PCA Two Ways

Nick Landolfi
Stanford University

Revision: August 2022, for EE263

Outline

- background: affine sets, projections, extremal trace
- minimum-residual affine set
- maximum-variance affine set
- examples with protein data

Affine sets

- a set $M \subset \mathbf{R}^n$ is *affine* if it contains the lines through any two of its points
  - i.e., $(1-\lambda)x + \lambda y \in M$ for all $x, y \in M$, $\lambda \in \mathbf{R}$
  - other terminology: affine subspace, linear variety, affine variety, flat set
- the affine sets are the solution sets of linear equations
  - given conforming A and b the set $\{x \in \mathbf{R}^n \mid Ax = b\}$ is affine, and vice versa
- the affine sets are translated subspaces
  - if M is affine, there exists a unique $a \in \mathbf{R}^n$ and subspace $S \subset \mathbf{R}^n$ so that $M = a + S$
  - notation $a + S$ means $\{a + x \mid x \in S\}$; dimension of M is dimension of S
- concrete representation for $M = a + S$ is $a + \mathrm{range}(U)$ where $\mathrm{range}(U) = S$ and $U^\top U = I$

Projection onto affine set

- given $a \in \mathbf{R}^n$ and $U \in \mathbf{R}^{n \times k}$ with $U^\top U = I$
- question: what is the projection of $x \in \mathbf{R}^n$ onto $a + \text{range}(U)$
- find $z \in \mathbf{R}^k$ to minimize
$$\|a + Uz - x\| = \|Uz - (x - a)\|$$
  - solution is $z^\star = U^\top (x - a)$
- projection is $Uz^\star + a = UU^\top x + (I - UU^\top)a$

Extremal trace problem

- problem: given $A = A^\top$, find $U \in \mathbf{R}^{n\times k}$ to

$$\begin{array}{ll} \text{maximize} & \operatorname{trace}(U^\top A U) \\ \text{subject to} & U^\top U = I \end{array}$$

- solution: pick first k (orthonormal) eigenvectors
  - let $A = Q\Lambda Q^\top$ be an eigendecomposition with $\lambda_1 \geq \cdots \geq \lambda_n$
  - then $U^\star = \left[\begin{array}{ccc} q_1 & \cdots & q_k \end{array}\right]$ is a solution
  - "a solution", since any permutation obtains same objective value

Extremal trace diagonalized problem

- $A = Q\Lambda Q^\top$ with $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$
- parameterize columns of U by basis Q; i.e., $U = QZ$ where $Z \in \mathbf{R}^{n \times k}$
  - columns of Z give coordinates of U in basis Q
  - there is one-to-one correspondence between U and Z
- in new coordinates, we find Z to

$$\begin{array}{ll} \text{maximize} & \text{trace}(Z^\top \Lambda Z) \\ \text{subject to} & Z^\top Z = I \end{array}$$

  - since $U^\top U = 1$ if and only if $Z^\top Z = 1$ and $U^\top A U = Z^\top \Lambda Z$
  - we have *diagonalized* the problem; changed coordinates to Q

Extremal trace diagonalized objective

- we have

$$\mathrm{trace}(Z^\top \Lambda Z) = \sum_{j=1}^{k} \lambda_i \tilde{z}_i^\top \tilde{z}_i = \sum_{j=1}^{n} \lambda_i ||\tilde{z}_i||^2 \leq \sum_{i=1}^{k} \lambda_i$$

  since

  - $||\tilde{z}_i||^2 \leq ||Z||^2 = 1$; i.e., the rows of an orthonormal matrix have norm bounded by 1
  - $\sum_{i=1}^{n} ||\tilde{z}_i||^2 = ||Z||_F = k$; i.e., the sum of squares elements of an orthonormal matrix is bounded by k

- we can can achieve this upper bound by selecting $Z^\star = \begin{bmatrix} e_1 & \cdots & e_k \end{bmatrix}$
- this choice corresponds to $U^\star = QZ^\star = \begin{bmatrix} q_1 & \cdots & q_k \end{bmatrix}$

Minimum-residual affine set

- given dataset $x_1, x_2, \ldots, x_m \in \mathbf{R}^n$
  - define $X = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}$, $\bar{x} = (1/m)X\mathbf{1}$, and $\bar{X} = X - (1/m)X\mathbf{1}\mathbf{1}^\top = (I - (1/m)\mathbf{1}\mathbf{1}^\top)X$
- we want to find the k-dimensional affine set "closest to" data
- problem: find $a \in \mathbf{R}^n$ and $U \in \mathbf{R}^{n \times k}$ (giving affine set $M_{a,U}$) to minimize

$$\sum_{i=1}^m \|x_i - \mathrm{proj}_{M_{a,U}}(x_i)\|^2$$

- solution: pick $a^\star = \bar{x}$ and U to have columns first k eigenvectors of $\bar{X}\bar{X}^\top$
  - eigenvectors of $\bar{X}\bar{X}^\top$ are first k left singular vectors of $\bar{X}$ (right singular vectors of $\bar{X}^\top$)

Minimum-residual affine set, offset

- fix $U \in \mathbf{R}^{n \times k}$, $U^\top U = I$
- find $a \in \mathbf{R}^n$ to minimize

$$\sum_{i=1}^m ||x_i - UU^\top x_i - (I - UU^\top)a||^2 = || \begin{bmatrix} I - UU^\top \\ \vdots \\ I - UU^\top \end{bmatrix} a - \begin{bmatrix} I - UU^\top & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & I - UU^\top \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} ||^2$$

  - solution $a^\star$ must satisfies normal equations, helps to notice $(I - UU^\top)^2 = (I - UU^\top)$
  - normal equations are $(I - UU^\top)a^\star = (I - UU^\top)\bar{x}$
  - $a^\star = \bar{x}$ works, and *does not depend on U*

Minimum-residual affine set, subspace

- assume $M = \bar{x} + \mathrm{range}(U)$
- find $U \in \mathbf{R}^{n\times k}$ to

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \|(I - UU^\top)(x_i - \bar{x})\|^2 \\ \text{subject to} & U^\top U = I \end{array}$$

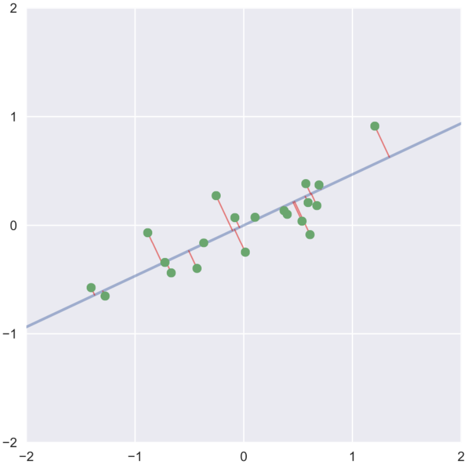
  - $\|(I - UU^\top)(x_i - \bar{x})\|^2 = \|(x_i - \bar{x})\|^2 - \|U^\top(x_i - \bar{x})\|^2$, first term constant w.r.t. U
  - $\sum_{i=1}^m \|U^\top(x_i - \bar{x})\|^2 = \|U^\top \bar{X}\|_F^2 = \mathrm{trace}(\bar{X}^\top U U^\top \bar{X}) = \mathrm{trace}(U^\top \bar{X}\bar{X}^\top U)$
- so we want to find $U \in \mathbf{R}^{n\times k}$ to

$$\begin{array}{ll} \text{maximize} & \mathrm{trace}(U^\top \bar{X}\bar{X}^\top U) \\ \text{subject to} & U^\top U = I \end{array}$$

  - an extremal trace problem, solution is first k eigenvectors of $\bar{X}\bar{X}^\top$

Total least squares

- measure distances orthogonal to line



Maximum-variance affine set

- given data set $x_1, \ldots, x_m \in \mathbf{R}^n$
- we want to find the k-dimensional affine subspace in which our data has "maximal variance"
- for affine subspace M, define the *projected mean* and *projected variance* by

$$\bar{x}(M) = \frac{1}{m}\sum_{i=1}^m \mathrm{proj}_M(x_i) \qquad \text{and} \qquad \nu(M) = \frac{1}{m}\sum_{i=1}^m \|\mathrm{proj}_M(x) - \bar{x}(M)\|^2$$

- problem: find $a \in \mathbf{R}^n$ and $U \in \mathbf{R}^{n\times k}$

$$\begin{array}{ll} \text{maximize} & \nu(a + \mathrm{range}\, U) \\ \text{subject to} & U^\top U = I \end{array}$$

- solution pick columns of U to be the first k eigenvectors $\bar{X}\bar{X}^\top$, any $a \in \mathbf{R}^n$ works

Maximum-variance affine set solution

- express $\bar{x}(a + \mathrm{range}\, U)$ as

$$\frac{1}{m}\sum_{i=1}^{m} UU^\top x_i + (I - UU^\top)a = UU^\top \bar{x} + (I - UU^\top)a$$

- drop the constant $1/m$ and write the objective

$$\sum_{i=1}^{m} \|\mathrm{proj}_M(x) - \bar{x}(S)\|^2 = \sum_{i=1}^{m} \|UU^\top x_i - UU^\top \bar{x}\|^2 = \sum_{i=1}^{m} \|UU^\top (x_i - \bar{x})\|^2$$

  - since U is orthonormal, $\|UU^\top(x_i - \bar{x})\| = \|U^\top(x_i - \bar{x})\|$, a familiar expression
  - the variance of the projected points does not depend on a

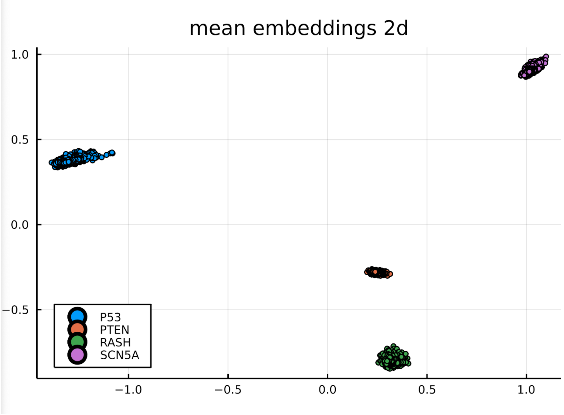
- so we want to find $U \in \mathbf{R}^{n\times k}$ to

$$\begin{array}{ll} \text{maximize} & \mathrm{trace}(U^\top \bar{X}\bar{X}^\top U) \\ \text{subject to} & U^\top U = I \end{array}$$

  - an extremal trace problem, pick the first k eigenvectors of $\bar{X}\bar{X}^\top$

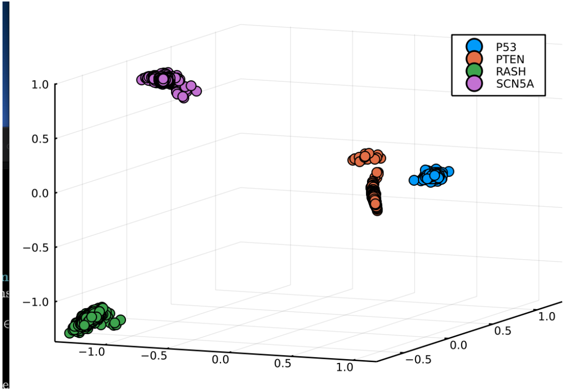
Protein embeddings 2d

- train a big neural network which maps proteins to vectors in $\mathbf{R}^{1024}$



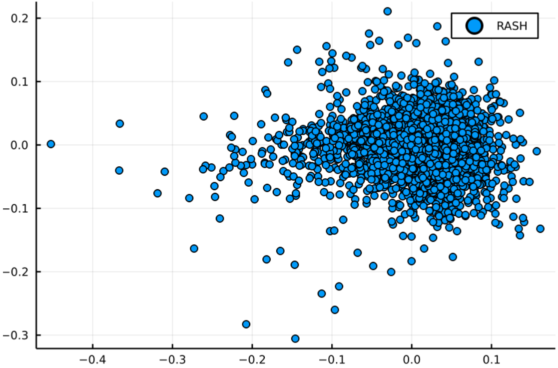
Protein embeddings 3D

- train a big neural network which maps proteins to vectors in $\mathbf{R}^{1024}$



Rash embeddings 3D

- RASH protein family



Rash embeddings 3D

- RASH protein family