Lecture 14: Principal Components Analysis

Introduction to Learning and Analysis of Big Data
Unsupervised Learning

- Until now: supervised
  - receive labeled data $(x, y)$
  - try to learn mapping $x \mapsto y$

- Current topic: **unsupervised**

- Receive unlabeled data only

- What’s a reasonable “learning” goal?
  - discover some structure in the data
  - does it clump nicely in space? (clustering)
  - is it well-represented by a small set of points? (sample compression/condensing)
  - does it have a sparse representation in some basis? (compressed sensing)
  - does it lie on a low-dimensional manifold? (dimensionality reduction)
Discovering structure in Data: Motivation

- Computational: “simpler” data is faster to search/store/process
- Statistical: “simpler” data allows for better generalization
- Dimensionality reduction has a denoising effect
- Example: human faces
  - 16x16 images and only grayscale: a 256-dimensional space
  - Curse of dimensionality: sample size exponential in $d$
  - Is learning faces hopeless?
  - Not if they have structure
  - Low-dimensional manifold [next slide]

- Common theme: learning of a high-dimensional signal is possible when its intrinsic structure is low-dimensional
Faces manifold

[credit: N. Vasconcelos and A. Lippman]
http://www.svcl.ucsd.edu/projects/manifolds
Principal Components Analysis (PCA)

- Dimensionality reduction technique for data in $\mathbb{R}^d$
- Problem statement [draw on board]:
  - given $m$ points $x_1, \ldots, x_m$ in $\mathbb{R}^d$
  - and target dimension $k < d$
  - find “best” $k$-dimensional subspace approximating the data
- Formally: find matrices $U \in \mathbb{R}^{d \times k}$ and $V \in \mathbb{R}^{k \times d}$
- that minimize

$$f(U, V) = \sum_{i=1}^{m} \| x_i - UVx_i \|_2^2$$

- $V : \mathbb{R}^d \to \mathbb{R}^k$ is “compressor”, $U : \mathbb{R}^k \to \mathbb{R}^d$ is “decompressor”
- Is $f : \mathbb{R}^{d \times k} \times \mathbb{R}^{k \times d} \to \mathbb{R}$ convex?
- No! $f(U, \cdot)$ and $f(\cdot, V)$ both convex, but not $f(\cdot, \cdot)$
- **Claim**: optimal solution achieved at $U = V^\top$ and $U^\top U = I$
  (columns of $U$ are orthonormal)
Principal Components Analysis: auxiliary claim

- **Optimization problem**: minimize $\sum_{i=1}^{m} \| x_i - UVx_i \|_2^2$
- **Claim**: optimal solution achieved at $U = V^\top$ and $U^\top U = I$
- **Proof**:
  - For any $U, V$, linear map $x \mapsto U V x$ has range $R$ of dimension $k$
  - Let $w_1, \ldots, w_k$ be orthonormal basis for $R$; arrange into columns of $W$.
  - Hence, for each $x_i$ there is $z_i \in \mathbb{R}^k$ such that $UVx_i = Wz_i$.
  - Note: $W^\top W = I$.
  - Which $z$ minimizes $f(x_i, z) := \| x_i - Wz \|_2^2$?
  - For all $x \in \mathbb{R}^d, z \in \mathbb{R}^k$,
    \[ f(x, z) = \| x \|_2^2 + z^\top W^\top Wz - 2z^\top W^\top x = \| x \|_2^2 + \| z \|_2^2 - 2z^\top W^\top x. \]
  - Minimize w.r.t. $z$: $\nabla_z f = 2z - 2W^\top x = 0 \implies z = W^\top x$.
  - Therefore
    \[ \sum_{i=1}^{m} \| x_i - UVx_i \|_2^2 = \sum_{i=1}^{m} \| x_i - Wz_i \|_2^2 \geq \sum_{i=1}^{m} \| x_i - WW^\top x_i \|_2^2. \]
  - $U, V$ are optimal, so $\sum_{i=1}^{m} \| x_i - UVx_i \|_2^2 = \sum_{i=1}^{m} \| x_i - WW^\top x_i \|_2^2$.
  - So instead of $U, V$ can take $W, W^\top$.  \( \square \)
- $WW^\top x$ is the **orthogonal projection** of $x$ onto $R$. [board]
PCA: reformulated

- **Optimization problem:**
  \[
  \text{minimize}_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \sum_{i=1}^{m} \|x_i - UU^\top x_i\|_2^2 \quad (\ast)
  \]

- For every \(x \in \mathbb{R}^d\) and \(U \in \mathbb{R}^{d \times k}\) with \(U^\top U = I\),
  \[
  \|x - UU^\top x\|^2 = \|x\|^2 - 2x^\top UU^\top x + x^\top UU^\top UU^\top x = \|x\|^2 - x^\top UU^\top x = \|x\|^2 - \text{trace}(x^\top UU^\top x) = \|x\|^2 - \text{trace}(U^\top xx^\top U).
  \]

- Recall **trace** operator:
  - Defined for any square matrix \(A\): sum of \(A\)'s diagonal entries
  - For any \(A \in \mathbb{R}^{p \times q}, B \in \mathbb{R}^{q \times p}\), we have \(\text{trace}(AB) = \text{trace}(BA)\)
  - \(\text{trace} : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}\) is a linear map

- Reformulating \((\ast)\):
  \[
  \text{maximize}_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \text{trace}\left(U^\top \sum_{i=1}^{m} x_i x_i^\top U\right).
  \]
Spectral theorem refresher

- Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric, $A = A^\top$
- and $Av = \lambda v$ for some $v \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$
- We say that $v$ is an eigenvector of $A$ with eigenvalue $\lambda$
- Fact: for symmetric $A$, eigenvectors of distinct eigenvalues are orthogonal. Proof:
  - if $Av = \lambda v$, $Au = \mu u$, $A = A^\top$
  - then $u^\top Av = \lambda \langle u, v \rangle = \mu \langle u, v \rangle$
  - hence, $\lambda \neq \mu \implies \langle u, v \rangle = 0$

- Theorem: for every symmetric $A \in \mathbb{R}^{n \times n}$ there is an orthogonal $V \in \mathbb{R}^{n \times n}$ (i.e., $V^\top = V^{-1}$) s.t. $V^\top AV = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.
- $V$ diagonalizes $A$
- $V = [v_1, \ldots, v_n]$ and $v_i$ is the eigenvector of $A$ corresponding to $\lambda_i$
- The $v_i$s are orthonormal ($V^\top V = I$) and span $\mathbb{R}^n$
- $A$ of the form $X^\top X$ or $XX^\top$ are positive semidefinite: $\Lambda \geq 0$
Maximizing the trace by $k$ top eigenvalues

- We want to

$$\maximize_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \text{trace} \left( U^\top \sum_{i=1}^{m} x_i x_i^\top U \right).$$

- Put $A = \sum_{i=1}^{m} x_i x_i^\top$.

- Diagonalize: $A = V \Lambda V^\top$
  - $V^\top V = VV^\top = I$
  - $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$

- So goal is:

$$\maximize_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \text{trace} \left( U^\top V \Lambda V^\top U \right).$$
Maximizing the trace by \( k \) top eigenvalues

- Goal:
  \[
  \maximize_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \text{trace} \left( U^\top V \Lambda V^\top U \right).
  \]

- We now show an upper bound on this value.
- Fix any \( U \in \mathbb{R}^{d \times k} \) with \( U^\top U = I \) and put \( B = V^\top U \).
- then \( U^\top V \Lambda V^\top U = B^\top \Lambda B \).
- Hence, goal is equal to \( \text{trace}(\Lambda BB^\top) = \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{k} (B_{ji})^2 \).
- Define \( \beta_j = \sum_{i=1}^{k} (B_{ji})^2 \).
- \( B^\top B = U^\top VV^\top U = U^\top U = I \). Therefore \( \|\beta\|_1 = \sum_{j=1}^{d} \beta_j = \text{trace}(BB^\top) = \text{trace}(B^\top B) = \text{trace}(I) = k \).
- Claim: \( \beta_j \leq 1 \). Proof:
  - choose \( \tilde{B} \in \mathbb{R}^{d \times d} \) s.t. \( \tilde{B}[\cdot, 1:k] = B \) and \( \tilde{B}^\top \tilde{B} = I \). Hence \( \tilde{B}^T = \tilde{B}^{-1} \).
  - So \( \tilde{B} \tilde{B}^\top = I \). Hence \( \forall j \in [d], \sum_{i=1}^{d} (\tilde{B}_{ji})^2 = 1 \implies \sum_{i=1}^{k} (B_{ji})^2 \leq 1 \).
- hence:
  \[
  \text{trace}(U^\top V \Lambda V^\top U) \leq \max_{\beta \in [0,1]^d : \|\beta\|_1 \leq k} \sum_{j=1}^{d} \lambda_j \beta_j = \sum_{j=1}^{k} \lambda_j.
  \]
PCA: solution

- We have:
  - $A = \sum_{i=1}^{m} x_i x_i^\top$
  - $A = V \Lambda V^\top$
  - $V^\top V = V V^\top = I$
  - $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$
  - $U \in \mathbb{R}^{d \times k}$ with $U^\top U = I$

- Under the above, we showed $\text{trace}(U^\top A U) \leq \sum_{j=1}^{k} \lambda_j$.

- Get this with equality: set $\hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k]$ where $A \hat{u}_i = \lambda_i \hat{u}_i$

- $i$th column of $\hat{U}$ is $A$'s eigenvector corresponding to $\lambda_i$

- then $\text{trace}(\hat{U}^\top A \hat{U}) = \sum_{j=1}^{k} \lambda_j$.

- So $\hat{U}$ is the optimal solution.
PCA: solution value

- Recall our original goal: minimize $U \in \mathbb{R}^{d \times k}: U^T U = I \sum_{i=1}^{m} \|x_i - UU^T x_i\|_2^2$.
- We showed, for $A = \sum_{i=1}^{m} x_i x_i^\top$,

  $$\arg\min_{U \in \mathbb{R}^{d \times k}: U^T U = I} \sum_{i=1}^{m} \|x_i - UU^T x_i\|_2^2 = \arg\max_{U \in \mathbb{R}^{d \times k}: U^T U = I} \text{trace}(U^T A^T U).$$

- We just showed that $\hat{U}$ maximizes $\text{trace}(U^T A U)$.
- Value of solution:
  - $\hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k]$ ($k$ top eigenvectors), $\text{trace}(\hat{U}^T A \hat{U}) = \sum_{j=1}^{k} \lambda_j$.
  - $\sum_{i=1}^{m} \|x_i - \hat{U}\hat{U}^T x_i\|_2^2 = \sum_{i=1}^{m} \|x_i\|_2^2 - \text{trace}(\hat{U}^T A \hat{U})$
  - $\sum_{i=1}^{m} \|x_i\|_2^2 = \text{trace}(A) = \text{trace}(V\Lambda V^\top) = \text{trace}(V^\top V\Lambda) = \text{trace}(\Lambda) = \sum_{i=1}^{d} \lambda_i$.
  - Optimal objective value: $\sum_{i=1}^{m} \|x_i - \hat{U}\hat{U}^T x_i\|_2^2 = \sum_{i=k+1}^{d} \lambda_i$.
- This is the “distortion” — how much data we throw away.
PCA: the variance view

- Suppose the data \( S = \{x_i, i \leq m\} \) is centered: \( \frac{1}{m} \sum_{i=1}^{m} x_i = 0 \)
- Let’s project it onto a 1-dim subspace w/ max variance [board]
- 1-dim projection operator: \( uu^\top \), for \( u \in \mathbb{R}^d \), \( u^\top u = 1 \)

\[
\text{Var}_{X \sim S}[u^\top X] = \frac{1}{m} \sum_{i=1}^{m} (u^\top x_i)^2 = \frac{1}{m} \sum_{i=1}^{m} u^\top (x_i x_i^\top) u
\]

- Optimization problem: maximize \( \sum_{u \in \mathbb{R}^d: u^\top u = 1} \sum_{i=1}^{m} u^\top (x_i x_i^\top) u \)
- Looks familiar? PCA!
- Maximized by eigenvector \( \hat{u}_1 \) of \( A = XX^\top \) corresponding to top eigenvalue \( \lambda_1 \)
- What about \( \hat{u}_2 \)?
- Subtract off data component spanned by \( \hat{u}_1 \) and repeat.
- PCA \( \equiv \) choosing the dimensions that maximize sample variance.
PCA computational complexity

- Computing $A = \sum_{i=1}^{m} x_i x_i^\top$ costs $O(md^2)$ time
- Diagonalizing $A = V \Lambda V^\top$ costs $O(d^3)$ time
- What if $d \gg m$?
- Write $A = X^\top X$, where $X = [x_1^\top; x_2^\top; \ldots; x_m^\top] \in \mathbb{R}^{m \times d}$
- Put $B = XX^\top \in \mathbb{R}^{m \times m}$; then $B_{ij} = \langle x_i, x_j \rangle$
- Diagonalize $B$ instead of $A$: get $B$’s eigenvalues and eigenvectors.
- Suppose $Bu = \lambda u$ for some $u \in \mathbb{R}^m$, $\lambda \in \mathbb{R}$
- then $A(X^\top u) = X^\top XX^\top u = \lambda X^\top u$.
- hence, $X^\top u$ is an eigenvector of $A$ with eigenvalue $\lambda$
- Conclusion: can diagonalize $B$ instead of $A$
- Pay $O(m^2d)$ time to compute $B$ and $O(m^3)$ time to diagonalize it

•
PCA and generalization

- PCA often used as a pre-processing step to supervised learning (e.g., SVM)
- Q: How to choose target dimension $k$? (At least in supervised setting.)

**Theorem**

Suppose $S = \{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}, i \leq m\}$, is drawn from $\mathcal{D}^m$. Consider hypothesis space $\mathcal{H}$ consisting of all hyperplanes $w \in \mathbb{R}^n$, and define hinge loss $\ell(y, y') = \max\{0, 1 - yy'\}$.

If for $k \leq d$ can run PCA on the sample and get distortion $\eta := \sum_{i=k+1}^{d} \lambda_i$. Then w/high prob., for all $w \in \mathcal{H}$

$$\text{risk}_\ell(w, \mathcal{D}) \leq \text{risk}_\ell(w, S) + O\left(\sqrt{\frac{k}{m}} + \sqrt{\frac{\eta}{m}}\right).$$

- bound does not depend on full dimension $d$!
- bound holds even without running PCA (sufficient for $w, k, \eta$ as above to exist)
- result suggests reasonable values for $k$
- other analyses use random matrix theory
PCA summary

- Unsupervised learning technique
- Often, pre-processing step for supervised; has denoising effect
- Performs dimensionality reduction from dim $d$ to dim $k < d$
- Optimization problem: find rank-$k$ orthogonal projection $UU^\top$ that minimizes mean square distortion on data $\sum_{i=1}^{m} \|x_i - UU^\top x_i\|_2^2$
- Solution: define data matrix $A = \sum_{i=1}^{m} x_i x_i^\top$, diagonalize it, choose $U = [u_1, \ldots, u_k]$ to be the top $k$ eigenvectors of $A$
- Equivalently: find $k$ orthogonal directions that capture most of the data variance
- Computational cost:
  - $O(md^2 + d^3)$ if $m \gg d$
  - $O(m^2d + m^3)$ if $d \gg m$
- Generalization (in some cases) depends only on $k$ and distortion but not on $d$