주성분 분석에 대한 자습서

Jonathon Shlens*

구글 리서치
마운틴 뷰, CA 94043

(일자: 2014년 4월 7일, 버전 3.02)

주성분 분석(PCA)은 현대 데이터 분석의 중심입니다. 널리 사용되지만 (때때로) 잘 이해되지 않는 블랙 박스입니다. 이 문서의 목표는 이 블랙 박스 뒤에 숨겨진 마법을 없애는 것입니다. 이 원고는 주성분 분석이 작동하는 방법과 이유에 대한 견고한 직관을 구축하는 데 중점을 둡니다. 이 원고는 PCA 이면의 수학인 단순한 직관에서 도출하여 이 지식을 결정화합니다. 이 자습서는 아이디어를 비공식적으로 설명하는 것을 부끄러워하지 않으며 수학을 부끄러워하지도 않습니다. 두 가지 측면을 모두 다루면 모든 수준의 독자가 PCA와 이 기술을 적용하는 시기, 방법 및 이유를 더 잘 이해할 수 있기를 바랍니다.

I. 서론

주성분 분석(PCA)은 혼란스러운 데이터 세트에서 관련 정보를 추출하기 위한 단순하고 비모수적인 방법이기 때문에 신경과학에서 컴퓨터 그래픽에 이르는 다양한 분야의 최신 데이터 분석에서 표준 도구입니다.

최소한의 노력으로 PCA는 복잡한 데이터 세트를 더 낮은 차원으로 축소하여 때때로 숨겨진 단순화된 구조를 드러내는 방법에 대한 로드맵을 제공합니다.

이 자습서의 목표는 PCA에 대한 직관적인 느낌과 이 주제에 대한 철저한 토론을 모두 제공하는 것입니다. 간단한 예부터 시작하여 PCA의 목표에 대한 직관적인 설명을 제공합니다. 명시적 솔루션을 제공하기 위해 선형 대수학의 프레임워크 내에 배치하기 위해 수학적 엄격함을 추가하여 계속할 것입니다. PCA가 SVD(Singular Value Decomposition)의 수학적 기법과 밀접하게 관련되는 방법과 이유를 살펴보겠습니다. 이러한 이해는 실제 세계에서 PCA를 적용하는 방법에 대한 처방과 기본 가정에 대한 감사로 이어질 것입니다. PCA에 대한 철저한 이해가 기계 학습 및 차원 축소 분야에 접근하기 위한 토대를 제공하기를 바랍니다.

이 백서의 토론과 설명은 자습서의 정신에 따라 비공식적입니다. 이 논문의 목표는 교육하는 것입니다.
때로는 부록으로 분류되더라도 엄격한 수학적 증명이 필요합니다. 자습서에 중요하지는 않지만 수학에 대한 보다 완전한 이해를 원하는 모험적인 독자를 위해 증명이 제공됩니다. 나의 유일한 가정은 독자가 선형 대수학에 대한 실무 지식을 가지고 있다는 것입니다. 내 목표는 주로 선형 대수학의 아이디어를 기반으로 하고 통계 및 최적화 이론의 도전적인 주제를 피함으로써 철저한 토론을 제공하는 것입니다(단, 토론 참조). 제안, 수정 또는 의견이 있으면 언제든지 저에게 연락하십시오.

전자주소: jonathon.shlens@gmail.com

II. 동기 부여: 장난감의 예

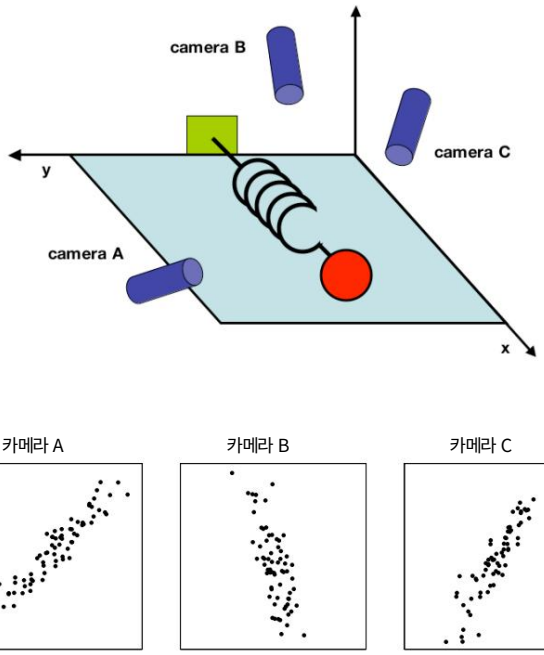
관점은 다음과 같습니다. 우리는 실험자입니다. 우리 시스템에서 다양한 양(예: 스펙트럼, 전압, 속도 등)을 측정하여 일부 현상을 이해하려고 합니다.

안타깝게도 데이터가 흐릿하고 불분명하며 심지어 중복되기 때문에 무슨 일이 일어나고 있는지 파악할 수 없습니다.
이것은 사소한 문제가 아니라 오히려 경험과학의 근본적인 장애물이다. 신경과학, 웹 인덱싱, 기상학 및 해양학과 같은 복잡한 시스템의 예는 많습니다. 기본 관계가 종종 매우 단순할 수 있기 때문에 측정할 변수의 수가 다루기 힘들고 때로는 기만적일 수도 있습니다.

예를 들어 그림 1에 표시된 물리학 다이어그램의 간단한 장난감 문제를 생각해 보십시오. 우리가 물리학자의 이상적인 용수철의 움직임을 연구하고 있다고 가정하십시오. 이 시스템은 질량이 없고 마찰이 없는 스프링에 부착된 질량 m의 볼로 구성됩니다. 볼이 평형 상태에서 약간 떨어진 곳에서 릴리스됩니다(예: 스프링이 늘어남). 용수철은 이상적이기 때문에 설정된 주파수에서 평형을 기준으로 x축을 따라 무한정 진동합니다.

이것은 x 방향을 따른 움직임이 명시적 시간 함수에 의해 해결되는 물리학의 표준 문제입니다.
즉, 기본 역학은 단일 변수 x의 함수로 표현될 수 있습니다.

그러나 무지한 실험자로서 우리는 이것에 대해 아무것도 모릅니다. 얼마나 많은 것은 고사하고 어떤 축과 치수가 측정에 중요한지 모릅니다. 따라서 우리는 3차원 공간에서 공의 위치를 측정하기로 결정합니다(우리는 3차원 세계에 살고 있기 때문입니다). 구체적으로 우리는 관심 있는 시스템 주변에 세 대의 무비 카메라를 배치합니다. 120Hz에서 각 동영상 카메라는 공의 2차원 위치(투영)를 나타내는 이미지를 기록합니다. 불행하게도 우리의 무지로 인해 실제 x, y, z 축이 무엇인지조차 모르기 때문에 시스템과 관련하여 임의의 각도에서 세 개의 카메라 위치 a, b 및 c 를 선택합니다. 측정 사이의 각도가 90o가 아닐 수도 있습니다 ! 이제 몇 분 동안 카메라로 녹화합니다. 큰 질문은 남아 있습니다. 이 데이터 세트에서 간단한 방정식으로 얻는 방법



무화과. 1 장난감 예. 진동하는 스프링에 부착된 공의 위치는 세 대의 카메라 A, B, C를 사용하여 기록됩니다. 각 카메라가 추적하는 공의 위치는 아래의 각 패널에 묘사되어 있습니다.

x의?

우리가 현명한 실험자라면 한 대의 카메라로 x축을 따라 위치를 측정했을 것이라는 것을 선험적으로 알고 있습니다. 그러나 이것은 현실 세계에서 일어나는 일이 아닙니다.
우리는 어떤 측정이 해당 시스템의 역학을 가장 잘 반영하는지 알지 못하는 경우가 많습니다. 또한 실제로 필요한 것보다 더 많은 차원을 기록하는 경우가 있습니다.

또한 성가신 실제 소음 문제도 해결해야 합니다. 장난감 예에서 이는 공기, 불완전한 카메라 또는 이상적이지 않은 스프링의 마찰을 처리해야 함을 의미합니다. 노이즈는 데이터 세트를 오염시켜 역학을 더 모호하게 만듭니다. 이 장난감 예는 실험자들이 매일 직면하는 도전 과제입니다. 추상적인 개념을 더 깊이 파고들 때 이 예를 염두에 두십시오. 바라건대, 이 논문이 끝날 때까지 주성분 분석을 사용하여 x를 체계적으로 추출하는 방법을 잘 이해할 수 있기를 바랍니다.

III. 프레임워크: 기본 변경

주성분 분석의 목표는 데이터 세트를 다시 표현하기 위해 가장 의미 있는 기준을 식별하는 것입니다. 이 새로운 기반이 노이즈를 걸러내고 숨겨진 구조를 드러낼 수 있기를 바랍니다. 용수철의 예에서 PCA의 명시적 목표는 "동역학이 x축을 따라 있습니다."를 결정하는 것입니다. 즉, PCA의 목표는 x , 즉 x축을 따라 있는 단위 기저 벡터가 중요한 차원임을 결정하는 것입니다.

이 사실을 확인하면 실험자는 어떤 역학이 중요하거나 중복되거나 노이즈인지 식별할 수 있습니다.

A. 소박한 근거

목표에 대한 보다 정확한 정의와 함께 데이터에 대한 보다 정확한 정의도 필요합니다. 우리는 모든 시간 샘플(또는 실험적 시도)을 데이터 세트의 개별 샘플로 취급합니다. 각 시간 샘플에서 여러 측정값(예: 전압, 위치 등)으로 구성된 데이터 세트를 기록합니다. 데이터 세트에서 한 시점에서 카메라 A는 해당 볼 위치 (xA, yA)를 기록합니다. 그런 다음 하나의 샘플 또는 시도를 6차원 열 벡터로 표현할 수 있습니다.

$$\text{엑스} = \begin{matrix} \text{xA} \\ \text{당신} \\ \text{xB} \\ \text{yB} \\ \text{xC} \\ \text{yC} \end{matrix}$$

여기서 각 카메라는 전체 벡터 X에 공 위치의 2차원 투영을 제공합니다. 120Hz에서 10분 동안 공의 위치를 기록하면 이 벡터의 10× 60×120 = 72000개를 기록한 것입니다.

이 구체적인 예를 통해 이 문제를 추상적인 용어로 재구성해 보겠습니다. 각 샘플 X는 m차원 벡터이며 여기서 m은 측정 유형의 수입니다. 마찬가지로, 모든 샘플은 직교 정규 기저에 의해 확장되는 m차원 벡터 공간에 있는 벡터입니다. 선형 대수학에서 우리는 모든 측정 벡터가 이 단위 길이 기본 벡터 세트의 선형 조합을 형성한다는 것을 알고 있습니다. 이 직교 기저는 무엇입니까?

이 질문은 일반적으로 종종 간과되는 암묵적인 가정입니다.
위의 장난감 예제 데이터를 수집했지만 카메라 A만 보았다고 가정합니다. (xA, yA)의 직교 정규 기저는 무엇입니까?
순진한 선택은 {(1,0),(0,1)}이지만 이것을 선택하는 이유 √ 2 {(tion? 그 이 $\frac{}{2,}$ $\frac{\sqrt{2}}{2}$),($\frac{\sqrt{2,2}}{}$ $\frac{}{2\sqrt{}}$ 2)} 또는 다른 임의의 로타 유는 순진한 기저가 우리가 데이터를 수집한 방법을 반영하기 때문입니다. 위치 (2,2)를 기록한다고 가정합니다. 우리는 √ 2 √ 2는 ( ) 방향으로 를 기록하지 않았고 0 당에서 $\frac{2\sqrt{}}{2,}$ $\frac{2}{2}$ 진자 방향. 오히려 우리는 카메라 창에서 왼쪽으로 2단위, 위로 2단위를 의미하는 카메라의 위치(2,2)를 기록했습니다. 따라서 원래의 기준은 데이터를 측정한 방법을 반영합니다.

이 순진한 기초를 선형 대수학에서 어떻게 표현합니까? 2차원의 경우 {(1,0),(0,1)}은 개별 행 벡터로 다시 캐스팅할 수 있습니다. 이러한 행 벡터로 구성되는 행렬은 2×2 항등 행렬 I입니다. m×m 항등 행렬을 구성하여 m 차원의 경우로 일반화할 수 있습니다.

$$B = \begin{matrix} \text{b1} \\ \text{b2} \\ \vdots \\ \text{비엠} \end{matrix} = \begin{matrix} 1\,0\cdots 0 \\ 0\,1\cdots 0 \\ \vdots\ \vdots\ \ddots\ \vdots \\ 0\,0\cdots 1 \end{matrix} = \text{나}$$

여기서 각 행은 m개의 성분을 갖는 직교 정규 기저 벡터 bi 입니다 . 순진한 기초를 효과적인 출발점으로 생각할 수 있습니다. 우리의 모든 데이터는 이 기준으로 기록되었습니다.

따라서 {bi}의 선형 조합으로 간단하게 표현할 수 있습니다.

나. 근거 변경

이 엄격함으로 이제 PCA가 요구하는 것을 더 정확하게 말할 수 있습니다. 원래 기저의 선형 조합인 다른 기저가 데이터 세트를 가장 잘 다시 표현합니까?

가까운 독자라면 선형이라는 단어가 눈에 띄게 추가된 것을 알아차렸을 것입니다. 실제로 PCA는 선형성이라는 엄격하지만 강력한 가정을 합니다. 선형성은 잠재 기반 집합을 제한하여 문제를 크게 단순화합니다. 이 가정으로 PCA는 이제 기본 벡터의 선형 조합으로 데이터를 다시 표현하는 것으로 제한됩니다.

X를 원래 데이터 세트라고 하고 각 열은 데이터 세트(예: X)의 단일 샘플(또는 순간)입니다. 장난감 예제에서 X는 m = 6이고 n = 72000인 m × n 행렬입니다.
Y를 선형 변환 P 와 관련된 또 다른 m × n 행렬이라고 합니다 . X 는 원래 기록된 데이터 세트이고 Y 는 해당 데이터 세트의 새로운 표현입니다.

$$PX = Y \tag{1}$$

또한 다음 수량을 정의하겠습니다.[1]

- pi 는 P 의 행입니다.
- xi 는 X (또는 개별 X) 의 열입니다 .
- yi 는 Y 의 열입니다.

방정식 1은 기저의 변화를 나타내므로 많은 해석이 가능합니다.

1. P 는 X를 Y 로 변환하는 행렬입니다.
2. 기하학적으로 P 는 회전과 늘이기입니다.
   X를 Y 로 변환합니다 .
3. P 의 행 {p1,...,pm}은 새로운 기저 vec의 집합입니다.
   X 의 열을 표현하기 위한 tors.

후자의 해석은 명확하지 않지만 PX의 명시적 내적을 작성하여 볼 수 있습니다 .

$$PX = \begin{matrix} p1 \\ \vdots \\ \text{오후} \end{matrix} \begin{matrix} \\ x1 \cdots xn \end{matrix}$$

$$Y = \begin{matrix} p1 \cdot x1 & \cdots & p1 \cdot xn \\ \vdots & \ddots & \vdots \\ \text{오후} \cdot x1 & \cdots & \text{오후} \cdot xn \end{matrix}$$

Y 의 각 열의 형식을 확인할 수 있습니다.

$$\text{이} = \begin{matrix} p1 \cdot xi \\ \vdots \\ \text{오후} \cdot \text{시} \end{matrix}$$

우리는 yi 의 각 계수가 다음의 내적임을 인식합니다.
xi 와 P의 해당 행. 즉, yi 의 j 계수 는 실제로 j에 대한 투영이며 yi 는 일
{p1,...,pm을 기준으로 한 투영입니다. }. 따라서 P 의 th 행 입니다. 이것은
P
의 행은 X 의 열을 나타내는 새로운 기저 벡터 집합입니다.

C. 남은 질문

선형성을 가정함으로써 문제는 적절한 기저 변화를 찾는 것으로 축소됩니다.
이 변환의 행 벡터 {p1,...,pm} 은 X 의 주성분이 됩니다. 이제 몇 가지 질문이 생깁니다.

- X를 재표현하는 가장 좋은 방법은 무엇입니까 ?
- 기저 P의 좋은 선택은 무엇입니까 ?

이러한 질문은 다음으로 Y 가 보여주길 원하는 기능이 무엇인지 스스로에게 물어봄으로써 답해야 합니다 . 분명히 합리적인 결과에 도달하려면 선형성을 넘어서는 추가 가정이 필요합니다. 이러한 가정의 선택은 다음 섹션의 주제입니다.

IV. 차이와 목표

이제 가장 중요한 질문이 있습니다. 데이터를 가장 잘 표현한다는 것은 무엇을 의미합니까? 이 섹션에서는 이 질문에 대한 직관적인 답변을 작성하고 추가 가정을 진행합니다.

A. 소음과 회전

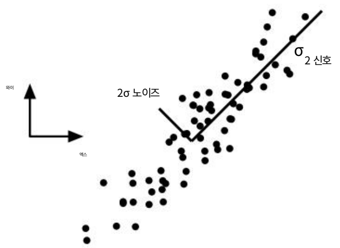
모든 데이터 세트의 측정 노이즈는 낮아야 합니다. 그렇지 않으면 분석 기술에 관계없이 신호에 대한 정보를 추출할 수 없습니다. 잡음에 대한 절대 척도는 존재하지 않지만 모든 잡음은 신호 강도에 비례하여 정량화됩니다. 일반적인 척도는 신호 대 잡음비(SNR) 또는 비율입니다.

분산 σ의 2 ,

$$SNR = \frac{2\sigma \text{ 신호}}{2\sigma \text{ 잡음}}.$$

높은 SNR(1)은 고정밀 측정을 나타내고 낮은 SNR은 매우 노이즈가 많은 데이터를 나타냅니다.

[1] 이 섹션에서 xi 와 yi 는 열 벡터이지만 미리 경고해야 합니다. 다른 모든 섹션에서 xi 와 yi 는 행 벡터입니다.



무화과. 2 카메라 A에 대한 (x, y)의 시뮬레이션된 데이터. 신호 및 노이즈 2 2 분 이터 클라우드에 대해서는 두 개의 신호 노이즈 선 으로 그래픽으로 표시됩니다 . 신호 및 노이즈 분산의 가장 큰 방 향은 기록 기준 (xA, yA) 이 아니라 가장 잘 맞는 선을 따릅니다.

카메라의 데이터를 자세히 살펴보겠습니다.

그림 2의 A. 스프링이 직선으로 이동한다는 점을 기억하면 모든 개별 카메라도 직선으로 움직임을 기록해야 합니다. 따라서 직선 운동에서 벗어나는 모든 퍼짐은 노이즈입니다. 신호 및 노이즈로 인한 분산은 다이어그램의 각 선으로 표시됩니다. 두 길이의 비율은 구름이 얼마나 얇은지를 측정합니다. 가능성에는 가는 선(SNR 1), 원(SNR = 1) 또는 그보다 더 나쁜 것이 포함됩니다. 합리적으로 좋은 측정값을 제시함으로써 정량적으로 측정 공간에서 분산이 가장 큰 방향에 관심 있는 역학이 포함되어 있다고 가정합니다. 그림 2에서 분산이 가장 큰 방향은 xA = (1,0)이나 yA = (0,1)이 아니라 구름의 장축 방향이다. 따라서 가정에 의해 관심 역학은 분산이 가장 크고 아마도 est SNR이 가장 높은 방향을 따라 존재합니다.

우리의 가정은 이러한 방향(즉, (xA, yA))이 가장 큰 분산의 방향과 일치하지 않기 때문에 우리가 검색하는 기저가 순진한 기저가 아님을 시사합니다 . 분산을 최대화(및 SNR을 가정하여)하는 것은 나이브 베이시스의 적절한 회전을 찾는 것과 일치합니다. 이 직관은 그림 2에서 라인 σ로 표시된 방향을 찾는 것과 일치합니다. 그림 2의 2차원 사례 신호 에서 가장 큰 분산의 방향은 데이터 클라우드에 가장 적합한 라인에 해당합니다. 따라서 순진한 기저를 회전하여 가장 적합한 선에 평행하게 놓이면 2D 경우에 대한 스프링의 동작 방향이 드러날 것입니다. 이 개념을 임의의 차원 수로 일반화하려면 어떻게 해야 합니까? 이 질문에 접근하기 전에 우리는 두 번째 관점에서 이 문제를 검토할 필요가 있습니다.

B. 중복

그림 2는 우리 데이터의 추가적인 교란 요인인 중복성을 암시합니다. 이 문제는 스프링의 예에서 특히 분명합니다. 이 경우 여러 센서가 동일한 동적 정보를 기록합니다. 그림 2를 재검토하고 2개의 변수를 기록하는 것이 정말 필요한지 묻습니다. 그림 3은 두 개의 임의 측정 유형 r1 과 r2 사이의 가능한 플롯 범위를 반영할 수 있습니다 . 왼쪽 패널은 두 가지를 묘사합니다.

무화과. 3 두 개의 개별 측정값 r1 및 r2 의 데이터에서 가능한 중복 스펙트럼 . 왼쪽의 두 측정값은 서로 예측할 수 없기 때문에 상관관계가 없습니다.

반대로, 오른쪽에 있는 두 측정값은 고도로 중복된 측정값을 나타내는 높은 상관관계가 있습니다.

명백한 관계가 없는 녹음. r2 에서 r1을 예측할 수 없기 때문에 r1 과 r2 는 상관관계가 없다고 말합니다 .

다른 극단에서 그림 3의 오른쪽 패널은 상관 관계가 높은 녹음을 보여줍니다. 이 극단은 여러 가지 방법으로 달성할 수 있습니다.

- 카메라 A와 B가 매우 가까운 경우 (xA, xB) 의 플롯 .
- (xA, x~A) 의 플롯에서 xA 는 미터 단위이고 ~xA 는 미터 단위입니다. 신장.

분명히 그림 3의 오른쪽 패널에서 둘 다가 아니라 단일 변수를 기록한 것이 더 의미가 있을 것입니다. 왜?
최적선을 사용하여 r2 에서 r1을 계산할 수 있기 때문입니다 (또는 그 반대). 하나의 응답만 기록하면 데이터가 더 간결하게 표현되고 센서 기록 횟수가 줄어듭니다(2 → 1 변수). 실제로 이것은 차원 축소의 핵심 아이디어입니다.

C. 공분산 행렬

변수가 2개인 경우 최적선의 기울기를 찾고 적합도를 판단하여 중복 사례를 식별하는 것은 간단합니다. 이러한 개념을 임의로 더 높은 차원으로 정량화하고 일반화하는 방법은 무엇입니까? 평균이 0인 두 가지 측정 세트를 고려하십시오.

$$A = \{a1,a2,\ldots,an\}, \qquad B = \{b1,b2,\ldots,bn\}$$

여기서 첨자는 샘플 번호를 나타냅니다. A와 B의 분산은 개별적으로 다음과 같이 정의됩니다.

$$\sigma_A^2 = \frac{1}{N}\sum_i a_i^2, \qquad \sigma_B^2 = \frac{1}{N}\sum_i b_i^2$$

A와 B 사이의 공분산은 간단한 일반화입니다.

$$\text{A와 B의 공분산} \equiv \sigma_{AB}^2 = \frac{1}{N}\sum_i a_i b_i$$

공분산은 두 변수 간의 선형 관계 정도를 측정합니다. 큰 양수 값은 양의 상관 관계가 있는 데이터를 나타냅니다. 마찬가지로 큰 음수 값은 음의 상관 관계가 있는 데이터를 나타냅니다. 공분산의 절대 크기는 중복 정도를 측정합니다. 공분산에 대한 몇 가지 추가 사실.

- σAB 는 A와 B가 상관관계가 없는 경우에만 0입니다(예: 그림 2, 왼쪽 패널).

AB A A = B인 경우 2 2. • σ = σ

A와 B를 대응하는 행 벡터로 동등하게 변환할 수 있습니다.

a = [a1 a2 ... an] b = [b1 b2 ... 억]

공분산을 내적 행렬 계산으로 표현할 수 있습니다.2

$$2\sigma \equiv ab \quad \frac{1}{N} abT \quad (2)$$

마지막으로 두 벡터에서 임의의 숫자로 일반화할 수 있습니다. 행 벡터 a 및 b의 이름을 각각 x1 및 x2 로 바꾸고 추가 인덱싱된 행 벡터 x3,...,xm을 고려하십시오.
새로운 m×n 행렬 X를 정의합니다.

$$\text{엑스} = \begin{matrix} x1 \\ \vdots \\ xm \end{matrix}$$

X 에 대한 한 가지 해석은 다음과 같습니다. X 의 각 행은 특정 유형의 모든 측정값에 해당합니다. X 의 각 열은 하나의 특정 시행(섹션 3.1의 X)의 측정 집합에 해당합니다. 이제 공분산 행렬 CX에 대한 정의에 도달했습니다 .

$$CX \equiv \frac{1}{N} XXT$$

행렬 CX = 를 고려하십시오. $\frac{1}{n}XXT$. CX 의 i 번째 j번째 요소 정 유형의 벡터와 j의 벡터 사이의 내적이며 CX의 여러 속성을 요약합니다. th 는 i번째 측 일 측정 유형. 우리는 할 수 있습니다

- CX 는 정사각형 대칭 m×m 행렬입니다(정리 2의 부록)
- CX 의 대각선 항은 특정 변수의 분산입니다. 측정 유형.
- CX 의 비대각선 항은 측정 유형 간의 공분산입니다.

CX는 가능한 모든 측정 쌍 간의 공분산을 캡처합니다. 공분산 값은 측정에서 노이즈와 중복성을 반영합니다.

- 대각 항에서 가정에 의해 큰 값은 흥미로운 구조에 반응합니다.
- 비대각선 용어에서 큰 크기는 해당합니다. 높은 중복성.

CX를 조작할 수 있는 옵션이 있다고 가정합니다 . 우리는 조작된 공분산 행렬 CY를 제안적으로 정의할 것입니다. CY 에서 어떤 기능을 최적화하고 싶습니까 ?

D. 공분산 행렬의 대각선화

우리의 목표는 (1) 공분산의 크기로 측정된 중복성을 최소화하고 (2) 분산으로 측정된 신호를 최대화하는 것이라고 말함으로써 마지막 두 섹션을 요약할 수 있습니다. 최적화된 공분산 행렬 CY는 어떻게 생겼습니까?

- CY 의 모든 비대각선 항은 0이어야 합니다. 따라서 CY는 대각 행렬이어야 합니다. 또는 달리 말하면 Y 는 상관관계가 없습니다.
- Y 의 각 연속 차원은 순위가 매겨져야 합니다. 분산에 따라.

CY를 대각화하는 방법에는 여러 가지가 있습니다 . PCA가 틀림없이 가장 쉬운 방법을 선택한다는 사실이 궁금합니다. PCA는 모든 기저 벡터 {p1,...,pm}이 직교 정규,
P 는 직교 행렬입니다. 이 가정이 가장 쉬운 이유는 무엇입니까?

PCA가 작동하는 방식을 상상해 보십시오. 그림 2의 간단한 예에서 P는 기본을 최대 분산 축과 정렬하기 위해 일반화된 회전 역할을 합니다. 여러 차원에서 이는 간단한 알고리즘으로 수행할 수 있습니다.

1. X 의 분산이 최대화되는 m차원 공간에서 정규화된 방향을 선택합니다 . 이 벡터를 p1로 저장합니다.
2. 분산이 최대화되는 다른 방향을 찾으십시오. 그러나 직교 정규성 조건 때문에 이전에 선택한 모든 방향과 직교하는 모든 방향으로 검색을 제한하십시오. 이 벡터를 파이 로 저장
3. m 벡터가 선택될 때까지 이 절차를 반복합니다.

결과적으로 정렬된 p의 집합이 주성분입니다.

원칙적으로 이 간단한 알고리즘이 작동하지만 정규 직교성 가정이 현명한 이유는 이것이 아닐 것입니다. 이 가정의 진정한 이점은

---

[2] 실제로 정규화 상수의 공분산 σ 약간의 변화는 추정 이론에서 발 $\frac{2}{AB}$ 다음과 같이 계산됩니다. $\frac{}{1n\ 1}$ Σ 아이 아이비 . 그만큼 생하지만 이 자습서의 범위를 벗어납니다.

이 문제에 대한 효율적이고 분석적인 솔루션입니다. 다음 섹션에서 두 가지 솔루션에 대해 논의할 것입니다.

rank-ordered variance의 규정으로 우리가 얻은 것을 주목하십시오. 주방향의 중요도를 판단하는 방법이 있습니다. 즉, 각 방향 pi 와 관련된 분산은 해당 분산에 따라 각 기본 벡터 pi를 순위 정렬하여 각 방향이 "주요"한 정도를 정량화합니다. 이제 여기에 도달하기 위해 만들어진 모든 가정의 의미를 검토하기 위해 잠시 멈출 것입니다. 수학적 목표.

E. 가정 요약

이 섹션에서는 PCA에 대한 가정에 대한 요약을 제공하고 이러한 가정이 제대로 수행되지 않는 경우에 대한 힌트를 제공합니다.

I. 선형성 선형성
은 문제를 기저의 변화로 구성합니다. 연구의 여러 영역에서 이러한 개념을 비선형 체제로 확장하는 방법을 탐구했습니다(토론 참조).

II. 큰 분산에는 중요한 구조가 있습니다.
이 가정은 또한 데이터가 높은 SNR을 갖는다는 믿음을 포함합니다. 따라서 관련 분산이 더 큰 주성분은 흥미로운 구조를 나타내는 반면 분산이 더 낮은 주성분은 노이즈를 나타냅니다. 이것은 강력하고 때로는 잘못된 가정이라는 점에 유의하십시오(토론 참조).

III. 주성분은 직교입니다.
이 가정은 PCA를 선형 대수 분해 기술로 해결하는 직관적인 단순화를 제공합니다. 이러한 기술은 다음 두 섹션에서 강조 표시됩니다.

우리는 PCA 도출의 모든 측면에 대해 논의했습니다. 남은 것은 선형 대수 솔루션입니다. 첫 번째 솔루션은 매우 간단한 반면 두 번째 솔루션은 중요한 대수적 분해를 이해하는 것과 관련됩니다.

V. 고유벡터 분해를 사용한 PCA 풀기

우리는 고유 벡터 분해의 중요한 속성을 기반으로 PCA에 대한 첫 번째 대수 솔루션을 도출합니다. 다시 한 번, 데이터 세트는 m × n 행렬인 X이며, 여기서 m은 측정 유형의 수이고 n은 샘플 수입니다. 목표는 다음과 같이 요약됩니다.

Y = PX 에서 정규 직교 행렬 P를 찾습니다 .

인 행. $\frac{1}{n}YY^T$ 는 대각 행렬입니다. P 의 CY ≡가 X 의 주성분 알 수 없는 변수로 CY를 다시 작성하는 것으로 시작합니다 .

$$C_Y = \frac{1}{n}YY^T$$
$$= \frac{1}{N}(PX)(PX)^T$$
$$= \frac{1}{N}PXX^TP^T$$
$$= P(\frac{1}{N}XX^T)P^T$$
$$C_Y = PC_XP^T$$

마지막 줄에서 X 의 공분산 행렬을 확인했습니다 .

우리의 계획은 모든 대칭 행렬 A가 고유 벡터의 직교 행렬(부록 A의 공식 3과 4에 의해)에 의해 대각선화된다는 것을 인식하는 것입니다 . 대칭 행렬 A 에 대해 정리 4는 A = EDET를 제공합니다 . 여기서 D 는 대각 행렬이고 E 는 열로 배열된 A 의 고유 벡터 행렬입니다.3 ,

이제 트릭이 나옵니다. 각 행 pi가 P ≡ E T 의 고유 벡터인 행렬이 되도록 행렬 P 를 선택합니다 . 이 관계와 부록 A의 정리 1 (P $\frac{1}{n}XX^T$. 이 선택으로

$P^{-1} = P^T$) CY 평가를 마칠 수 있습니다 .

$$C_Y = PC_XP^T$$
$$= P(E\,TDE)P^T$$
$$= P(P\,TDP)P^T$$
$$= (PP^T)D(PP^T)$$
$$= (PP^{-1})D(PP^{-1})$$
$$C_Y = D$$

P 의 선택이 CY를 대각화한다는 것은 명백합니다 . 이것이 PCA의 목표였습니다. 행렬 P 와 CY에서 PCA 결과를 요약할 수 있습니다.

- X 의 주성분은 nXXT 의 고유 벡터입니다 .
  $C_X = \frac{1}{n}$

일 • i CY 의 대각선 값은 X 의 분산입니다 .
파이 .

실제로 데이터 세트 X 의 PCA를 계산 하는 것은 (1) 각 측정 유형의 평균을 빼는 것과 (2) CX의 고유 벡터를 계산하는 것을 수반합니다. 이 솔루션은 부록 B에 포함된 매트 랩 코드에 설명되어 있습니다.

---

[g] 행렬 A는 r ≤ m개의 정규 직교 고유 벡터를 가질 수 있습니다. 여기서 r은 행렬의 순위입니다. A 의 순위가 m보다 작 으면 A는 퇴화되거나 모든 데이터가 r ≤ m 차원의 부분 공간을 차지합니다. 직교성의 제약을 유지하면서 행렬 E를 "채우기" 위해 (m-r) 추가 정규 직교 벡터를 선택하여 이 상황을 해결할 수 있습니다. 이러한 추가 벡터

이러한 방향과 관련된 분산이 0이므로 최종 솔루션에 영향을 주지 않습니다.

VI. SVD를 사용하는 보다 일반적인 솔루션

이 섹션은 수학적으로 가장 복잡하며 연속성을 많이 잃지 않고 건너뛸 수 있습니다. 완전성을 위해서만 제공됩니다. 우리는 PCA에 대한 또 다른 대수적 솔루션을 도출하고 그 과정에서 PCA가 SVD(singular value decomposition)와 밀접하게 관련되어 있음을 발견합니다. 사실, 이 둘은 매우 밀접하게 관련되어 있어서 종종 상호 교환적으로 사용되는 이름입니다. 하지만 우리가 보게 될 것은 SVD가 베이시스의 변화를 이해하는 보다 일반적인 방법이라는 것입니다.

우리는 신속하게 분해를 도출하는 것으로 시작합니다. 다음 섹션에서 우리는 분해를 해석하고 마지막 섹션에서 이러한 결과를 PCA와 관련시킵니다.

A. 특이값 분해

X는 임의 의 n × m 행렬[4] 이고 X TX 는 랭크 r, 정사각형, 대칭 m × m 행렬이라고 합니다. 의욕이 없어 보이는 방식으로 관심 있는 모든 양을 정의해 보겠습니다.

- {v 1,v 2,...,v r}는 대칭 행렬 X TX 에 대한 관련 고유값 {λ1,λ2,...,λr}을 갖는 정규 직교 m × 1 고유 벡터의 집합입니다 .

$$(X^{T}X)v_i = \lambda_i v_i$$

- $\sigma_i \equiv \sqrt{\lambda_i}$ 는 양의 실수이며 특이 값이라고 합니다.
  우.
- {u 1,u 2,...,u r}는 Xv i 로 정의되는 n × 1 벡터의 집합입니다 .
  $u_i \equiv \frac{1}{\sigma_i}$

최종 정의에는 두 가지 새롭고 예상치 못한 속성이 포함됩니다.

- $u_i \cdot u_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$
- $\|Xv_i\| = \sigma_i$

이러한 속성은 모두 정리 5에서 입증되었습니다. 이제 분해를 구성할 모든 조각이 있습니다. 특이값 분해의 스칼라 버전은 세 번째 정의의 재진술일 뿐입니다.

$$Xv_i = \sigma_i u_i \tag{3}$$

이 결과는 꽤 많은 것을 말해줍니다. X TX 의 고유 벡터를 곱한 X는 스칼라 곱하기 다른 벡터와 같습니다.

[4] 이 섹션에서만 우리는 m×n에서 n×m으로 관례를 뒤집고 있습니다. 이 파생의 이유는 섹션 6.3에서 명확해집니다.

고유벡터 집합 {v 1,v 2,...,v r} 과 벡터 집합 {u 1 , u 2,...,u r} 은 둘 다 r차원 공간에서 정규 직교 집합 또는 밑입니다.

그림 4의 규정된 구성을 따라 하나의 행렬 곱셈에서 모든 벡터에 대한 이 결과를 요약할 수 있습니다. 새로운 대각 행렬 Σ를 구성하는 것으로 시작합니다.

$$\Sigma \equiv \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & 0 & \\ & & \sigma_r & & \\ & 0 & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix}$$

여기서 σ1 ≥ σ2 ≥ ... ≥ σr 은 순위가 지정된 특이 값 집합입니다. 마찬가지로 수반되는 직교 행렬을 구성합니다.

$$V = [v_1 \; v_2 \; \dots \; v_m]$$

$$U = [u_1 \; u_2 \; \dots \; u_n]$$

V 와 U 에 대한 행렬을 각각 "채우기" 위해 (즉, 축퇴 문제를 처리하기 위해) 추가 (m-r) 및 (n-r) 또는 정상 벡터를 추가했습니다. 그림 4는 SVD의 매트릭스 버전을 형성하기 위해 모든 조각이 어떻게 함께 맞춰지는지를 그래픽으로 보여줍니다.

$$XV = U\Sigma$$

여기서 V 및 U 의 각 열은 분해의 스칼라 버전을 수행합니다(공식 3). V 는 직교 이므로 최종 형식에 도달하려면 T = V 입니다.
분해의 양변에 V를 곱할 수 있습니 $^{-1}$
다.

$$X = U\Sigma V^{T} \tag{4}$$

동기 부여 없이 파생되었지만 이 분해는 매우 강력합니다. 방정식 4는 임의의 행렬 X 가 직교 행렬, 대각 행렬 및 다른 직교 행렬(또는 회전, 늘이기 및 두 번째 회전)로 변환될 수 있음을 나타냅니다. 방정식 4를 이해하는 것은 다음 섹션의 주제입니다.

B. SVD 해석

SVD의 최종 형식은 간결하지만 두툼한 문장입니다. 대신 방정식 3을 다음과 같이 재해석해 보겠습니다.

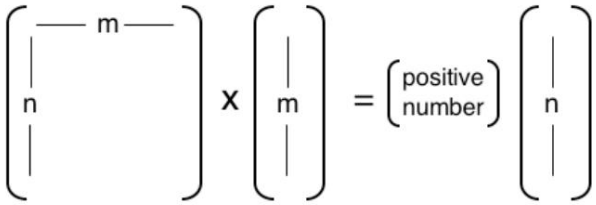
$$Xa = \text{킬로바이트}$$

여기서 a 와 b 는 열 벡터이고 k는 스칼라 상수입니다. 집합 {v 1,v 2,...,v m} 은 a 와 유사 하고 집합 {u 1,u 2,...,u n} 은 b와 유사합니다. 그러나 고유한 점은 {v 1,v 2,...,v m} 및 {u 1,u 2,...,u n}이 각각 m 또는 n 차원 공간에 걸쳐 있는 정규 직교 벡터 집합이라는 것입니다. 특히 느슨하게 말하면 이러한 세트는

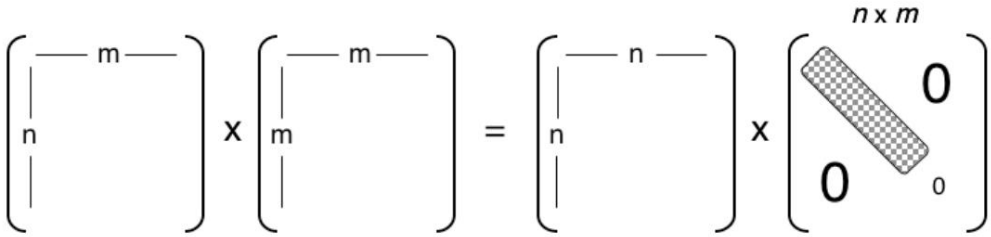
SVD의 스칼라 형식은 방정식 3으로 표현됩니다.

$$Xv_i = \sigma_i u_i$$

행렬 형식의 구성 뒤에 있는 수학적 직관은 모든 n개의 스칼라 방정식을 단 하나의 방정식으로 표현하기를 원한다는 것입니다 . 이 프로세스를 그래픽으로 이해하는 것이 가장 쉽습니다. 방정식 3의 행렬을 그리면 다음과 같습니다.



세 개의 새로운 행렬 V, U 및 Σ를 구성할 수 있습니다. 모든 특이값은 1순위 σ1 ~ ≥ σ2 ~ ≥ ... ≥ σr~ 입니다 . 그리고 올바른 일
응답 벡터는 동일한 순위 순서로 인덱싱됩니다. 관련 벡터 v i 및 u i 의 각 쌍은 해당 행렬을 따라 i 열에 쌓입니다 . 해당 특이값 σi는 Σ의 대각선(i번째 위치)을 따라 배치됩니다 . 이렇게 하면 다음과 같은 방정식 XV = UΣ 가 생성됩니다.



행렬 V 및 U 는 각각 m×m 및 n×n 행렬이고 Σ 는 대각선을 따라 0이 아닌 값(바둑판으로 표시됨)이 몇 개 있는 대각선 행렬입니다. 이 단일 행렬 방정식을 풀면 모든 n "값" 형식 방정식이 풀립니다.

무화과. 4 스칼라 형식(수식 3)에서 SVD의 행렬 형식(수식 4) 구성.

가능한 모든 "입력"(예: a) 및 "출력"(예: b). {v 1,v 2,...,v n} 및 {u 1,u 2,...,u n}이 가능한 모든 "입력" 및 "출력"에 걸쳐 있다는 견해를 공식화할 수 있습니까 ?

방정식 4를 조작하여 이 모호한 가설을 더 정확하게 만들 수 있습니다.

$$X = U\Sigma V^{T}$$
$$UTX = \Sigma V^{T}$$
$$UTX = Z$$

여기서 Z ≡ ΣV를 정의했습니다. ^{T}. 이전 . 이것을 비교
열 {u 1,u 2,...,u n}은 이제 방정식 1에 대한 ^{T}
U 방정식의 행이며, {u 1,u 2,...,u n}은 동일한 역할을 수행합니다. {p 1,p 2,...,p m} 로 . 따라서 U는 X 에서 Z 로의 기저 변화입니다. 이전에 열 벡터를 변환한 것처럼 열 벡터를 변환하고 있음을 다시 추론할 수 있습니다. 직교 정규 기저 U (또는 P)가 열 T 벡터를 변환한다는 T 사실은 U 가 X 의 열에 걸쳐 있는 기저임을 의미합니다.

열에 걸쳐 있는 기준을 X의 열 공간이라고 합니다. 열 공간은 모든 행렬의 가능한 "출력"에 대한 개념을 공식화합니다.

SVD에는 다음을 정의할 수 있는 재미있는 대칭이 있습니다.

유사한 수량 - 행 공간.

$$XV = \Sigma U$$
$$(XV)^{T} = (\Sigma U)^{T}$$
$$VTXT = U T\Sigma$$
$$VTXT = Z$$

여기서 Z ≡ U TΣ를 정의했습니다 . 다시 V 의 행 (또는 V의 열)은 T X 를 Z 로 변환하기 위한 정규 직교 기저입니다 . X 에 대한 전치로 인해 V 는 X의 행 공간에 걸친 정규 직교 기저가 됩니다. 행 공간도 마찬가지로 공식화됩니다. 임의의 행렬에 가능한 "입력"이 무엇인지에 대한 개념.

우리는 SVD의 전체 의미를 이해하기 위해 표면을 긁는 것일 뿐입니다. 그러나 이 자습서의 목적을 위해 PCA가 이 프레임워크에 포함되는 방법을 이해하는 데 충분한 정보가 있습니다.

C. SVD 및 PCA

PCA와 SVD가 밀접하게 관련되어 있음이 분명합니다. 원래의 m × n 데이터 행렬 X로 돌아가 보겠습니다. 다음을 정의할 수 있습니다.

PCA 요약 1. 데이터를 m×n
행렬로 구성합니다. 여기서 m 은 측정 유형의 수이고 n 은 샘플의 수입니다.
2. 각 측정 유형에 대한 평균을 뺍니다.
3. 공분산의 SVD 또는 고유 벡터를 계산합니다.

무화과. 5 주성분 분석을 수행하는 방법에 대한 단계별 지침 목록

새로운 행렬 Y를 n×m 행렬로.5

$$Y \equiv \frac{1}{\sqrt{n}} X^{T}$$

여기서 Y 의 각 열은 평균이 0입니다. Y TY를 분석하면 Y 의 선택이 명확해집니다 .

$$Y^{T}Y = \left(\frac{1}{\sqrt{n}} X^{T}\right)^{T} \left(\frac{1}{\sqrt{n}} X^{T}\right)$$

$$= \frac{1}{N} XX^{T}$$

$$Y^{T}Y = CX$$

구성에 의해 Y TY는 X 의 공분산 행렬과 같습니다. 섹션 5에서 우리는 X 의 주성분이 CX 의 고유 벡터라는 것을 알고 있습니다. Y 의 SVD를 계산하면 행렬 V 의 열에는 Y TY = CX 의 고유 벡터가 포함됩니다 .

따라서 V 의 열은 X 의 주성분입니다. 이 두 번째 알고리즘은 부록 B에 포함된 Matlab 코드로 캡슐화됩니다.

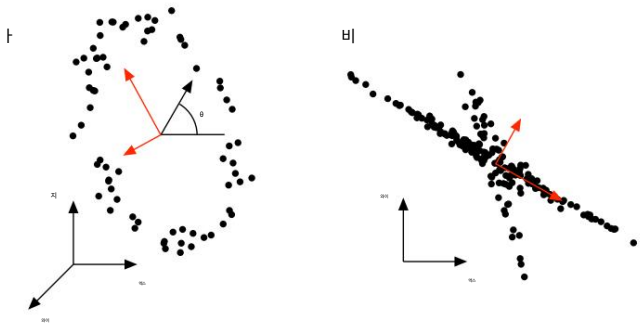
1 이것은 무엇을 의미합니까? V는 $Y \equiv \frac{1}{\sqrt{N}} X^{T}$ X 의 행 공간에 걸쳐 있습니다.
1 따라서 V는 또한 X 의 열 공간에 걸쳐 있어야 하며 주성분을 찾는 것
이 X 의 열 공간 6 에 걸쳐 있는 정규 직교 기저를 찾는 것과 같다는 결론을 내릴 수 있습니다.

VII. 논의

주성분 분석(PCA)은 선형 대수학의 분석 솔루션을 사용하여 복잡한 데이터 세트에서 간단한 기본 구조를 나타내기 때문에 널리 응용됩니다.

그림 5는 PCA 구현에 대한 간략한 요약을 제공합니다.

PCA의 주요 이점은 데이터 세트의 가변성을 설명하기 위한 각 차원의 중요성을 정량화하는 것에서 발생합니다. 특히, 각각에 따른 분산의 측정은



무화과. 6 PCA가 실패한 경우의 예(빨간색 선). (a) 대관람차에서 사람을 추적(검은색 점). 모든 역학은 순진한 기초의 비선형 조합인 바퀴 θ의 위상으로 설명될 수 있습니다. (b) 이 예제 데이터 세트에서 비가우시안 분산 데이터 및 비직교 축으로 인해 PCA가 실패합니다. 분산이 가장 큰 축은 적절한 답에 해당하지 않습니다.

주요 구성 요소는 각 차원의 상대적 중요도를 비교하는 수단을 제공합니다. 이 방법을 사용하는 이면의 암묵적인 희망은 소수의 주성분(즉, 측정 유형의 수보다 적음)에 따른 분산이 전체 데이터 세트의 합리적인 특성을 제공한다는 것입니다. 이 진술은 적극적인 연구의 광대한 영역인 차원 축소 방법의 배후에 있는 정확한 직관입니다. 용수철의 예에서 PCA는 6차원이 기록되더라도 대부분의 변동이 단일 차원(운동 방향 x)을 따라 존재한다는 것을 식별합니다 .

PCA가 수많은 실제 문제에 대해 "작동"하지만 부지런한 과학자나 엔지니어라면 언제 PCA가 실패하는지 질문해야 합니다. 이 질문에 답하기 전에 이 알고리즘의 주목할 만한 특징을 살펴보겠습니다. PCA는 완전히 비모수적입니다. 모든 데이터 세트를 연결할 수 있고 답이 나오므로 조정할 매개변수가 필요하지 않으며 데이터가 기록된 방식에 관계가 없습니다. 한 관점에서 PCA가 비모수적(또는 플러그 앤 플레이)이라는 사실은 답이 고유하고 사용자와 독립적이기 때문에 긍정적인 기능으로 간주될 수 있습니다. 또 다른 관점에서 PCA가 데이터 소스에 대해 불가지론적이라는 사실도 약점입니다. 예를 들어, 그림 6a에서 대관람차에 탄 사람을 추적한다고 생각해 보십시오. 데이터 포인트는 단일 변수인 바퀴의 세차 각도 θ로 명확하게 설명될 수 있지만 PCA는 이 변수를 복구하지 못합니다.

A. 차원축소의 한계와 통계

PCA의 한계를 더 깊이 이해하려면 기본 가정에 대한 고려와 동시에 데이터 소스에 대한 보다 엄격한 설명이 필요합니다. 일반적으로 말하면, 이 방법의 기본 동기는 데이터 세트의 상관관계를 해제하는 것, 즉 2차 종속성을 제거하는 것입니다. 이 목표에 접근하는 방식은 미국 서부에 있는 마을을 탐험하는 방법과 비슷합니다. 언제

[5] Y 는 섹션 6.1의 유도에 제시된 적절한 n×m 치수입니다. 이것이 6.1과 그림 4에서 차원의 "뒤집기"에 대한 이유입니다. 6 최종 목표가 X 의 열 공간에 대한 직교 정규 기저를 찾는 것이라면 Y 를 구성하지 않고 직접 계산할 수 있습니다. 대칭

√1n 의 SVD에 의해 생성된 U 의 열 구성 요소.

X 도 보안 주체여야 합니다.

다른 큰 길을 보고 좌회전이나 우회전을 하고 이 길을 따라 운전하는 식입니다. 이 비유에서 PCA는 탐색된 각 새 도로가 이전 도로와 수직이어야 하지만 분명히 이 요구 사항은 지나치게 엄격하며 데이터(또는 마을)는 그림 6b와 같이 비직교 축을 따라 정렬될 수 있습니다. 그림 6은 PCA가 만족스럽지 못한 결과를 제공하는 이러한 유형의 데이터에 대한 두 가지 예를 제공합니다.

이러한 문제를 해결하려면 최적의 결과로 간주되는 것을 정의해야 합니다. 차원 축소의 맥락에서 성공의 한 가지 척도는 축소된 표현이 원본 데이터를 예측할 수 있는 정도입니다. 통계 용어로 오차 함수(또는 손실 함수)를 정의해야 합니다. 일반적인 손실 함수인 평균 제곱 오차(예: L2 표준)에서 PCA가 데이터의 최적 감소 표현을 제공한다는 것을 증명할 수 있습니다. 이는 주성분에 대해 직교 방향을 선택하는 것이 원본 데이터를 예측하는 최상의 솔루션임을 의미합니다. 그림 6의 예에서 이 진술이 어떻게 사실일 수 있습니까? 그림 6의 직관에 따르면 이 결과는 오해의 소지가 있습니다.

이 역설에 대한 해결책은 우리가 분석을 위해 선택한 목표에 있습니다. 분석의 목표는 데이터의 상관 관계를 해제하는 것입니다. 즉, 데이터에서 2차 종속성을 제거하는 것이 목표입니다. 그림 6의 데이터 세트에서 변수 사이에 상위 종속성이 존재합니다. 따라서 2차 종속성을 제거하는 것은 데이터의 모든 구조를 드러내는 데 불충분합니다.[7]

상위 종속성을 제거하기 위한 여러 가지 솔루션이 있습니다. 예를 들어, 문제에 대한 사전 지식이 알려진 경우 비선형성(즉, 커널)이 데이터에 적용되어 데이터를 보다 적절한 순진한 기반으로 변환할 수 있습니다. 예를 들어, 그림 6a에서 데이터의 극좌표 표현을 검사할 수 있습니다. 이 파라메트릭 접근법은 종종 커널 PCA라고 합니다.

또 다른 방향은 예를 들어 감소된 차원에 따른 데이터가 통계적으로 독립적이어야 한다는 요구와 같이 데이터 세트 내에서 종속성에 대한 보다 일반적인 통계적 정의를 부과하는 것입니다. ICA(Independent Component Analysis)라고 하는 이 알고리즘 클래스는 PCA가 실패하는 많은 도메인에서 성공하는 것으로 입증되었습니다. ICA는 신호 및 이미지 처리의 많은 영역에 적용되었지만 솔루션을 (때때로) 계산하기 어렵다는 사실로 인해 어려움을 겪고 있습니다.

이 논문을 쓰는 것은 저에게 매우 교육적인 경험이었습니다. 이 문서가 PCA의 동기와 결과, 그리고 이 중요한 분석 기법 뒤에 숨겨진 가정을 이해하는 데 도움이 되기를 바랍니다. 이것이 도움이 되었으면 계속 글을 쓰도록 영감을 주는 메모를 보내주세요!

[7] 데이터 세트의 모든 종속성을 나타내기에 2차 종속성은 언제 충분합니까? 이 통계 조건은 1차 및 2차 통계가 데이터의 충분한 통계일 때 충족됩니다. 예를 들어 데이터 세트가 가우시안 분포인 경우에 발생합니다.

부록 A: 선형 대수학

이 섹션에서는 선형 대수학에서 이 백서에 중요한 몇 가지 불명확한 정리를 증명합니다.

1. 직교 행렬의 역행렬은 전치행렬입니다.

A 는 m×n 직교 행렬이고 여기서 ai 는 i 벡터입니다. A TA 의 i 번째 j 열 _ 번째 요소 는

$$(TA)i\,j \quad = AI^{T} \text{ 에이지 } = \begin{cases} 1 \text{ ifi} = j \\ 0 \text{ 그렇지 않으면} \end{cases}$$

따라서 A TA = I 이므로 A $^{-1}$ = A ^{티}.

2. 임의의 행렬 A에 대해 ATA 와 AAT 는 대칭입니다.

$$(AAT)^{티} = A^{TT} = AAT$$
$$(타)^{티} = A\,TA\,TT = TA\_$$

3. 행렬은 직교 대각화가 가능한 경우에만 대칭입니다.

이 문은 양방향이기 때문에 두 부분으로 구성된 "if-and-only-if" 증명이 필요합니다. "if-then"의 정방향 및 역방향 사례를 증명해야 합니다.

정방향 사례부터 시작하겠습니다. A 가 직교 대각선 가능 하면 A 는 대칭 행렬입니다. 가설에 따르면, 직교 대각화 가능하다는 것은 A = EDET (D 는 대각 행렬이고 E 는 A를 대각화하는 특수한 행렬) 와 같은 E가 존재한다는 것을 의미합니다. T A 를 계산해 봅시다. ,

.

$$ㅏ^{티} = (에데트)^{티} = E \;\; TTD\;TE\;T = 에데트 = A$$

분명히 A 가 직각으로 대각화 가능하다면 대칭이어야 합니다.

반대의 경우는 더 복잡하고 덜 깨끗하므로 독자에게 맡길 것입니다. 이것 대신에, 다소 설득력이 있지는 않더라도 "전방" 사례가 시사적이기를 바랍니다.

4. 대칭 행렬은 직교 고유 벡터의 행렬로 대각선화됩니다.

A를 관련 고유 벡터 {e1, e2,..., en} 가 있는 정사각형 n×n 대칭 행렬이라고 합니다 . Let E = [e1 e2 ... en] 여기서 th i 이 정리는 A = EDET를 만족하는 대각 행렬 D 가 존재한다고 주장합니다 ㄷ 와 열은 고유벡터 ei 입니다.

.

이 증명은 두 부분으로 나뉩니다. 첫 번째 부분에서 행렬의 고유 벡터가 모두 선형 독립인 경우에만 행렬이 직교 대각선화될 수 있음을 알 수 있습니다. 증명의 두 번째 부분에서 우리는 대칭 행렬이

모든 고유 벡터가 선형적으로 독립적일 뿐만 아니라 직교한다는 특별한 속성을 가지고 있으므로 증명을 완료합니다.

증명의 첫 번째 부분에서 A는 반드시 대칭이 아닌 일부 행렬이고 독립적인 고유 벡터를 갖는다고 가정합니다(즉, 축퇴 없음). 또한 E = [e1 e2 ... en]을 열에 배치된 고유 벡터의 행렬이라고 합니다. D를 고유값으로 두고 i번째 에 배치합니다.

i번째 위치에 대각 행렬. 일

이제 AE = ED 임을 보여드리겠습니다 . 방정식의 우변과 좌변의 열을 검사할 수 있습니다.

왼쪽 : AE = [Ae1 Ae2 ... Aen]
우측 : ED = [λ1e1 λ2e2 ... λnen]

분명히, AE = ED 이면 모든 i에 대해 Aei = λiei입니다 . 이 방정식은 고유값 방정식의 정의입니다. 따라서 AE = ED 가 되어야 합니다 . 약간의 재정렬은 A = EDE 1 을 제공하여 첫 번째 부분 증명을 완료합니다.

증명의 두 번째 부분에서는 대칭 행렬이 항상 직교 고유 벡터를 가짐을 보여줍니다. 일부 대칭 행렬의 경우 λ1 과 λ2를 고유 벡터 e1 과 e2 의 고유 고유값이라고 합니다 .

$$\begin{aligned}\lambda 1e1 \cdot e2 &= (\lambda 1e1) \quad \text{테 } 2\\ &= (Ae1) = \quad \text{테 } 2\\ &\quad e1\ TA\ T = \quad e2\\ &\quad e1\ TAe2\\ &\quad \text{티} = e1 \quad (\lambda 2e2)\end{aligned}$$

$$\lambda 1e1 \cdot e2 = \lambda 2e1 \cdot e2$$

마지막 관계에 의해 우리는 (λ1 λ2)e1 · e2 = 0임을 동일시할 수 있습니다. 고유값이 실제로 고유하다고 추측했으므로 e1 · e2 = 0 인 경우여야 합니다. 따라서 대칭 행렬의 고유 벡터는 직교합니다.

이제 A가 대칭 행렬 이라는 원래 가정으로 돌아가 보겠습니다 . 증명의 두 번째 부분에서 우리는 A 의 고유 벡터가 모두 정규 직교임을 압니다(우리는 정규화할 고유 벡터를 선택합니다). 이것은 E 가 직교 행렬이므로 정리 1, E 에 의해 최종 결과를 다시 작성할 수 있음을 의미합니다.

$$\text{티} = E \quad ^{-1}$$

A = 에데트

. 따라서 대칭 행렬은 고유 벡터의 행렬에 의해 대각선화됩니다.

5. 임의의 m × n 행렬 X에 대해 대칭 행렬 XTX는 {v 1,v 2,...,v n} 의 정규 직교 고유 벡터 집합과 관련 고유값 {λ1,λ2,...,λn} 집합을 가집니다. . 그런 다음 벡터 집합 {Xv 1,Xv 2,...,Xv n}은 직교 기저를 형성하며, 각 벡터 Xv i 의 길이는 √ λi 입니다 .

이러한 속성은 모두 이 세트의 두 벡터의 내적에서 발생합니다.

$$\begin{aligned}(Xv \quad i) \cdot (Xv \quad j) &= (Xv \quad i)^{\text{티}} (Xv \quad j)\\ &= v \quad i^{\text{티}} X\ TXv \quad j\\ &= v \quad i^{\text{티}} (\lambda jv \quad j)\\ &= \lambda jv \quad i \cdot v \quad j\end{aligned}$$

$$(Xv \quad i) \cdot (Xv \quad j) = \lambda j \delta i\ j$$

마지막 관계는 X 의 고유 벡터 집합이 크로네커 델타를 초래하는 직교이기 때문에 발생합니다. 보다 간단한 용어로 마지막 관계는 다음과 같이 말합니다.

$$(Xv \quad i) \cdot (Xv \quad j) = \begin{matrix}\lambda j & i = j\\ 0 & i \neq j\end{matrix}$$

이 방정식은 세트의 두 벡터가 직교임을 나타냅니다.

두 번째 속성은 각 벡터의 길이 제곱이 다음과 같이 정의됨을 실현하여 위의 방정식에서 발생합니다.

$$Xv \quad i \quad ^2 = (Xv \quad i) \cdot (Xv \quad i) = \lambda i$$

부록 B: 코드

이 코드는 Mathworks8의 Matlab 6.5(릴리스 13)용으로 작성되었습니다 .
이 코드는 계산적으로 효율적이지는 않지만 설명적입니다(간결한 주석은 %로 시작).

이 첫 번째 버전은 데이터 세트의 공분산을 검사하여 섹션 5를 따릅니다.

```
함수 [신호,PC,V] = pca1(데이터)
% PCA1: 공분산을 사용하여 PCA를 수행합니다. % 데이터 - 입
력 데이터의 MxN 행렬(M 차원, N 시행) % 신호 - 투사된 데이터
%                의 MxN 행렬 % PC - 각 열은 PC입니다.

%                V - Mx1 분산 행렬

[M,N] = 크기(데이터);

각 차원의 평균에서 % 빼기
mn = 평균(데이터,2); 데이터 =
데이터 - repmat(mn,1,N);

% 공분산 행렬 계산 공분산 = 1 / (N-1) * 데이터 * 데이터';

% 고유 벡터와 고유 값 찾기
```

[8] http://www.mathworks.com

```
[PC, V] = eig(공분산);

% 행렬의 대각선을 벡터로 추출
V = 진단(V);

% 분산을 내림차순으로 정렬 [junk, rindices] = sort(-1*V);

V = V(린디스);
PC = PC(:,인덱스);

% 원본 데이터 세트 신호 투사 = PC' * 데이터;
```

이 두 번째 버전은 SVD를 통한 PCA 컴퓨팅 섹션 6을 따릅니다.

```
함수 [신호,PC,V] = pca2(데이터)
% PCA2: SVD를 사용하여 PCA를 수행합니다. % 데이
터 - 입력 데이터의 MxN 행렬(M 차원, N 시행) % 신호 - 투사된 데이터
%                       의 MxN 행렬 % PC - 각 열은 PC입니다.
%                       V - Mx1 분산 행렬

[M,N] = 크기(데이터);

각 차원의 평균에서 % 빼기
mn = 평균(데이터,2); 데이터 =
데이터 - repmat(mn,1,N);

% 행렬 Y 구성
Y = 데이터' / sqrt(N-1);

% SVD는 모든 작업을 수행합니다.
[u,S,PC] = svd(Y);

% 분산 계산 S = diag(S); V = S.*S;

% 원래 데이터 신호 투사 = PC' * 데이터;
```