PP index 정리

2022.09.02. 예지혜

[1주차] 차원축소 리뷰 논문 2021-Review of Dimension Reduction Methods

※ Projection Pursuit (PP)
- 데이터 탐색에 주로 쓰이는 비지도, non-parametric, 선형 차원축소 방법론
- EDA에 많이 쓰임 :
  저차원의 선형 투영을 찾기
  흥미로운 패턴 탐색 (interestingness가 PP index)
- 인덱스에 따라 다양한 패턴 감지 :
  분류, 군집화, 분포 추정, 회귀
- 장점 :
  다른 패턴 적합에도 유연하게 적용할 수 있으며, 샘플에 없는 점도 매핑할 수 있어 투영 공간에서 새로운 예시를 보여줄 수 있다.
  지도 학습에도 적용이 되었다.
  비정규성에 초점을 둔 많은 PP 인덱스가 제안되었다. ex) Legendre index[1]), Hermite index, natural Hermite index, entropy index, moment index[2])
- 단점 :
  computational difficulty --> gradient 방법, Newton-Raphson method, GA, SA 등의 최적화 방법이 사용된다.

---

1) Friedman, J.H. and Tukey, J.W. (1974) A Projection Pursuit Algorithm for Explorratory Data Analysis. IEEE Transactions on Computers, C-23, 881-890. https://doi.org/10.1109/T-C.1974.224051
2) Jones, M.C. and Sibson, R. (1987) What Is Projection Pursuit? Journal of the Royal Statistical Society: Series A (General), 150, 1-37. https://doi.org/10.2307/2981662

[3주차]

논문 1. 1974-A Projection Pursuit Algorithm for Exploratory Data Analysis

※ 선형 차원축소의 장점
- 비선형 방법보다 해석이 쉽고, 계산 비용이 적다.
- 기존 데이터에 없던 새로운 데이터에 대해서도 매핑이 가능하다.

※ 선형 차원축소의 단점
- global 트렌드에만 집중하여, 그 안에 local한 클러스터들의 방향은 무시하고, global trend 방향으로만 축소하는 경향이 있다.

※ PP의 장점
- global과 local 특성을 결합하여 유용한 선형 매핑을 찾아낸다.[3] [4]
- trimmed global measure를 사용해 이상치에 민감하지 않다는 장점이 있다. (robustness)

[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 최대화에 hill-climbing 알고리즘을 사용할 수 있어 계산 비용이 적게 든다.
- **Isolation과 결합하면, PP는 클러스터를 찾는 데에 유용한 방법이다.**[5] 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 어떤 식으로 데이터의 특징을 파악하려 하는지를 인덱스에 반영하였다. 보통 클러스터 간 거리는 더 멀고, 클러스터 내부의 거리는 좁게 만드는 방향을 찾으려 했으며 인덱스 식은 다음과 같다. ($\hat{k}$: projection axis)

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

$$s(\hat{k}) = \left[\sum_{i=pN}^{(1-p)N} (X_i \cdot \hat{k} - \bar{X}_k)^2/(1-2p)N\right]^{1/2} \quad \text{where} \quad \bar{X}_k = \sum_{i=pN}^{(1-p)N} X_i \cdot \hat{k}/(1-2p)N.$$

---

3) J. B. Kruskal, "Toward a practical method which helps un‑cover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation," in Statistical Computation, R. C. Milton and J. A. Nelder, Ed. New York: Academic, 1969.
4) -, "Linear transformation of multivariate data to reveal clustering," in Multidimensional Scaling: Theory and Applica‑tion in the Behavorial Sciences, vol. 1, Theory. New York and London: Seminar Press, 1972.
5) R. L. Maltson and J. E. Dammann, "A technique for deter‑mining and coding subclasses in pattern recognition problems," IBM J., vol. 9, pp. 294-302, July 1965.

: trimmed standard deviation. k 방향으로 투영했을 때 데이터의 퍼진 정도.
(N이 데이터 개수, p가 양쪽으로 자를 비율. i는 pN부터 N-pN까지만 거침. Xik는 각 데이터가 k방향으로 투영되었을 때의 값. 따라서 Xk는 trimmed mean을 의미한다. 이러한 trimmed sd를 사용함으로써 outlier에 민감하지 않은 robustness를 가진다.)

$$d(\hat{k}) = \sum_{i=1}^{N} \sum_{j=1}^{N} f(r_{ij}) 1(R - r_{ij}) \quad \text{where} \quad r_{ij} = | X_i \cdot \hat{k} - X_j \cdot \hat{k} |$$

: local density. k 방향으로 투영했을 때의 로컬 밀집도.
(r은 각 데이터가 투영됐을 때의 거리. 1 함수를 통해 거리 r이 cutoff R보다 작은 경우만 더하므로, 로컬에 대해서만 계산된다. 일종의 window 역할이다. f는 거리에 monotonically decreasing이므로 반비례한다. 즉, d는 로컬 내에서 얼마나 뭉쳐있는지를 의미한다. - 거리가 가까울수록 큰 값을 가지므로)

※ 2차원으로의 투영
두 개의 방향 $\hat{k}$와 $\hat{l}$을 탐색하고, 서로 직교한다.

$$s(\hat{k}, \hat{l}) = s(\hat{k}) s(\hat{l})$$

$r_{ij} = [(X_i \cdot \hat{k} - X_j \cdot \hat{k})^2 + (X_i \cdot \hat{l} - X_j \cdot \hat{l})^2]^{1/2}$ (거리는 k 방향으로의 거리와 l 방향으로의 거리 합)

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

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

둘의 차이를 잘 모르겠음

반복 실험을 통해, 로컬 밀도 탐색에 영향을 주는 건 함수 f(r)이 아니라 characteristic width (앞에서의 R로 추정됨)였다. 이 값은 로컬의 범위를 의미하므로, 알고리즘이 분포 변화에 얼마나 민감하게 반응할지를 결정하기 때문이다.

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)[6]이 있다. $I(\hat{k})$를 최대화하는 대신 동일한 함수 $I[SAT(\hat{k})] \in E^{n-1}(-\infty, \infty)$를 찾아 최대화하는 방법

---

6) J. H. Friedman and S. Steppel, "Non-linear constraint elimination in high dimensionality through reversible transformations," Stanford Linear Accelerator Cen., Stanford, Calif., Rep. SLAC PUB-1292, Aug. 1973.

으로, 제약조건이 사라져 더 안정적인 최적화를 할 수 있다. 똑같이 n-1개의 파라미터를 찾는다.

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

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 l에 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

[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 투영할 지 결정해야 하는데 2차원 투영이 1차원보다는 불안정하지만 더 많은 정보를 가지고 있다. 보통 1차원 투영 결과를 2차원 투영의 시작점으로 사용하는 것이 좋은 전략이었다.

논문 2. Projection Pursuit Indexes Based on Orthonormal Function Expansions

투영된 데이터의 분포와 정규분포 간의 weighted L2 거리를 나타내는 인덱스 Legendre index와 Hermite index의 차이를 살펴보고, 두 인덱스의 general form이면서도 정규분포와의 차이에 민감하도록 고안한 Natural Hermite index를 살펴본다.

normality(=randomness, 큰 엔트로피)를 null model로 두고, 엔트로피가 가장 작은 투영을 찾음으로써 정규분포와 가장 거리가 먼 투영을 찾는 것이 목적이다.

※ Polynomial expansion은 각 계수 자체를 하나의 인덱스로써 주목
- low-order indexes : long-sighted vision, 멀리서 큰 구조를 파악.
- higher-order indexes : short-sighted vision, 가까이서 내부의 복잡한 구조를 파악.

※ 기본 형태

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

$Y = T(X)$, $T: R \to R$ (Y는 X의 transformation. 특정 구조에 민감하게 반응하도록 진행)
X의 분포함수 F(x), 밀도함수 f(x), Y의 분포함수 G(Y), 밀도함수 g(y)

X 관점에서 표현하면,

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

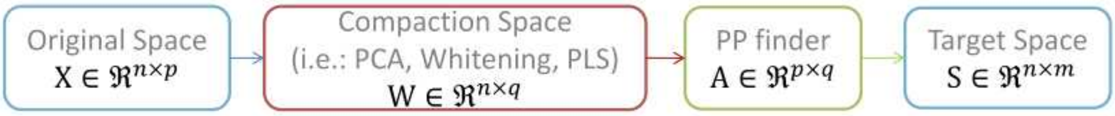
이 인덱스는 X와 표준정규분포 $\phi(x)$간 가중치 L2 거리이며, 가중치는 $\phi(x)/T'(x)^2$이다.
- 인덱스 I는 f에 대한 함수.
- f는 투영 벡터 α에 의존하므로, PP는 모든 가능한 α에 대한 인덱스의 local maxima를 찾는 과정을 포함한다.

Transformation family of T를 다음과 같이 정의하면 세 인덱스를 이렇게 정리할 수 있다.
$T_\sigma(X) = \sqrt{2\pi}\sigma(\Phi_\sigma(X) - 1/2)$

Legendre Index (by Friedman)	Hermite Index (by Hall)	Natural Hermite Index
$I^L = \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 \frac{1}{2\phi(x)} dx,$	$I^H = \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 dx.$	$I^N = \int_{\mathbb{R}} \{f(x) - \phi(x)\}^2 \phi(x) dx.$
$T(X) = 2\Phi(X) - 1$ $T'(X) = 2\Phi(X)$	$T(X) \propto \Phi_{\sigma=\sqrt{2}}(X)$ $T'(X) = \sqrt{\phi(x)}$	$T(X) = X$, identity 변형
$T_{\sigma=1}$	$T_{\sigma=\sqrt{2}}$	T_∞
가중치는 정규분포의 역수로, 꼬리에 대한 가중치	가중치 항상 1	가중치는 정규분포 식으로, center에 대한 관심

[5주차] 2015 - a projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets

유전자 데이터와 같은 $n \ll p$ 데이터는 PP를 적용하기에 계산 비용이 크고, poor local을 찾는다는 단점이 있어, 이러한 데이터에 맞게 데이터를 먼저 압축하고 PP를 적용하는 PP framework를 제안한 논문이다. 이 논문에서는 PCA, whitening, PLS 3가지의 차원축소 방법과 5가지의 PP 인덱스를 결합하여 그 성능을 비교한다.



※ PP 인덱스

- clustering을 위한 인덱스[7][8][9][10][11][12][13]
- supervised 분석을 위한 인덱스 : 원래 그룹 정보가 존재하여 이를 활용
- regression을 위한 인덱스[14][15]

이 논문에서는 supervised classification에 초점을 두어 다음과 같은 5개의 인덱스를 사용한다. 이 인덱스값들을 maximize하는 투영 a를 찾는 것이 PP의 목적이다.

1) Index Bhattacharya (Bat)[16][17]

그룹간 Bhattacharya 거리로, 각 i와 j는 서로 다른 그룹을 의미한다.

μ_i : 그룹 i에 속하는 데이터들의 투영 Xa의 평균 / σ_i : 분산

$$\Im_{Bat} = \min_{i,j \in C}\left\{ \frac{1}{4}\frac{(\mu_i - \mu_j)^2}{\sigma_i + \sigma_j} + \frac{1}{2}\log\left(\frac{\sigma_i + \sigma_j}{2\sqrt{\sigma_i \sigma_j}}\right)\right\},$$

---

7) [42] I. Perisic, C. Posse, Projection pursuit indices based on the empirical distribution function, J. Comput. Graph. Stat. 14 (3) (2005) 700-715.
8) [19] D. Pena, F. Prieto, Cluster identification using projections, J. Am. Stat. Assoc. 96 (456) (2001) 1433-1445.
9) [38] C. Posse, Projection pursuit exploratory data analysis, Comput. Stat. Data Anal. 20 (1995) 669-687.
10) [39] C. Posse, Tools for two-dimensional exploratory projection pursuit, J. Comput. Graph. Stat. 4 (2) (1995) 83-100.
11) [43] D. Cook, A. Buja, J. Cabrera, Projection pursuit indexes based on orthonormal function expansions, J. Comput. Graph. Stat. 2 (3) (1993) 225-250.
12) [30] M.C. Jones, R. Sibson, What is projection pursuit? J. R. Stat. Soc. Ser. A: General 150 (1) (1987) 1-37.
13) [13] J.H. Friedman, J.W. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Comput. 23 (9) (1974) 881-890.
14) [46] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Stat. Assoc. 76 (1981) 817-823.
15) [47] P. Hall, On projection pursuit regression, Ann. Stat. 17 (2) (1989) 573-588.
16) [28] E. Rodriguez-Martinez, J.Y. Goulermas, T. Mu, J.F. Ralph, Automatic induction of projection pursuit indices, IEEE Trans. Neural Netw. 21 (8) (2010) 1281-1295.
17) [5] L. Jimenez, D. Landgrebe, Hyperspectral data analysis and supervised feature reduction via projection pursuit, IEEE Trans. Geosci. Remote Sens. 37 (6) (1999) 2653-2667.

2) Index quality projected clusters (qpc)[18][19]

compact pure한 클러스터를 찾는 인덱스

α : 각 데이터 i, j에 대해 같은 그룹이면 α는 양수, 다른 그룹이면 음수 값을 가진다.

함수 G(.) : x=0에서 가장 큰 값을 가지는 함수로, 가우시안 함수를 사용할 수 있다.

투영된 거리가 가까울수록 큰 값을 가지므로, 이 인덱스를 최대화하면 같은 그룹끼리 잘 뭉쳐지고, 다른 그룹끼리 잘 분리되는 투영을 찾을 수 있다.

$$\Im_{qpc} = \sum_{i,j=1}^{n} \alpha_{i,j} G((x_i - x_j)\mathbf{a}),$$

3) Index Fisher linear discriminant analysis (lda)[20][21][22]

LDA 방법으로, 그룹 간 거리는 크게, 그룹 내 거리는 작게 하는 투영을 찾는다.

$$\Im_{lda} = 1 - \frac{|\mathbf{A}^T \mathbf{W} \mathbf{A}|}{|\mathbf{A}^T (\mathbf{W} + \mathbf{B}) \mathbf{A}|},$$

4) Index neighborhood components analysis (nca)[23]

nca 방법론의 비용함수로 사용되는 인덱스로, 올바른 그룹으로 분류되었을 확률을 의미한다.

$$\Im_{nca} = \sum_{i}^{n} \sum_{j \in \Omega_i} p_{ij},$$

$$p_{ii} = 0 \text{ and } p_{ij} = \exp(-\|x_i\mathbf{A} - x_j\mathbf{A}\|^2) / \sum_{k \neq i} \exp(-\|x_i\mathbf{A} - x_k\mathbf{A}\|^2)$$

5) Index locality preserving (Lp)[24]

Locality Preserving Projections (Lpp) 방법의 비지도 학습 인덱스이다.

인접한 데이터들을 함께 모으는 투영을 찾아낸다.

---

18) [24] M. Grochowski, W. Duch, Fast projection pursuit based on quality of projected clusters, in: A. Dobnikar, U. Lotric, B. Ster (Eds.), Adaptive and Natural Computing Algorithms Lecture Notes in Computer Science, vol. 6594, Springer, Berlin, Heidelberg, 2011, pp. 89-97.
19) [44] M. Grochowski, W. Duch, Projection pursuit constructive neural networks based on quality of projected clusters, in: V. Kurkova, R. Neruda, J. Koutnik (Eds.), Artificial Neural Networks—ICANN 2008, PT II, Lecture Notes In Computer Science, vol. 5164, 2008, pp. 754-762.
20) [27] J.R. Jee, Projection pursuit, Wiley Interdiscip. Rev.: Comput. Stat. 1 (2) (2009) 208-215.
21) [21] E. Lee, D. Cook, S. Klinke, T. Lumley, Projection pursuit for exploratory supervised classification, J. Comput. Graph. Stat. 14 (4) (2005) 831-846.
22) [45] E.-K. Lee, D. Cook, A projection pursuit index for large p small n data, Stat. Comput. 20 (3) (2010) 381-392.
23) [48] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighbourhood components analysis, in: Advances in Neural Information Processing Systems, vol. 17, 2005, pp. 513-520.
24) [52] X. He, S. Yan, Y. Hu, P. Niyogi, H. Jiang Zhang, Face recognition using Laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 328-340.

L : Laplacian matrix of k-neighborhood graph

$$\Im_{Lp} = 1/(\mathbf{A}^T \mathbf{XLA}),$$

(오타로 추정됨. $1/(A^{T}XLX^{T}A)$)

이 논문의 결과, PCA+LDA 조합이 가장 성능이 좋았고, qpc와 Lp 인덱스는 대체로 성능이 좋지 않았다.

[6주차] 2018-PPtreeViz-An R package for visualizing projection pursuit classification trees

classification을 위한 PP를 진행하기 위해 만든 R 패키지. 최적화에는 속도를 위해 Rcpp와 RcppArmadillo 패키지를 사용.

※ 패키지에서 제공하는 인덱스

1) LDA 인덱스 – 5주차 논문과 동일

$$I_{\mathrm{LDA,W}}(\mathbf{A}) = 1 - \frac{|\mathbf{A}^\top \mathbf{W} \mathbf{A}|}{|\mathbf{A}^\top (\mathbf{W} + \mathbf{B}) \mathbf{A}|}$$

2) PDA index

LDA를 적용하고 싶은데 변수간 상관관계가 높을 때 사용하는 인덱스.

상관관계가 높으면 LDA 인덱스의 분모가 0으로 가기 때문에 off-diagonal 행렬에 패널티 λ를 부여하여 식을 수정함. λ가 0이면 LDA 인덱스와 동일, 1이면 모든 변수간 상관도 제거.

$$I_{\mathrm{PDA,W}}(\mathbf{A}, \lambda) = 1 - \frac{|\mathbf{A}^\top \mathbf{W}_{\mathrm{PDA}} \mathbf{A}|}{|\mathbf{A}^\top \left(\mathbf{W}_{\mathrm{PDA}} + \mathbf{B}\right) \mathbf{A}|}$$

where

$$\mathbf{W}_{\mathrm{PDA}} = \mathrm{diag}(\mathbf{W}) + (1 - \lambda)\mathrm{offdiag}(\mathbf{W})$$

$$\begin{aligned} \mathrm{diag}(\mathbf{W}) &= \mathrm{diag}(w_{11}, \ldots, w_{pp}) \\ \mathrm{offdiag}(\mathbf{W}) &= \mathbf{W} - \mathrm{diag}(\mathbf{W}) \end{aligned}$$

3) L_r index

4) 1D Gini index

5) 1D entropy index

[코드] PPtreeViz 패키지에 5주차 논문의 다섯 가지 인덱스를 적용한 실습 파일 존재.

[8주차] Using_Projection-Based_Clustering_to_Find_Distance

고차원 데이터에서 거리와 밀도 구조(DDS, distance and density structures)에 의해 형성되는 클러스터를 잡아내기 위해 투영과 클러스터링을 동시에 진행하는 PBC를 제안한 논문이다. PBC는 topographic map을 통해 클러스터의 경향과 개수 둘다 잘 파악하게 한다. 실험을 통해 32개의 방법론과 비교했으며, PBC는 항상 우수한 성능을 보였다.

1) Combining DR with Clustering k-means (9개)
- PPCI 패키지[25) : PP를 클러스터링과 결합. **PPC_MD[26) (MinimumDensity), PPC_MC[27) (MaximumClusterbility), PPC_NC[28) (NormalisedCut)**, Kernel_PCA_Clust[29)
- tandem clustering : 변수 일부만 중요할 때, PCA나 FA를 적용한 후 k-means로 클러스터링을 진행하는 방법. clustrd 패키지 사용. RKM, FKM, KM, KM_I12

2) Conventional Clustering Algorithms (5개)
- K-means, SL, Spectral, Ward, PAM, MoG

3) Benchmarking of 18 Clustering Algorithms (18개)
- SOM, ADP, AP, DBscan, fuzzy, Markov, QTC, SOTA, CLARA, neural gas, HCL, hierarchical (complete, average, mcquitty, median, centroid linkage), DIANA

: 모두 FCPS 패키지 사용.

2019-PPCI an R Package for Cluster Identification using Projection Pursuit

PP가 찾은 최적의 투영으로 계층적 클러스터링을 반복하는 방법을 R 패키지로 구현한 논문이다. (PP, clustering) 조합으로 (mdh, mddc), (mch, mcdc), (ncuth, ncutdc) 3개의 조합을 제안하였다.

---

25) Hofmeyr, D., & Pavlidis, N. (2019). PPCI: an R package for cluster identification using projection pursuit. The R Journal, 11, 152. (https://CRAN.R-project.org/package=PPCI) (Hofmeyr and Pavlidis 2019).
26) Pavlidis, N. G., Hofmeyr, D. P., & Tasoulis, S. K. (2016). Minimum density hyperplanes. The Journal of Machine Learning Research, 17, 5414-5446.
27) Hofmeyr, D., & Pavlidis, N. (2015). Maximum clusterability divisive clustering. In 2015 IEEE symposium series on computational intelligence (pp. 780-786). Piscataway, NJ: IEEE.
28) Hofmeyr, D. P. (2016). Clustering by minimum cut hyperplanes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1547-1560.
29) Hofmeyr, D., & Pavlidis, N. (2019). PPCI: an R package for cluster identification using projection pursuit. The R Journal, 11, 152.

[10주차] A Section Pursuit Index for Finding Hidden Structure in Multiple Dimensions
선형 투영이 데이터의 중심에 있는 패턴을 잡아내지 못하는 문제를 해결하기 위해, section pursuit 방법으로 슬라이스를 찾아내는 방법을 제안한다.

※ Tour
- Ground Tour
  데이터를 계속 회전시키며 저차원에 투영하여 살펴보는 방법으로, 주로 2차원 projection plane에 투영시켜 사람이 직접 그 변화를 확인한다.
- Slicing
  Ground Tour의 projection plane으로부터 각 데이터 포인트 간의 직교 거리를 계산하여, 그 거리가 cutoff h보다 짧으면 슬라이스 내부, 그 바깥에 존재하면 슬라이스 외부로 간주한다. 이러한 방식으로 슬라이스 내부와 외부의 분포를 비교하여 holes와 grains를 찾아낼 수 있다.
- Guided Tour
  Grand tour에 PP를 결합한 개념으로, PP 인덱스를 최적화시키며 projection plane을 선택하고, 인덱스에 따라 원하는 패턴을 찾아낼 수 있다.

※ Index 정의[30]
$Y = XA$, p차원의 데이터 X를 A 벡터를 통해 d(보통 2)차원의 Y plane으로 투영한다.
- 각 데이터 포인트로부터 Y 평면까지의 유클리디안 거리 h

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

k개의 bin이 있다고 정의하여, 해당 bin의 슬라이스에 들어가면 S_k에 1을 더하고, 해당 bin이지만 슬라이스에 들어가지 않으면 C_k에 1을 더한다. (즉, S와 C는 개수)

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

Hole Index	Grain Index
$I_A^{\text{low}} = \sum_k [(c_k - s_k)]_{>\varepsilon}$,	$I_A^{\text{up}} = \sum_k [(s_k - c_k)]_{>\varepsilon}$
(외부-내부) 개수이므로, 이 인덱스가 크면 외부의 데이터가 많아 hole이다.	(내부-외부) 개수이므로, 이 인덱스가 크면 내부에 데이터가 많아 grain이다.

ϵ은 노이즈를 피하고 bin 개수에 의한 인위적인 영향을 피할 수 있게 한다.
특정 bin을 강조하거나 민감도를 조정하기 위해 다음과 같이 일반화될 수 있다.

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

30) Gous, A., and Buja, A. (2004), “Visual Comparison of Datasets Using Mixture Decompositions,” Journal of Computational and Graphical Statistics, 13, 1-19. [3]