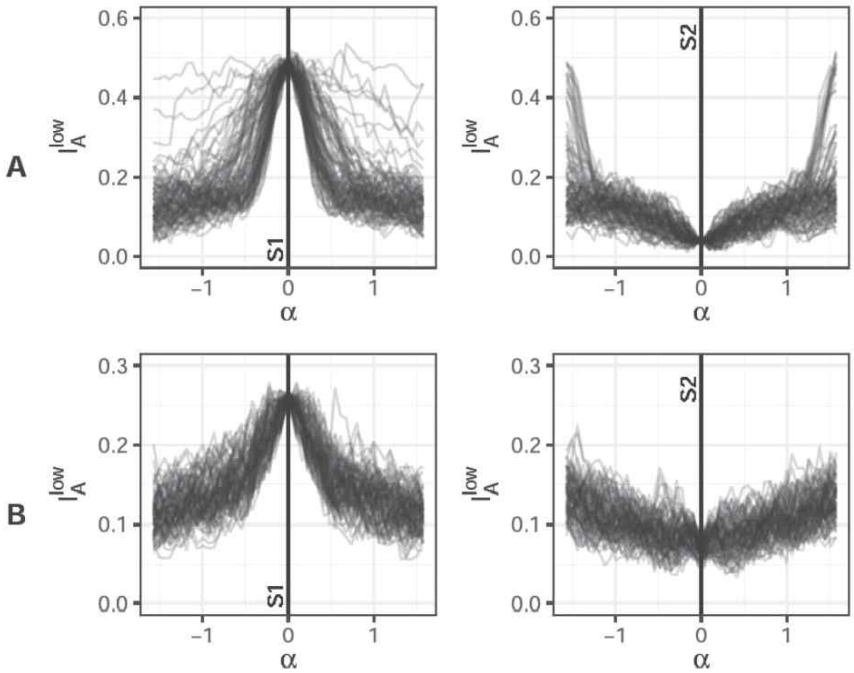
이 plot을 통해 로컬 maximum이나 ridge, smoothness, squint angle 같은 요소들을 확인할 수 있다. squint angle이 크면 최적화하기 쉬워진다. 최적화된 투영에서는 마치 산 꼭대기에서 모든 방향을 보는 듯한 효과를 낼 수 있다.



A, B 두 데이터셋에 대한 topotrace plot이다. informative slice S1에서 최댓값을 확인할 수 있다. A 데이터셋은 일부 trace들이 다양한 각도에서 높은 값을 유지하고 있다.

This suggests that the function has ridges. This is expected for this data, as a result of a symmetry in the simulation distribution, the structure remains visible so long as the first variable is dominant along one direction in the plane. This also makes the structure easier to detect, and we find that among the 100 random directions, several traces

reach index values close to that of the ideal view.



q=2일 때의 그림인데, squint angle이 더 작고 S1 슬라이스 근처에서 인덱스가 급격히 변화하는 것을 확인할 수 있다.

4.7. Index in Practice

1) 데이터가 하이퍼스피어 안에 있도록 centering과 scaling을 진행하고, 최대 r 밖에 존재하는 데이터포인트들은 삭제
2) 데이터 사이즈와 슬라이스 두께에 따라 적절한 bin size를 선택
3) 적절한 슬라이스 radius h를 선택
4) 앞서 주어진 식을 통해 ϵ 계산
5) 각 슬라이스에 대한 인덱스 계산
6) guided tour의 수정된 버전을 최적화에 사용

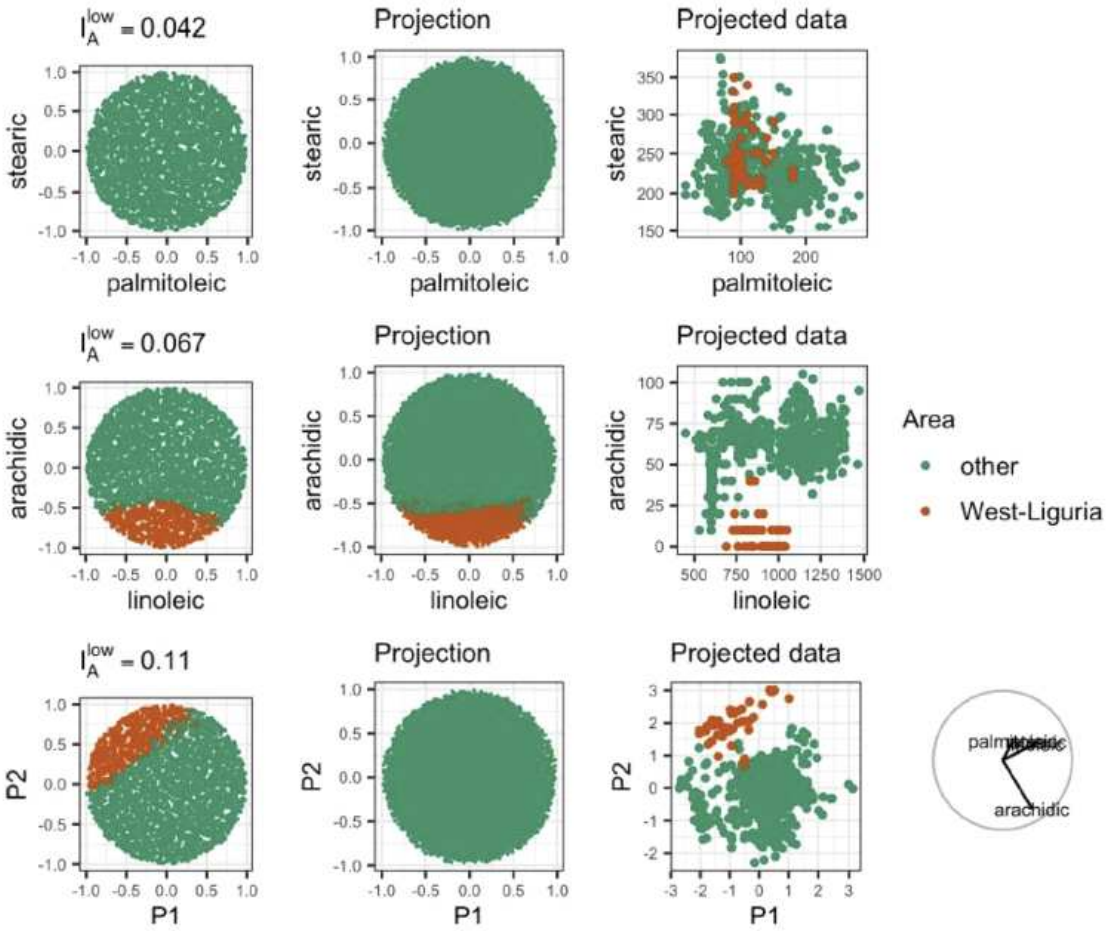
5. Applications

5.1. Index Settings

5.2. Classification Boundaries

분류 결정경계를 찾기 위해 예측하고자 하는 샘플을 모두 제거한 후 section pursuit을 진행할 수 있다. 그 결과 찾아낸 슬라이스를 각 분류에 할당된 데이터 포인트들과 함께 확인하여 결정경계를 찾는 것이다.

5.2.1. Olives data



West-Liguria을 분류하기 위해 이에 해당하는 데이터 포인트들을 모두 제거함으로써 빈 공간을 만들어 section pursuit을 진행한다.

위의 두 행은 설명 변수 2개를 축으로 이용한 것이고, 마지막 행은 section pursuit을 진행한 것이다. 첫 번째 행은 WL을 구분해내지 못하나, 두 번째 행은 비선형 결정경계로 분류 가능하다. 마지막 행은 선형 결정경계로 분류 가능하다. 신기한 것은 프로젝션은 어떠한 결정경계도 주지 않으나 실제로 투영된 데이터는 선형 결정경계로 분류가 가능하다는 것이다. (각 열에서 나타내고 있는 것이 무엇인지 정확히 모르겠음. The model predictions are shown in slices and projections in the first two columns of Figure 11, the last column shows the projected data.)

5.3. Inequality Condition

물리학적 실험을 위한 세팅인 거 같은데 전반적으로 이해 안 됨.