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 do we want to do with this unlabeled data?
  - 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

- Why do we care about structure in data?
- Computational: “simpler” data is faster to search/store/process
- Statistical: “simpler” data allows for better generalization
- Sometimes the structure is the goal (e.g. sorting photos into albums).
- Common theme: learning of a high-dimensional signal is possible when its intrinsic structure is low-dimensional
- 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]
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
  \[ f(U, V) = \sum_{i=1}^{m} \| x_i - UVx_i \|_2^2 \]
- $V : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is “compressor”, $U : \mathbb{R}^k \rightarrow \mathbb{R}^d$ is “decompressor”
- Is $f : \mathbb{R}^{d \times k} \times \mathbb{R}^{k \times d} \rightarrow \mathbb{R}$ convex?
- No! $f(U, \cdot)$ and $f(\cdot, V)$ both convex, but not $f(\cdot, \cdot)$
Principal Components Analysis: auxiliary claim

- Optimization problem: minimize $\sum_{i=1}^{m} \|x_i - UVx_i\|_2^2$
  
- **Claim:** There is an optimal solution with $U = V^\top$ and $U^\top U = I$.

- **Proof:**
  - Let $U, V$ be an optimal solution.
  - Linear map $x \mapsto UVx$ 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 $x$ there is $z \in \mathbb{R}^k$ such that $UVx = Wz$.
  - Note: $W^\top W = I$.
  - So for each $x$ there is a $z$ s.t. $\|x - UVx\|_2^2 = \|x - Wz\|_2^2$.
  - Claim: $\|x - UVx\|_2^2 \geq \|x - WW^\top x\|_2^2$.
  - **Proof:**
    - Define $f(u) := \|x - Wu\|_2^2$. Then $f(z) = \|x - UVx\|_2^2$.
    - Minimize $f(u)$: $\nabla f = 2W^\top(x - W) = 2W^\top x - 2u = 0$
      $\implies u^* = W^\top x$.
    - So for all $x$, $\|x - WW^\top x\|_2^2 \leq f(W^\top x) \leq f(z) \equiv \|x - UVx\|_2^2$.
  - Therefore $\sum_{i=1}^{m} \|x_i - UVx_i\|_2^2 \geq \sum_{i=1}^{m} \|x_i - WW^\top x_i\|_2^2$.
  - $U, V$ are optimal, so this holds with equality.
  - So instead of $U, V$ can take $W, W^\top$. □
PCA: reformulated

- $WW^\top x$ is the **orthogonal projection** of $x$ onto $R$. [board]
- PCA equivalent optimization problem:

$$\min_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \sum_{i=1}^{m} \|x_i - UU^\top x_i\|^2$$

- We show that this problem is related to the **trace** operator. Recall:
  - Trace is defined square matrices $A$: 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
- For every $x \in \mathbb{R}^d$ and $U \in \mathbb{R}^{d \times k}$ with $U^\top U = I$,

$$\|x - UU^\top x\|^2 = \|x\|^2 - 2x^\top UU^\top x + x^\top UU^\top UU^\top x = \|x\|^2 - x^\top UU^\top x = \|x\|^2 - \text{trace}(x^\top UU^\top x) = \|x\|^2 - \text{trace}(U^\top xx^\top U).$$

- So PCA is equivalent to: maximize $\max_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \text{trace} \left( U^\top \sum_{i=1}^{m} x_i x_i^\top U \right).$
Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric, $A = A^\top$
and $A\mathbf{v} = \lambda \mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$
We say that $\mathbf{v}$ is an eigenvector of $A$ with eigenvalue $\lambda$
Fact: for symmetric $A$, eigenvectors of distinct eigenvalues are orthogonal. Proof:
\begin{itemize}
  \item if $A\mathbf{v} = \lambda \mathbf{v}$, $A\mathbf{u} = \mu \mathbf{u}$, $A = A^\top$
  \item then $\mathbf{u}^\top A \mathbf{v} = \lambda \langle \mathbf{u}, \mathbf{v} \rangle = \mu \langle \mathbf{u}, \mathbf{v} \rangle$
  \item hence, if $\lambda \neq \mu$ then $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.
\end{itemize}

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 $A$

$V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ and $\mathbf{v}_i$ is the eigenvector of $A$ corresponding to $\lambda_i$

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

- Let $A := \sum_{i=1}^{m} x_i x_i^\top$.
- Diagonalize: let orthogonal $V$ and diagonal $\Lambda$ s.t. $A = V \Lambda V^\top$.
  \[
  \begin{align*}
  V^\top V &= VV^\top = I \\
  \Lambda &= \text{diag}(\lambda_1, \ldots, \lambda_d), \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0
  \end{align*}
  \]
- 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 $U \in \mathbb{R}^{d \times k}: U^\top U = I$ trace $(U^\top V \Lambda V^\top U)$.
- We now show an upper bound on this value.
- Fix any $U \in \mathbb{R}^{d \times k}$ with $U^\top U = I_k$ and put $B = V^\top U$. Note $B \in \mathbb{R}^{d \times k}$.
- then $U^\top V \Lambda V^\top U = B^\top B$.
- Hence, objective is equal to $\text{trace}(\Lambda B B^\top) = \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{k} (B_{ji})^2$.
- Define $\beta_j = \sum_{i=1}^{k} (B_{ji})^2$. (squared norm of row $j$)
- So objective is equal to $\sum_{j=1}^{d} \lambda_j \beta_j$.
- What are the values for $\beta_j$ that maximize the objective?
- $\sum_{j=1}^{d} \beta_j = B B^\top = \text{trace}(B B^\top)$.
- What is $\text{trace}(B B^\top)$?
- $B^\top B = U^\top V V^\top U = U^\top U = I_k$.
- Therefore $\text{trace}(B^\top B) = \text{trace}(I_k) = k$. So $\sum_{j=1}^{d} \beta_j = k$.

Continue on next slide...
Maximizing the trace by $k$ top eigenvalues

- We showed that the value of any solution can as $\text{trace}(\Lambda B B^\top)$, where $B \in \mathbb{R}^{d \times k}$ has orthonormal columns.
- We showed that for $\beta_j := \sum_{i=1}^{k} (B_{ji})^2$, the value of the objective is $\sum_{j=1}^{d} \lambda_j \beta_j$.
- and that $\sum_{j=1}^{d} \beta_j = k$.

Claim: $\beta_j \leq 1$. Proof:
  
  ▶ Extend $B$ with $d - k$ orthonormal columns to get an orthogonal matrix $\tilde{B} \in \mathbb{R}^{d \times d}$.
  ▶ So $\tilde{B} \tilde{B}^\top = I_d$.
  ▶ Hence $\forall j \in [d], \sum_{i=1}^{d} (\tilde{B}_{ji})^2 = 1$.
  ▶ Therefore $\sum_{i=1}^{k} (B_{ji})^2 \leq 1$.

- Upper bound for objective value: the maximal value for $\sum_{j=1}^{d} \lambda_j \beta_j$, under the constraints:
  
  ▶ $\sum_{j=1}^{d} \beta_j = k$
  ▶ $0 \leq \beta_j \leq 1$

- Maximal solution: set $\beta_1 = \beta_2 = \ldots = \beta_k = 1$. 

Kontorovich and Sabato (BGU) Lecture 14
Conclusion: The optimal solution to the PCA optimization problem has a value of

\[
\text{trace} \left( U^\top V \Lambda V^\top U \right) = \sum_{j=1}^{d} \lambda_j \beta_j \leq \sum_{j=1}^{k} \lambda_j.
\]

Claim: the maximal solution gets this with equality.

- Recall: \( A = V \Lambda V^\top \).
- Let \( \hat{u}_i \) be the \( i \)'th eigenvector of \( A \), corresponding to \( \lambda_i \).
- So \( A\hat{u}_i = \lambda_i \hat{u}_i \).
- Set the solution \( \hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k] \).
- 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 $\sum_{i=1}^{m} \| x_i - UU^T x_i \|^2$.  
  We showed, for $A = \sum_{i=1}^{m} x_i x_i^T$, 
  $$\argmin_{U \in \mathbb{R}^{d \times k} : U^T U = I} \sum_{i=1}^{m} \| x_i - UU^T x_i \|^2 = \argmax_{U \in \mathbb{R}^{d \times k} : U^T U = I} \text{trace}(U^T A^T U).$$

- We just showed that $\hat{U}$ maximizes $\text{trace}(U^T AU)$.

- Value of solution:
  - $\hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k]$ ($k$ top eigenvectors), $\text{trace}(\hat{U}^T \hat{A} \hat{U}) = \sum_{j=1}^{k} \lambda_j$.
  - $\sum_{i=1}^{m} \| x_i - \hat{U}\hat{U}^T x_i \|^2 = \sum_{i=1}^{m} \| x_i \|^2 - \text{trace}(\hat{U}^T \hat{A} \hat{U})$  
  - $\sum_{i=1}^{m} \| x_i \|^2 = \text{trace}(A) = \text{trace}(V\Lambda V^T) = \text{trace}(V^T V\Lambda) = \text{trace}(\Lambda) = \sum_{i=1}^{d} \lambda_i$.
  - Optimal objective value: $\sum_{i=1}^{m} \| x_i - \hat{U}\hat{U}^T x_i \|^2 = \sum_{i=k+1}^{d} \lambda_i$.

- This is the “distortion”: how much “energy” we throw away.

- The distortion is equal to the sum of the lowest $d - k$ eigenvalues.
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 $\sum_{u \in \mathbb{R}^d: u^\top u = 1} u^\top (x_i x_i^\top) u$
- 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 $B_{ij} = \langle x_i, x_j \rangle$.
- Claim: can diagonalize $B$ instead of $A$.
  - 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 $\lambda$
- 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)
- How to choose target dimension $k$ in this case?

**Theorem**

Suppose $S = \{(x_i, y_i)\}_{i=1}^m$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$, $S \sim \mathcal{D}^m$.

If we do PCA with $k$ dimensions, and $\eta_k := \sum_{i=k+1}^d \lambda_i$, then with high probability, for all $w \in \mathcal{H}_L^k$ (linear predictors)

$$\ell^h(w, \mathcal{D}) \leq \ell^h(w, S) + O \left( \sqrt{\frac{k}{m}} + \sqrt{\frac{\eta_k}{m}} \right).$$

- Bound does not depend on full dimension $d$!
- Bound holds for all $k$, no need to actually run PCA.
- Result suggests reasonable values for $k$ if want to do PCA.
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 for soft-SVM depends on $k$ and distortion, not on $d$