[3주차] Projection Pursuit 요약

2022.02.03 예지혜

논문 1. A Projection Pursuit Algorithm for Exploratory Data Analysis

[Introduction]

다변량 데이터를 저차원으로 매핑하여 살펴보는 것은 자주 사용되는 기법이다. 비록 데이터의 모든 정보를 담고 있지 않더라도 시각화를 통해 인간의 눈으로 패턴을 인지할 수 있기 때문에 많은 방법이 연구되어 왔다.

세 가지 범주로 나누어 보면, 1) 선형 차원 축소 2) 전체 데이터의 비선형 차원 축소 3) 일부 주어진 점들만 비선형 차원 축소로 나눌 수 있다.

비선형 차원축소의 단점 (선형 차원축소의 장점):
- 해석이 어렵다.
- 파라미터가 매우 많아 분석에 이용된 데이터에서만 정의되는 경우가 있어 새로운 데이터를 기존 모델에 매핑할 수 없다.
- 계산 비용이 매우 많이 든다.

선형 차원축소의 단점:
전반적인 global 패턴을 잡다보니, 여러 local 패턴으로 형성된 global 패턴을 잘못 잡는 경우가 생긴다.

PP의 장점:
- global과 local 특성을 결합하여 유용한 선형 매핑을 찾아낸다.
- trimmed global measure를 사용하기 때문에 이상치에 민감하지 않다는 장점이 있다.

[Projection Pursuit]

PP는 최적의 투영을 찾기 위해 분산 뿐 아니라 데이터 간 거리(interpoint distance)를 사용하는 선형 매핑 알고리즘이다. 다차원의 공간에서 각 방향에 대해 "usefulness"를 나타내는 연속형의 인덱스를 찾는다. 이 인덱스는 sufficiently continuous라 최대화할 때 hill-climbing 알고리즘을 사용할 수 있어 계산 비용이 적게 든다.

PP: 메모리 O(N), cpu cycle O(NlogN)
비선형: 메모리 $O(N^2)$, cpu cycle $O(N^2)$

또한 선형 알고리즘이라 해석이 쉽고 새로운 데이터를 매핑할 수 있다.

Isolation과 결합하면, PP는 클러스터를 찾는 데에 유용한 방법이다. PP는 각 클러스터에 대해 개별적으로 적용될 수 있어 몇 번이고 데이터를 분리할 수 있고, 계산 비용 또한 경제적이다.

[The Projection Index]

PP 인덱스는 인간 연구자와 컴퓨터 간 상호작용으로 결정되었는데, 연구자가 데이터를 회전시키면서 변화하는 투영을 보면서 데이터의 특징을 파악할 수 있었다. 보통 클러스터 간 거리는 더 멀고, 클러스터 내부의 거리는 좁게 만드는 방향을 찾으려 했으며 인덱스 식은 이렇게 나타낼 수 있다. ($\hat{k}$: projection axis)

$$I(\hat{k}) = s(\hat{k})d(\hat{k})$$

$$s(\hat{k}) = \left[\sum_{i=pN}^{(1-p)N} (\boldsymbol{X}_i \cdot \hat{k} - \bar{X}_k)^2/(1-2p)N\right]^{1/2} \qquad d(\hat{k}) = \sum_{i=1}^{N}\sum_{j=1}^{N} f(r_{ij})\,1(R - r_{ij})$$

$$\bar{X}_k = \sum_{i=pN}^{(1-p)N} \boldsymbol{X}_i \cdot \hat{k}/(1-2p)N. \qquad r_{ij} = \left| \boldsymbol{X}_i \cdot \hat{k} - \boldsymbol{X}_j \cdot \hat{k} \right|$$

$s(\hat{k})$는 투영시킨 데이터의 trimmed sd로, 데이터의 퍼진 정도를 나타내고, $d(\hat{k})$는 nearness를 나타내는 함수의 평균치로, 데이터 간 로컬 밀집 정도를 나타낸다.

Cutoff radius R이 무엇인지.. d 함수가 이해 안 됨

2차원으로 투영시키면 두 개의 방향 $\hat{k}$와 $\hat{l}$을 찾게 되고 서로 직교한다. 식은 다음과 같다.

$$s(\hat{k},\hat{l}) = s(\hat{k})s(\hat{l}), \quad r_{ij} = \left[(\boldsymbol{X}_i \cdot \hat{k} - \boldsymbol{X}_j \cdot \hat{k})^2 + (\boldsymbol{X}_i \cdot \hat{l} - \boldsymbol{X}_j \cdot \hat{l})^2\right]^{1/2}$$

$$\bar{r} = \int_0^R rf(r)\,dr \Big/ \int_0^R f(r)\,dr \qquad \text{(one dimension)}$$

$$\bar{r} = \int_0^R rf(r)r\,dr \Big/ \int_0^R f(r)r\,dr \qquad \text{(two dimensions)}$$

characteristic width

알고리즘을 반복적으로 적용해보면서 이 방식이 함수 $f(r)$에 민감하기 보다는 characteristic width에 주로 의존하는 것을 알게 되었다. 이 값은 local density를 의미하며 투영된 부분 공간 간의 거리를 측정한다. (Its value establishes the distance in the projected subspace over which the local density is averaged, and thus establishes the scale of density variation to which the algorithm is sensitive.)

결국 인덱스 I는 로컬 집중도(d가 큰 것)와 글로벌 분산(s가 큰 것)을 동시에 측정하며, 실험을 진행하며 대부분의 연구자가 궁금해할 척도임을 확인할 수 있었다. 따라서 이 인덱스를 최대화하는 프로젝션을 찾는 것이 자연스러운 목표이다.

[One-dimensional PP]

n차원에서 1차원으로의 프로젝션은 그것의 방향을 결정하는 n-1개의 파라미터로 이루어진 함수이다. 각 방향의 코사인을 나타내며, n개의 코사인들의 제곱합이 1이 되도록 하는 제약조건을 사용한다. 이를 구하는 방법 중 하나로 Solid Angle Transform (SAT)이 있다. $I(\hat{k})$를 최대화하는 대신 동일한 함수 $I[SAT(\hat{k})]\ in\ E^{n-1}(-\infty,\infty)$를 찾아 최대화하는 방법으로, 제약조건이 사라져 더 안정적인 최적화를 할 수 있다. 똑같이 n-1개의 파라미터를 찾는다.

이러한 목적함수 최적화를 위해 사용되는 수렴 기준이 목적함수 평가에 영향을 많이 미치고, 이는 계산 비용에도 많은 영향을 미쳐 몇 개의 평가를 사용할 지는 매번 달라진다.

또한 다양한 starting direction으로부터 솔루션이 결정될 수 있고, (e.g. larger principal axes, 원점, 랜덤) 여러 개의 솔루션에 대해 데이터 해석의 유용성을 평가하여 답을 결정한다.

추가적인 다양한 답을 얻기 위해 탐색하는 공간의 차원을 줄이기도 한다. 작은 분산을 가진 방향들을 제거하기도 하고 이미 얻어진 방향 k들을 이용해 임의적으로 방향을 선택하기도 한다. (m개의 차원을 제거하여 n-m차원에서 시작)

[Two-dimensional PP]

- 앞서 정의한 식을 최적화
- k를 상수로 고정하고 I에 대해 최대화. k는 문제와 관련하여 직접 정해줄 수도 있고, 1차원 projection 결과를 사용할 수도 있다.
- 하나를 고정하고 다른 하나를 구하는 과정을 두 방향이 모두 수렴할 때까지 반복

[Some Experimental Results]

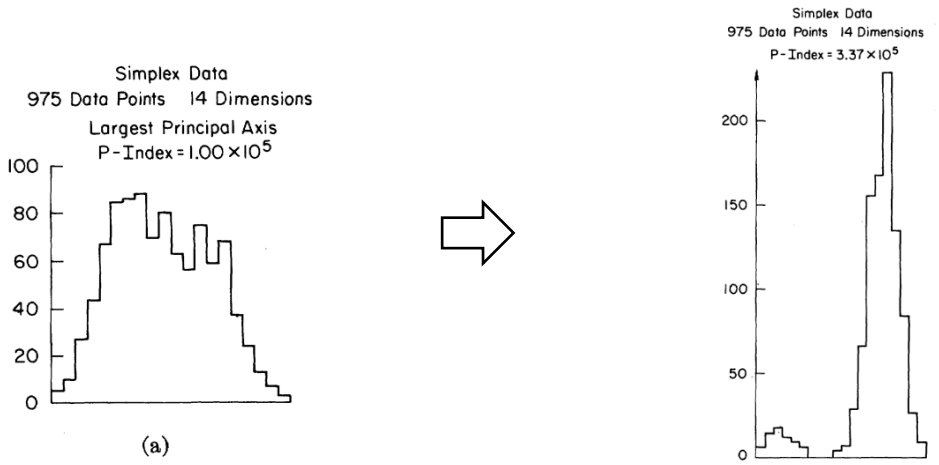
두 개의 인공 데이터와 iris data, physics 실험 데이터를 다루어 보았는데, 1차원 프로젝션에 대해서는 $f(r) = R - r$, 2차원 프로젝션에 대해서는 $f(r) = R^2 - r^2$을 사용하였다.

A. 유니폼 분포 랜덤 데이터

Preferred projection이 없는 인공 데이터로의 projection pursuit을 실험하기 위해 14차원의 유니폼 분포에서 975개의 데이터를 발생시켰다. 1차원 프로젝션에 대해 14개의 original axes와 14개의 principal axes를 starting direction으로 사용하였다.

결과는 이해 안 됨.

B. Gaussian data distributed at the Vertices of a simplex



이번에는 preferred projection이 있는 경우를 살펴보기 위해 15차원의 다면체의 꼭짓점에서 정규분포를 랜덤으로 발생시켜 975개의 데이터에 대해 PP를 적용하였다. 분산이 가장 큰 principal axis를 시작점으로 잡아 PP를 적용하니 P-index가 3배로 증가하며 데이터들이 확연히 나누어지는 것을 확인할 수 있다. 이렇게 나누어진 각 클러스터에 대해 PP를 반복 적용할 수 있으며 그때마다 클러스터가 잘 분리되었다.

[Discussion]

PP는 데이터에 구조가 있으면 잘 찾아내고, 없으면 잘 찾지 못하는 것으로 보아 잘 작동하는 것을 확인하였다. PP는 선형 매핑이기 때문에 curved surface는 잘 찾지 못하는 단점이 있었으나 measurement variable로 비선형 transformation을 사용하면 spherical 클러스터링을 찾아내기도 하였다.

PP는 분산을 최대화하면서도 클러스터 내의 거리를 줄이기 때문에 데이터 구조에 방해만 되는 무관한 프로젝션을 찾는 경우를 피할 수 있었다. 또한, characteristic radius r만 결정하면 되기 때문에 사전지식이 많이 필요하지 않다는 장점이 있다. 그 외에도 몇 차원으로 프로젝션 할 지 결정해야 하는데 2차원 프로젝션이 1차원보다는 불안정하지만 더 많은 정보를 가지고 있다. 보통 1차원 프로젝션 결과를 2차원 프로젝션의 시작 포인트로 사용하는 것이 좋은 전략이었다.

논문 2. Projection Pursuit Indexes Based on Orthonormal Function Expansions

[초록]

실험을 통해 orthogonal polynomial index인 Legendre index와 Hermite index의 차이를 살펴본다. 이 두 방법은 투영된 데이터의 분포와 정규분포 간의 weighted L2 거리를 나타낸다. 또한 이 두 방법의 conceptual problem을 완화한 방법인 Natural Hermite index도 살펴본다.

Polynomial expansion은 각 계수 자체를 하나의 인덱스로써 주목하게 하는데, 이 중 첫번째 두 개의 계수가 매우 유용한 정보로 알려져 있다. 이들은 데이터의 구조를 멀리서 크게 바라볼 수 있게 하며, 이들을 보완하며 내부 구조를 살펴보는 데에는 낮은 차수의 계수들이 사용된다.

[Introduction]

PP는 p차원의 데이터를 k차원(1또는 2, 많으면 3)으로 투영시킨 것의 "interestingness"를 평가하는 기준 함수(즉, 인덱스)를 정의하고 이것의 전역, 지역 최대값을 모두 찾는다.

nonnormal
- "interesting" 투영을 찾는 것은 가장 nonnormal한 투영을 찾는 것과 같다고 볼 수 있다. 대부분의 투영은 가우시안과 닮았기 때문에 특이한 투영을 찾으려면 정규성을 따르지 않는 것을 찾아야 하기 때문이다.
- Huber도 structured 또는 nonrandomness를 흥미로운 것이라 보고 엔트로피를 정의했는데, 엔트로피가 높으면 랜덤성도 높아진다. 그런데 엔트로피는 정규분포에서 가장 커지기 때문에, 엔트로피가 작은 것을 찾는다는 건 정규성을 보이지 않는 투영을 찾는다는 것과 같다.
- normality를 null model로 사용하는 것도 강조되는데, 이러한 접근은 위치, 척도, 공분산 등 전통적인 방법에서 조사되던 구조들을 버리는 것이라 볼 수 있다.

$let\ k = 1,\ Z \to X = \alpha' Z\ \in R\ (\alpha \in S^{p-1}),\ where\ S^{p-1}\ is\ a\ unit\ (p-1)\ sphere\ in\ R^p$
$in\ the\ null\ case,\ if\ Z \sim N(0, I_p),\ then\ X \sim N(0,1).$
인덱스 I는 X의 $density\ f(x)$와 표준 정규분포 간의 거리.

→ 적용: Friedman's proposed index "Legendre index"

$$I^L = \int_{-1}^{1} \left\{ g(y) - \frac{1}{2} \right\}^2 dy.$$

$Y = 2\Phi(X) - 1$ 변환을 통해 X를 [-1,1] 구간으로 매핑한다. 이를 통해 center로 집중하게 하여 꼬리 변동에 민감하게 반응하지 않도록 하였다.

$in\ the\ null\ case, if\ X \sim N(0,1), then\ Y \sim U(-1,1)$. 인덱스는 Y의 $density\ g(y)$와 $U(-1,1)$ 간의 L2 거리. 이후 $g(y)$는 natural polynomial basis로 확장된다.

Projection pursuit의 index I에 대해 짚고 넘어갈 것

1. I는 f에 대한 함수이다.
2. f는 투영 벡터 α에 의존하므로, PP는 모든 가능한 α에 대한 I의 local maxima를 찾는 과정을 포함한다.

[2. Transformation Approach]

$Y = T(X)$, $T : R \to R$ (즉, Y는 X의 transformation)

X의 분포함수 $F(x)$, 밀도함수 $f(x)$, Y의 분포함수 $G(y)$, 밀도함수 $g(y)$

$$I = \int_{\mathbb{R}} \{g(y) - \psi(y)\}^2 \psi(y) dy,$$

즉, 인덱스는 변형된 Y의 밀도함수와 null distribution 간의 거리를 의미한다.

alternative orthonormal basis 형태의 추정에 적합하도록 형태를 바꾸거나, 특정 구조에 대한 민감성을 조절하기 위해 transformation을 진행한다.

$$\begin{aligned} I &= \int_{\mathbb{R}} \left\{ \frac{f(x)}{T'(x)} - \frac{\phi(x)}{T'(x)} \right\}^2 \phi(x) dx \\ &= \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 \frac{\phi(x)}{T'(x)^2} dx. \end{aligned}$$

변형된 내용을 다시 돌려 X에 대한 함수로 표현하면 f(x)와 표준정규분포 간의 가중치 거리라고 이해할 수 있으며 여기서 가중치는 $\phi(x)/T'(x)^2$이다.

- Legendre Index (by Friedman)

$$I^L = \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 \frac{1}{2\phi(x)} dx,$$

아이러니하게도 Friedman은 꼬리 변동에 의한 영향을 줄이고자 제안되었는데, 이 인덱스의 가중치 $1/\phi(x)$는 오히려 꼬리 관측치에 더 가중치를 부여해 영향을 많이 받게 된다.

- Hermite Index (by Hall)

$$I^H = \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 dx.$$

이러한 문제를 해결하기 위해 Hall이 제안한 인덱스는 f(x)와 표준정규분포 간의 L2 거리를 측정

한다. 이 또한 결국 앞선 가중치 인덱스에서 가중치 $\phi(x)/T'(x)^2$이 1인 경우로 볼 수 있다.

- Natural Hermite Index

$$I^N = \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 \phi(x) dx.$$

이는 T(X)=X identity 변형을 이용한 인덱스로 Natural Hermite index라 한다.

We call this index the Natural Hermite index and Hall's index the Hermite index, because both use Hermite polynomials in the expansion of f(x), and I^N is "natural" because the distance from the normal density is taken with respect to Normal measure.

Transformation family of T를 다음과 같이 규정하면 앞서 살펴본 3개의 인덱스를 모두 표현할 수 있다.

$T_\sigma(X) = \sqrt{2\pi}\sigma(\Phi_\sigma(X) - 1/2)$
$T'_\sigma(0) = 1$ for all $\sigma > 0$. The limit for $\sigma \to \infty$ is $T_{\sigma\to\infty}(X) = X$.

$T_{\sigma=1}$	$T_{\sigma=\sqrt{2}}$	T_∞
Legendre index, I^L	Hermite index, I^H	Natural Hermite index, I^N

이 세 개의 인덱스는 꼬리 부분에 가중치를 어느 정도로 부여했는지에 따라 해석할 수 있는데, σ가 작을수록 꼬리의 가중치가 더 크다고 볼 수 있다.

This is more a conceptual stumbling block than a practical deficiency because the problem is somewhat moot for finite function expansions.