PCA

- We want to reduce the dimensionality of the data, i.e. find a projection.
- We want a projection with high variance.
- Assume our data is $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in R^D$.
- We will assume the mean of the data is zero, i.e. $\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = 0$. If that's not the case, subtracting the mean will result in zero-mean data.
- Let $\mathbf{u} \in R^D$ be the projection vector, i.e. this is the vector on which we'll project the data (in the 1D case). Then the projections of the data are $p_1, p_2, \ldots, p_N \in R^1$ (scalars) s.t. $p_i = \mathbf{x}_i^T \mathbf{u}$.
- We want to maximize the variance of the projection, which depends on $\mathbf{u}$: $\mathbf{u} = \underset{\mathbf{u}}{\operatorname{argmax}} \sum_i \| \mathbf{x}_i^T \mathbf{u} - \mu \|^2$ where μ is the mean of the projections. We will also require that $\| \mathbf{u} \| = 1$, i.e. $\mathbf{u}$ is unit-length.
- Some norms: $\| \mathbf{u} \|_1 = \sum_{j=1}^{d} |u_j|$ a.k.a. L1 norm

$$\| \mathbf{u} \|_2 = \sqrt{\mathbf{u}^T \mathbf{u}} = \sqrt{\sum_{j=1}^{d} u_j^2}$$ a.k.a. L2 norm

- We have: $\mu = \frac{1}{N}\sum_{i=1}^{N} p_i = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^T \mathbf{u} = \frac{1}{N}\left(\sum_{i=1}^{N} \mathbf{x}_i\right)^T \mathbf{u}$
  $= 0$ (because the data is zero-mean).

- Then $\mathbf{u} = \underset{\mathbf{u}}{\operatorname{argmax}} \sum_i \| \mathbf{x}_i^T \mathbf{u} \|^2$ subject to $\|\mathbf{u}\| = 1$.

- In matrix notation, we have:
  $\mathbf{u} = \underset{\mathbf{u}}{\operatorname{argmax}} \| \mathbf{X}\mathbf{u} \|^2$ subject to $\mathbf{u}^T\mathbf{u} = 1$ where
  $\mathbf{X}$ is the $N \times D$ data matrix, and $\mathbf{X}\mathbf{u}$ is the $N \times 1$ vector of projections.

- To write both the objective and constraint with one expression, we will use a "Lagrangian". If we have a goal to maximize $f(x)$ subject to $g(x) = 0$, we formulate a Lagrangian as $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$, and take its derivative with respect to (wrt) x to find the solution x that maximizes the objective subject to the constraint.

- For PCA, we get this Lagrangian:
  $\mathcal{L}(\mathbf{u}, \lambda) = \| \mathbf{X}\mathbf{u} \|^2 + \lambda(\mathbf{u}^T\mathbf{u} - 1)$.

- Taking a derivative wrt $\mathbf{u}$:
  $$\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = \frac{\partial[(\mathbf{X}\mathbf{u})^T \mathbf{X}\mathbf{u} + \lambda \mathbf{u}^T\mathbf{u} - \lambda]}{\partial \mathbf{u}} = 2\mathbf{X}^T\mathbf{X}\mathbf{u} + 2\lambda\mathbf{u}$$
  (using the chain rule and matrix derivatives - see next page)

An Aside: Matrix Calculus, Useful Formulas

If y is ... ,	then $\frac{\partial y}{\partial x}$ is ...
$\mathbf{A}x$	$\mathbf{A}^T$
$x^T\mathbf{A}$	$\mathbf{A}$
x^Tx	$2x$

Chain rule for $\frac{\partial z}{\partial x}$ where $z = f(y)$, $y = g(x)$ i.e. $z = f(g(x))$:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} * \frac{\partial y}{\partial x} \quad \text{(scalars)} \qquad \frac{\partial z}{\partial x} = \frac{\partial y}{\partial x}\frac{\partial z}{\partial y} \quad \text{(matrices)}$$

Back to PCA:

- We set the derivative to 0 to find the optimal u, i.e.

$$\frac{\partial L}{\partial u} = 0 \Rightarrow (X^TX)u = (-\lambda)u$$ which is an <u>eigenvalue problem.</u>

(There are packages available to solve it.)

- Note that X^TX is called the "data covariance matrix":

$$[X^TX]_{ij} = \sum_n x_{ni}x_{nj}$$

- In conclusion, the optimal projection (leading to maximum variance) is some eigenvector of the data covariance matrix. Below we'll show why that's the <u>first</u> eigenvector (corresponding to the largest eigenvalue).

- How to find a second projection direction?

$$\mathbf{v} = \arg\max_{\mathbf{v}} \|\mathbf{X}\mathbf{v}\|^2 \quad \text{subject to} \quad \mathbf{v}^T\mathbf{v} = 1, \ \mathbf{v}^T\mathbf{u} = 0$$

($\mathbf{v}^T\mathbf{u} = 0$ is required to ensure non-redundancy).

- The second projection direction $\mathbf{v}$ corresponds to the second eigenvector of the data covariance matrix.

Why best projection eigenvector corresponds to the largest eigenvalue:
Remember that we want to maximize the variance of the projected data, i.e. $\|\mathbf{X}\mathbf{u}\|^2 = (\mathbf{X}\mathbf{u})^T\mathbf{X}\mathbf{u} = \mathbf{u}^T\underbrace{\mathbf{X}^T\mathbf{X}\mathbf{u}}_{=-\pi\mathbf{u}} = \underbrace{(-\pi)}\underbrace{\mathbf{u}^T\mathbf{u}}_{=1}$, i.e. the variance, when the projection is $\mathbf{u}$, is the eigenvalue corresponding to the eigenvector $\mathbf{u}$. Thus, to obtain the largest possible variance of the projected data, we should use the eigenvector corresponding to the largest eigenvalue.

Are we guaranteed that the condition $g(x)=0$ is met when we use a Lagrangian multiplier?

No. First, we have to solve both $\frac{\partial L}{\partial x} = 0$ and $\frac{\partial L}{\partial \pi} = 0$.
Second, the Lagrangian formulation only tells us what the necessary conditions are to solve are problem, but the sufficient conditions might be different.