20 教師なし学習と主成分分析

教師なし学習

サンプル点はあるがラベルはない！クラスもY値もなく、予測するものは何もない。ゴール：データの構造を発見する。

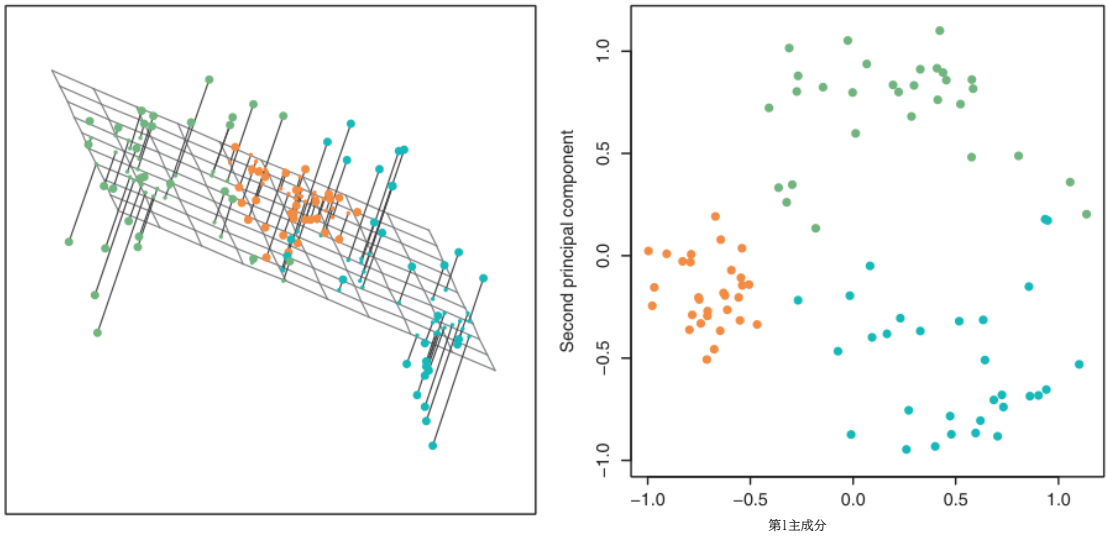
Examples:

- Clustering: partition data into groups of similar/nearby points.
- 次元削減：データは多くの場合，特徴空間内の低次元の部分空間（または多様体）の近くにある．[クラスタリングが類似したサンプルポイントをグループ化することであるのに対して， 次元削減は以下のことを目的としている．

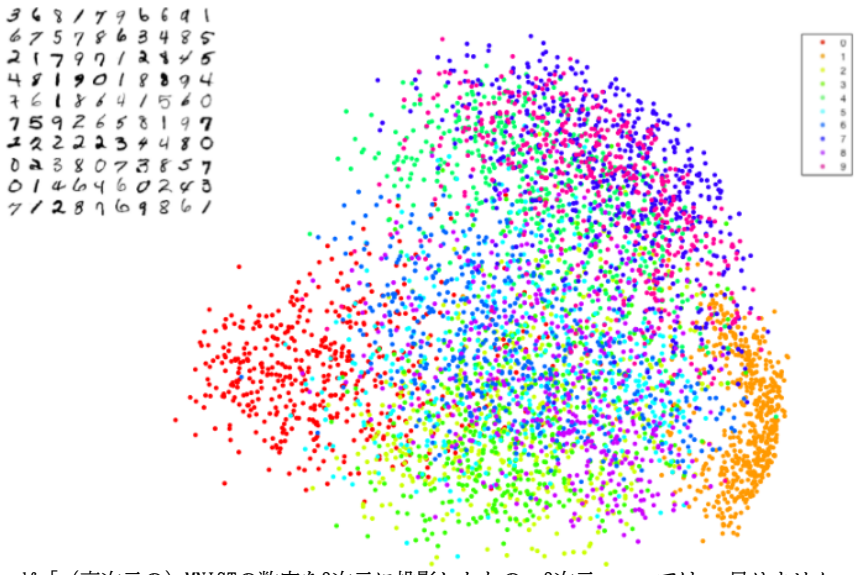
  identifying a continuous variation from sample point to sample point.]
- Density estimation: fit a continuous distribution to discrete data.
  [最尤推定を使って標本点にガウシアンを当てはめるとき，それは密度推定であるが，ガウシアンよりも複雑な関数，より局所的な変化を持つ関数を当てはめることもできる]．

主成分分析(PCA)（Karl Pearson, 1901）

目標：R^dのサンプル点が与えられたとき、最も多くのバリエーションを捉えるk個の方向を見つける。(次元削減)。



3dpca.pdf [PCAによって3次元点を2次元に投影した例]．



pcadigits.pdf [（高次元の）MNISTの数字を2次元に投影したもの。2次元Laurensではvan足りませんderMaaten完全にGeoffrey桁を分離するにはHinton, JMLR、 2008(MCLab)}だが、t − SNEの観測によると、0(赤)Octoberと30, 12014(オレンジ)4¤の桁は十分に分離できる。]

Why?

- Find a small basis for representing variations in complex things, e.g. faces, genes.
- Reducing # of dimensions makes some computations cheaper, e.g. regression.
- Remove irrelevant dimensions to reduce overfitting in learning algs.
  Like subset selection, but the “features” aren’t axis-aligned;
  they’re linear combos of input features.

[回帰や分類の前の前処理としてPCAが使われることがあります。]

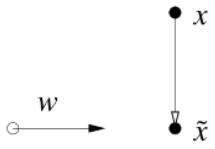
Let X be $n \times d$ design matrix. [No fictitious dimension.]

平均X_jはゼロである。[いつものように、x値の平均を計算し、各標本点から平均を引くことで、データをセンタリングできる]。

w を単位ベクトルとする.

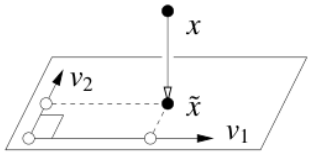
The orthogonal projection of point x onto vector w is $\tilde{x} = (x \cdot w)\, w$

If w not unit, $\tilde{x} = \frac{x \cdot w}{|w|^2} w$



[考え方は、最適な方向wを選び、すべてのデータをwに投影して、1次元空間で分析できるようにする。もちろん、d次元から1次元に投影すると、多くの情報が失われる。そこで、いくつかの方向を選ぶとする。それらの方向は部分空間にまたがっており、点を部分空間に直交投影したい。方向が互いに直交していれば簡単である.]

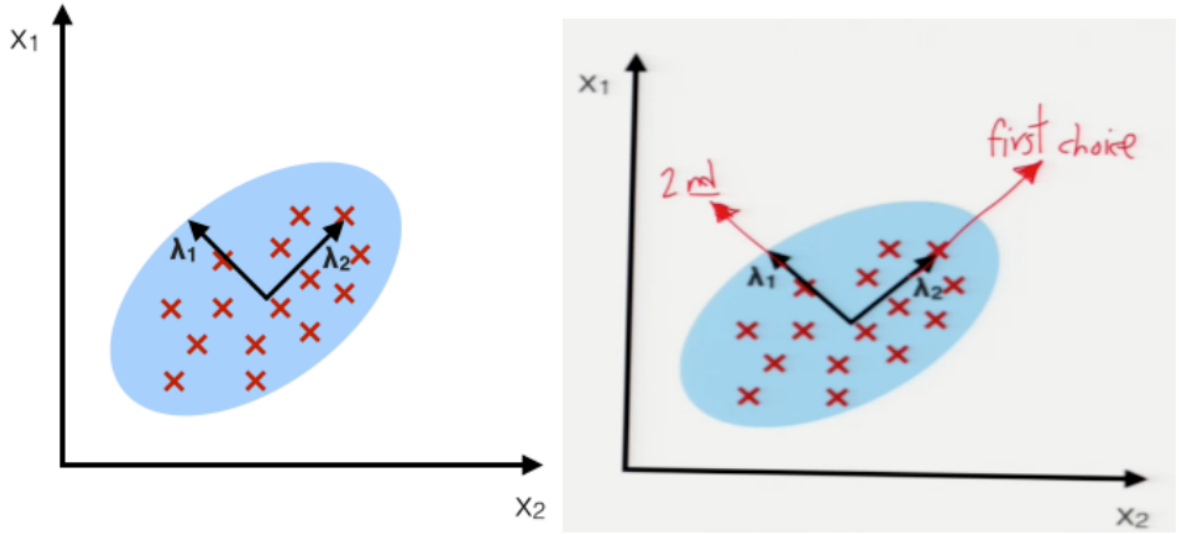
Given orthonormal directions $v_1, \ldots, v_k$, $\tilde{x} = \sum_{i=1}^{k} (x \cdot v_i)\, v_i$
[The word "orthonormal" implies they're mutually orthogonal and length 1.]



[通常，R^dの投影点は求めないが，主成分空間の主座標は求める]. $x \cdot v_i$

$X^\top X$ is square, symmetric, positive semidefinite, $d \times d$ matrix. [As it's symmetric, its eigenvalues are real.]
Let $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$ be its eigenvalues. [sorted]
Let $v_1, v_2, \ldots, v_d$ be corresponding orthogonal **unit** eigenvectors.
[It turns out that the principal directions will be these eigenvectors, and the most important ones will be the ones with the greatest eigenvalues. I will show you this in three different ways.]

PCA derivation 1: Fit a Gaussian to data with maximum likelihood estimation.
Choose k Gaussian axes of greatest variance.

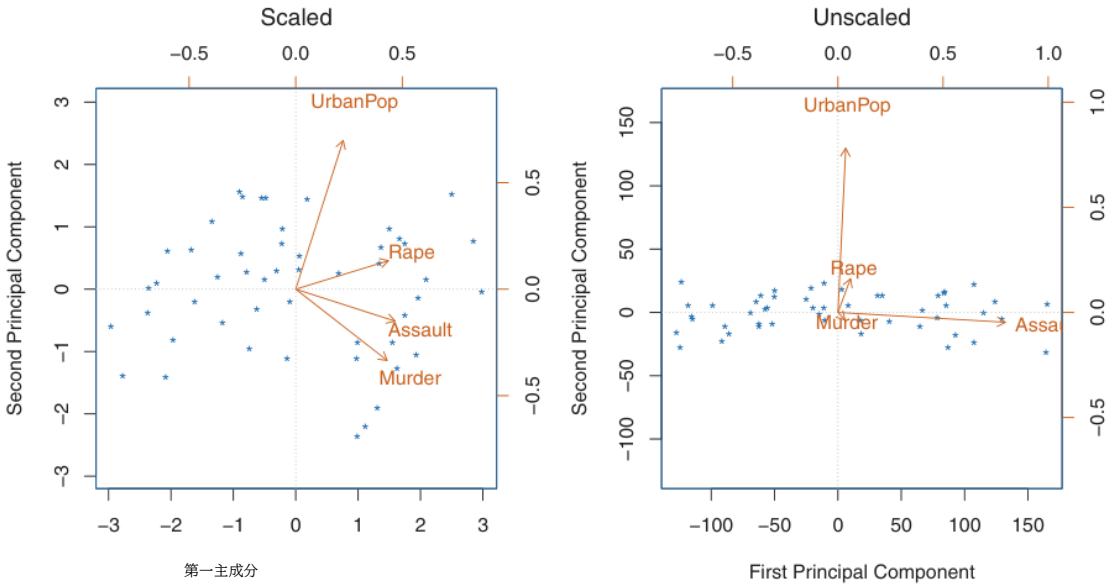


gaussfitpca.png [A Gaussian fitted to sample points.]

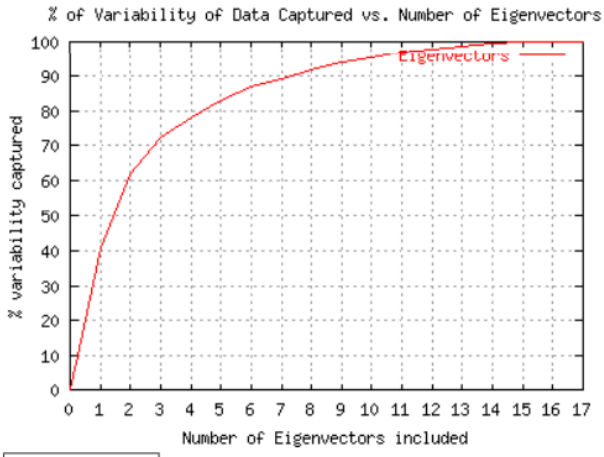
Recall that MLE estimates a covariance matrix $\hat{\Sigma} = \frac{1}{n} X^\top X$. [Presuming X is centered.]

PCA Alg:
- Center X.
- Optional: Normalize X. Units of measurement different?
  - Yes: Normalize.
    [Bad for principal components to depend on arbitrary choice of scaling.]
  - No: Usually don't.
    [If several features have the same unit of measurement, but some of them have smaller variance than others, that difference is usually meaningful.]
- Compute unit eigenvectors/values of $X^\top X$.

- 任意：固有値の大きさに基づいてkを選択する．
- 最適なk次元部分空間について、固有ベクトルv_{dk+1}，．．．
- トレーニング／テストデータの主成分空間における座標 x - v_i を計算する。[この射影を行う場合、2つの選択肢があります：入力された訓練データを射影する前にセンタリングを解除するか、あるいは訓練データをセンタリングしたときに使ったのと同じベクトルでテストデータを平行移動するかです]。



normalize.pdf [4次元データの2次元部分空間への射影。左が正規化されたデータ、右が正規化されていないデータ。矢印は、2つの主成分に投影された元の4つの軸を示す。データが正規化されていない場合、殺人のようなまれな事象は主軸方向にはほとんど影響しない。どちらが良いのだろうか？それは、殺人やレイプのような低頻度の出来事が、不釣り合いな影響を持つべきだと考えるかどうかによる。]

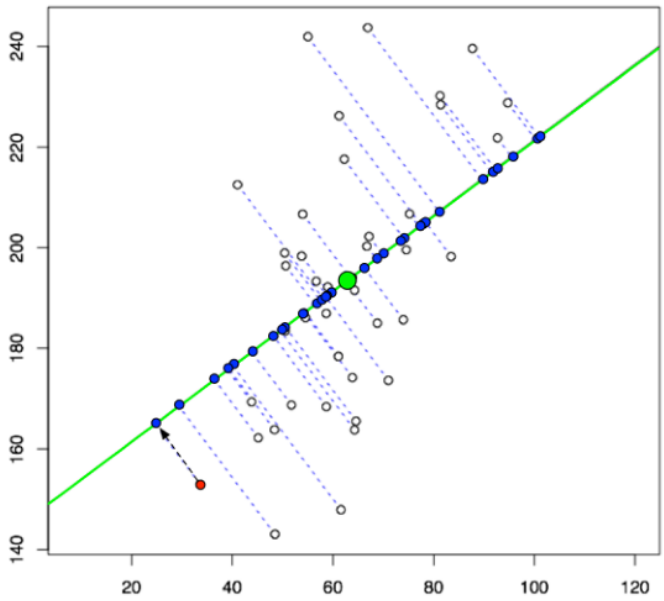


$$\%\ \text{of variability} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

variance.pdf [Plot of # of eigenvectors vs. percentage of sample variance captured for a 17D data set. In this example, just 3 eigenvectors capture 70% of the variance.]

[If you are using PCA as a preprocess for a supervised learning algorithm, there's a more effective way to choose k: (cross-)validation.]

PCAの導出2：投影されたデータの標本分散を最大化する方向wを見つける［言い換えれば、データを下に投影するとき、すべてを束にするのではなく、できるだけ広げたいのです］。 keep it as



project.jpg［点を線上に投影。青い点の標本分散が最大になるように緑の線の向きを選びたい］。

$$\text{Maximize } \mathrm{Var}(\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n\}) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i \cdot \frac{w}{|w|}\right)^2 = \frac{1}{n}\frac{|Xw|^2}{|w|^2} = \frac{1}{n}\underbrace{\frac{w^\top X^\top X w}{w^\top w}}_{\text{Rayleigh quotient of } X^\top X \text{ and } w}$$

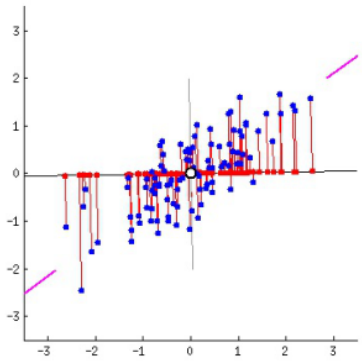
この分数はレイリー商と呼ばれるよく知られた構造である。この分数を見ると、近くに固有ベクトルの匂いがするはずです。これを最大化するには？］

If w is an eigenvector v_i, Ray. quo. $= \lambda_i$
→ of all eigenvectors, v_d achieves maximum variance λ_d/n.
One can show v_d beats every other vector too.
すべてのベクトルwは固有ベクトルの線形結合なので、そのレイリー商は固有値の凸結合になる。これを証明するのは簡単だが、時間がない。証明はウィキペディアで 「Rayleigh quotient 」を調べてください］。

［つまり、一番上の固有ベクトルが最適な方向を示してくれる。しかし、我々は通常k個の方向が欲しい。1つの方向を選んだら、その最適な方向と直交する方向を選ばなければならない。しかし、その制約のもとで、我々は再び標本分散を最大にする方向を選ぶのです］。wをv_dに直交するように制約したらどうでしょうか。するとv_{d1}が最適になります。

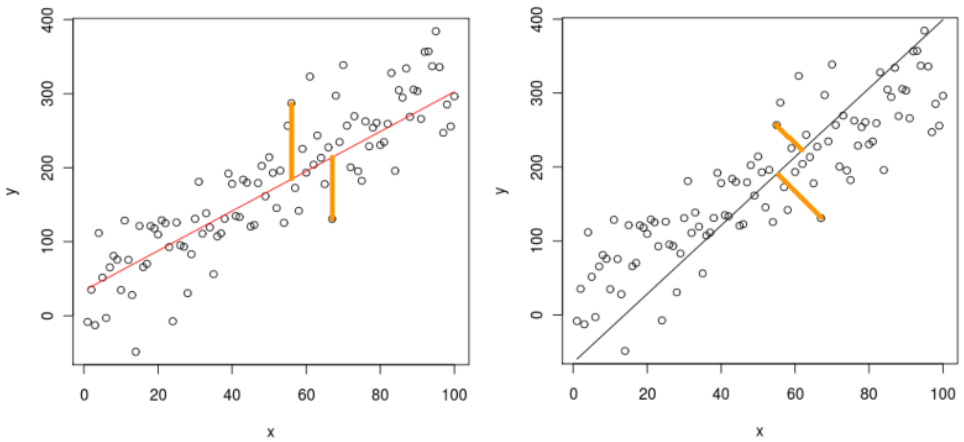
[And if we need a third direction orthogonal to v_d and v_{d-1}, the optimal choice is v_{d-2}. And so on.]

PCA derivation 3: Find direction w that minimizes "projection error"



PCAanimation.gif [これはGIFアニメーションです。残念ながら、アニメーションはPDFの講義ノートに含めることができません。赤い線の長さの二乗和が最も小さくなる黒い線の方向を求めよ] # # PCAanimation.gif [これはGIFアニメーションです。

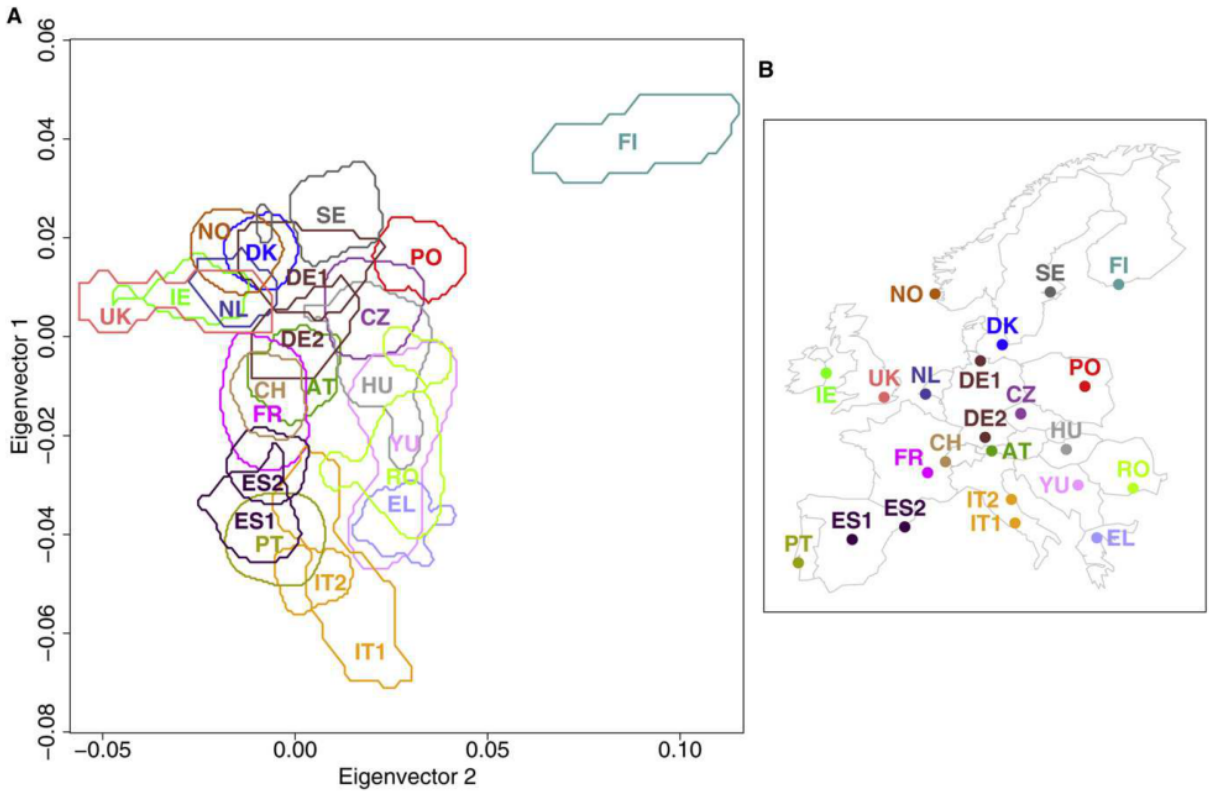
[これは一種の最小二乗直線回帰と考えることができるが、微妙だが重要な変更が1つある。固定の垂直方向の誤差を測定する代わりに、選択した主成分の方向に直交する方向の誤差を測定しているのだ]。



projlsq.png, projpca.png [最小2乗線形回帰とPCAの比較。線形回帰では，射影方向は常に垂直であるが，PCAでは，射影方向は射影超平面に直交する．しかしどちらの手法でも，射影距離の2乗和を最小化する．]

$$\text{Minimize} \sum_{i=1}^{n} \left|X_i - \tilde{X}_i\right|^2 = \sum_{i=1}^{n} \left|X_i - \frac{X_i \cdot w}{|w|^2} w\right|^2 = \sum_{i=1}^{n} \left(|X_i|^2 - \left(X_i \cdot \frac{w}{|w|}\right)^2\right)$$
$$= \text{constant n (導出2による分散).}$$

[投影誤差の最小化 = 分散の最大化．[この時点から，我々は同じ推論を続ける derivation 2.]



europegenetics.pdf (Lao et al., Current Biology, 2008.) [様々なヨーロッパ人の遺伝子の一塩基多型(SNP)行列の最初の2つの主成分の図。入力マトリックスには、ヨーロッパのこれらの場所（右）から来た2,541人と、309,790個のSNPがある。各SNPは2値なので、0か1の309,790次元と考える。出力（左）は、最初の2つの主成分上で、特定の国タイプの出身者が高密度に投影されたスポットを示している。これに関して驚くべきことは、投影された遺伝子型がいかに密接にヨーロッパの地理に似ているかということである。]

Eigenfaces

X それぞれdピクセルのn枚の顔画像がある。

[If we have a 200 × 200 image of a face, we represent it as a vector of length 40,000, the same way we represent the MNIST digit data.]

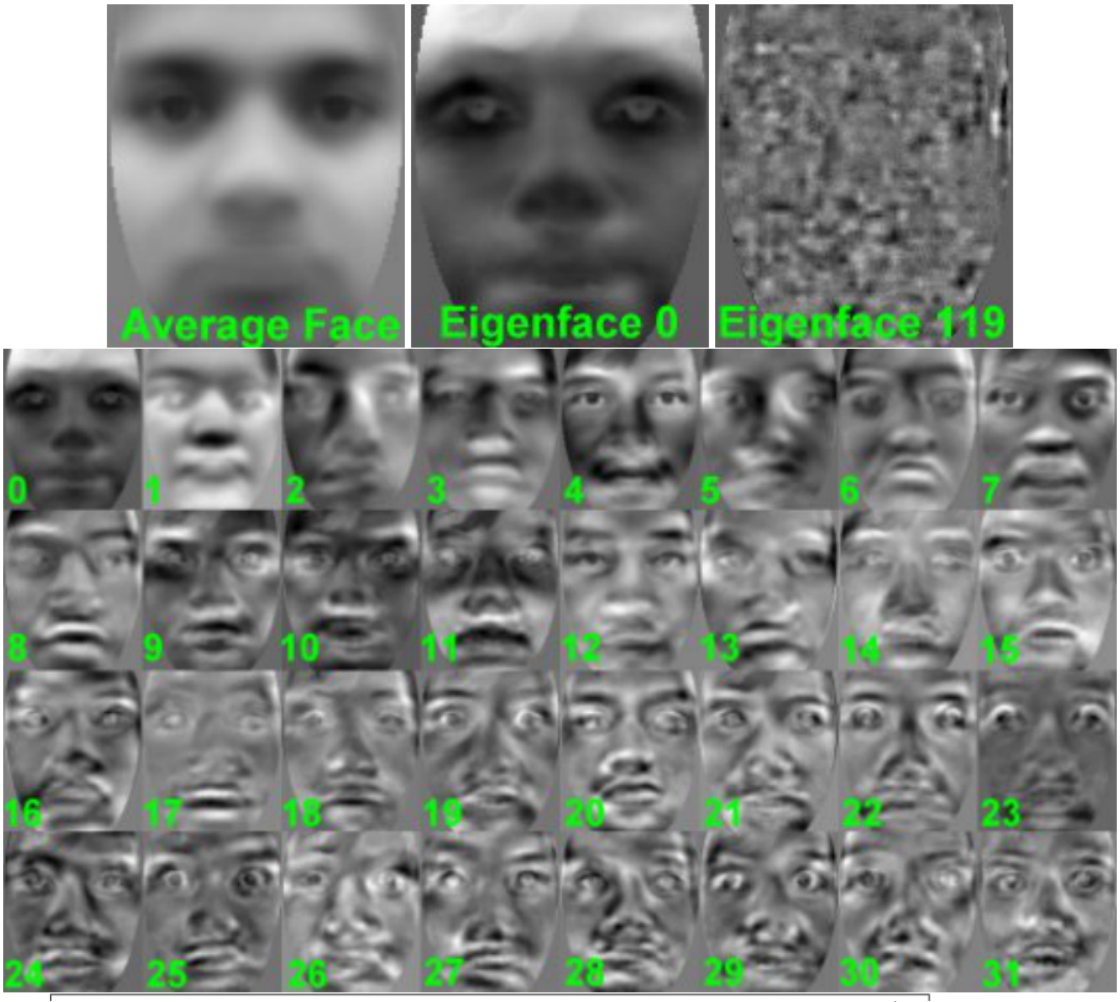
Face recognition: Given a query face, compare it to all training faces; find nearest neighbor in $\mathbb{R}^d$.

[This works best if you have several training photos of each person you want to recognize, with different lighting and different facial expressions.]

Problem: Each query takes $\Theta(nd)$ time.

Solution: Run PCA on faces. Reduce to much smaller dimension d'.
Now nearest neighbor takes $O(nd')$ time.
[Possibly even less. We'll talk about speeding up nearest-neighbor search at the end of the semester. If the dimension is small enough, you can sometimes do better than linear time.]

[それぞれ4万ピクセルの顔が500個保存されていて、それらを40の主成分に減らすとすると、各クエリの顔は、2000万ピクセルの代わりに、保存されている2万ピクセルの主座標を読み取る必要がある] # 。



facerecaverage.jpg, facereceigen0.jpg, facereceigen119.jpg, facereceigen.jpg が固有顔である。平均顔 ”は、データの中央揃えに使われた平均値である]。 [Images of

eigenfaceproject.pdf [最初の4つの固有ベクトルと50の固有ベクトルに投影された顔（左）の画像。最後の画像はぼやけているが、顔認識には十分である]。

最良の結果を得るためには、まず強度分布を均等化しましょう。

facerecequalize.jpg [Image equalization.]

[固有顔は完璧ではない。固有顔は顔の形と照明の両方を符号化している。理想的なのは、照明を除外して顔の形だけを分析する方法があることだが、それは難しい。最初の3つの固有顔は通常照明に関するものであり、最初の3つの固有顔を削除することで、より良い顔認識を得ることができると言う人もいる]。

[Blanz-Vetter face morphing video (morphmod.mpg)を表示].

BlanzとVetterは、3D顔モデリングにPCAをより洗練された方法で使用している。彼らは人々の顔の3Dスキャンを取り、人々の顔と理想化されたモデルとの対応関係を見つける。例えば、鼻先、口角、その他の顔の特徴を特定する。ピクセルの配列をPCAに入力する代わりに、顔のさまざまな点の3次元位置をPCAに入力する。これはより確実に機能する。]