Week 9: Dimensionality Reduction

Instructor: Sergey Levine

1 Dimensionality Reduction Recap

In the last lecture, we saw how we could construct a probabilistic model for dimensionality reduction. In this model, as with all other generative probabilistic models, our objective is to maximize the log-likelihood of the data:

$$\log p(\mathcal{D}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i),$$

where the data is $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ (so we're doing unsupervised learning – no labels), and the particular probabilistic model we use is a Gaussian model $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{U}\mathbf{z} + \mu_0, \sigma^2)$ ($\mu_0 = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$), where we model $\mathbf{x}$ as originating from a Gaussian distribution with the mean given by $\mathbf{U}\mathbf{z}$, where $\mathbf{z}$ is the low-dimensional point that corresponds to $\mathbf{x}$, and $\mathbf{U}$ is our basis (which we wish to learn).

This model basically says that any information in the datapoint $\mathbf{x}$ that is not modeled by $\mathbf{U}\mathbf{z} + \mu_0$ is the consequence of Gaussian noise. The resulting objective can be written out as

$$\log p(\mathcal{D}) = \sum_{i=1}^{N} -\frac{1}{2}||\mathbf{U}\mathbf{z}_i - \bar{\mathbf{x}}_i||^2 + \text{const},$$

which means that our goal is to find a basis $\mathbf{U}$ that allows us to reconstruct the high-dimensional datapoints $\mathbf{x}_i$ from their low-dimensional features $\mathbf{z}_i$ with the minimum error: that is, we wish to retain as much information as possible about the original datapoints.

When defining our hypothesis space, we also constrained the columns of $\mathbf{U}$, which we call $\mathbf{u}_1, \dots, \mathbf{u}_K$, to be orthonormal, such that $\mathbf{u}_i^T\mathbf{u}_i = 1$ and $\mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$. We saw that these constraints do not in any way limit the information about $\mathbf{x}$ that we can pack into $\mathbf{z}$, so it is fine to impose these constraints without loss of information (without increasing the error/decreasing the probability).

When $\mathbf{U}$ is orthonormal, we know that $\mathbf{z}_i = \mathbf{U}^T(\mathbf{x}_i - \mu_0)$, and by using $\bar{\mathbf{x}}_i = \mathbf{x}_i - \mu_0$ for convenience, we showed that we can rewrite the maximum likelihood optimization of $\mathbf{U}$ as

$$\mathbf{U} \leftarrow \arg\min_{\mathbf{U}} \frac{1}{2N}\sum_{i=1}^{N} ||\mathbf{U}\mathbf{z}_i - \bar{\mathbf{x}}_i||^2 \text{ such that } \mathbf{u}_i^T\mathbf{u}_i = 1 \text{ and } \mathbf{u}_i^T\mathbf{u}_j = 0 \;\; \forall i \neq j,$$

which reduces to

$$\mathbf{U} \leftarrow \arg\min_{\mathbf{U}} -\frac{1}{2}\sum_{k=1}^{K} \mathbf{u}_k^T \Sigma \mathbf{u}_k \text{ such that } \mathbf{u}_i^T \mathbf{u}_i = 1 \text{ and } \mathbf{u}_i^T \mathbf{u}_j = 0 \ \ \forall i \neq j,$$

where $\Sigma = \frac{1}{N}\sum_{i=1}^{N} \bar{\mathbf{x}}_i$. This simplification follows directly from substituting $\mathbf{z}_i = \mathbf{U}^T \bar{\mathbf{x}}_i$ and expanding the square $||\cdot||^2$, following by the identity

$$\bar{\mathbf{x}}_i^T \mathbf{U}\mathbf{U}^T \bar{\mathbf{x}}_i = \sum_{k=1}^{K} \mathbf{u}_k^T \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{u}_k.$$

To solve this optimization problem, first, let's see what happens in the simple case where $K = 1$ and $\mathbf{U}$ has just one column. Add in the Lagrange multipliers (we multiply by $\frac{1}{2}$ for convenience):

$$\mathcal{L}(\mathbf{u}, \lambda) = -\frac{1}{2}\mathbf{u}^T \Sigma \mathbf{u} + \frac{1}{2}\lambda(\mathbf{u}^T \mathbf{u} - 1).$$

Now take the derivative:

$$\frac{d\mathcal{L}}{d\mathbf{u}} = -\Sigma \mathbf{u} + \lambda \mathbf{u} = 0 \Rightarrow \Sigma \mathbf{u} = \lambda \mathbf{u}$$

That's interesting! It's not immediately straightforward to solve for $\mathbf{u}$, until we recognize that this is exactly the definition of eigenvectors and eigenvalues! So $\mathbf{u}$ must be an eigenvector of Σ, and λ must be an eigenvalue. All eigenvectors $\mathbf{u}$ will satisfy the constraint, so if we substitute the solution into the objective, we get

$$\min_{\mathbf{u}} -\frac{1}{2}\mathbf{u}^T \Sigma \mathbf{u} = \min_{\mathbf{u}} -\frac{1}{2}\mathbf{u}^T \mathbf{u} \lambda = \min_{\mathbf{u}} -\frac{1}{2}\lambda,$$

where the last step follows from the fact that $\mathbf{u}^T \mathbf{u} = 1$. So our goal is simply to maximize the eigenvalue that corresponds to the eignevector $\mathbf{u}$! Therefore, we have only one basis vector, it should be the eigenvector of Σ that corresponds to the largest eigenvalue.

What happens if we have more than one basis vector ($K > 1$)? Well, we could repeat the same exercise with Lagrange multipliers, but we could observe that, were it not for the orthogonality constraints $\mathbf{u}_i^T \mathbf{u}_j = 0$ for $i \neq j$, the rest of the Lagrangian factorizes additively (that is, all $\mathbf{u}_k$ vectors are independent), so we always have

$$\Sigma \mathbf{u}_k = \lambda_k \mathbf{u}_k,$$

and we always want to maximize the corresponding eigenvalues λ_k, since we have

$$-\frac{1}{2}\sum_{k=1}^{K} \mathbf{u}_k^T \Sigma \mathbf{u}_k = -\frac{1}{2}\sum_{k=1}^{K} \lambda_k.$$

Therefore, since all $\mathbf{u}_k$ vectors must be orthogonal, they must all be *different* eigenvectors, and we should pick the ones that correspond to the K *largest* eigenvalues. We therefore recover the simple algorithm for obtaining the optimal basis to maximize $\log p(\mathcal{D})$: compute an eigenvalue decomposition of the empirical covariance

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mu_0)(\mathbf{x}_i - \mu_0)^T,$$

and then populate the columns of $\mathbf{U}$ with the eigenvectors of Σ corresponding to the K largest eigenvalues. This is called principal component analysis (PCA).

2 Analyzing the Error

We saw how we could figure out the best basis $\mathbf{U}$. But how do we choose the dimensionality K? Oftentimes, we might just choose K based on the constraints of our problem. For example, if we want to visualize our data in 2D or 3D, we might choose $K = 2$ or $K = 3$ to make the reduced dimension points $\mathbf{z}_i$ interpretable. However, for other applications, like preprocessing our data for supervised learning, we might want to choose K so as to retain some fraction of the information in the data. The typical unit of measure for this information is "variance." First, we'll introduce an intuitive idea of what this means, and then we'll do some math to see how we can compute it.

In the lecture slides (slide 4) we can see an example of a 2D dataset and a visualization of two principal components. The biggest principal component (eigenvector with the largest eigenvalue) points along the direction in which the data is most spread out. If we were to fit a covariance matrix Σ to this data, this direction would correspond to the longest axis of the unit variance ellipse. Thus, the direction with the largest variance is the first principal component. The second principal component is orthogonal to the first (eigenvectors are always orthogonal), and points in the direction with the second largest variance.

If we keep one of the two principal components (meaning we choose $K = 1$), the error we incur will correspond to the distance between the actual 2D points and the line given by the first principal component. The expected squared error *is* the variance, hence the error is equal to the sum of the variances along all of the principal components we *didn't* choose.

The variance along each principal component $\mathbf{u}_k$ is given by

$$E[(\mathbf{u}_k^T\bar{\mathbf{x}}_i)^2] = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{u}_k^T\bar{\mathbf{x}}_i)^2,$$

which is simply the equation for the variance of the random variable $\mathbf{u}_k^T\bar{\mathbf{x}}$, which is zero mean by construction (because $\bar{\mathbf{x}} = \mathbf{x} - \mu_0$) and corresponds to the projection of the data onto the principal component $\mathbf{u}_k$. However, note that this can be written as

$$E[(\mathbf{u}_k^T\bar{\mathbf{x}}_i)^2] = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{u}_k^T\bar{\mathbf{x}}_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_k^T\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\mathbf{u}_k = \mathbf{u}_k^T\Sigma\mathbf{u}_k.$$

This is exactly the quantity that PCA maximizes! So the goal of PCA is to find principal components that *explain* as much variance as possible in the data. Furthermore, we saw above that, since $\Sigma \mathbf{u}_k = \lambda_k \mathbf{u}_k$, we have

$$\mathbf{u}_k^T \Sigma \mathbf{u}_k = \lambda_k \mathbf{u}_k^T \mathbf{u}_k = \lambda_k,$$

so the variance along the principal component $\mathbf{u}_k$ is exactly the corresponding eigenvalue λ_k. This is another explanation for why we choose the eigenvectors that correspond to the largest eigenvalues.

To understand *how much* of the variance is explained by the first K eigenvalues, it helps to think about all the *other* eigenvectors that we don't choose. If we consider the full basis $\bar{\mathbf{U}}$, which consists of $\mathbf{U}$ followed by all the other $D-K$ eigenvectors, we know that the total variance is simply the variance along all of the eigenvectors:

$$\sum_{k=1}^{D} \bar{\mathbf{u}}_k^T \Sigma \bar{\mathbf{u}}_k = \sum_{k=1}^{K} \mathbf{u}_k^T \Sigma \mathbf{u}_k + \sum_{k=K+1}^{D} \bar{\mathbf{u}}_k^T \Sigma \bar{\mathbf{u}}_k.$$

But because this is the *total* variance, we also know that it doesn't matter which orthonormal basis $\bar{\mathbf{U}}$ we select, since any orthonormal basis with D dimensions spans the entire space. In particular, we know that

$$\sum_{k=1}^{D} \bar{\mathbf{u}}_k^T \Sigma \bar{\mathbf{u}}_k = \sum_{k=1}^{D} \mathbf{e}_k^T \Sigma \mathbf{e}_k,$$

where $\mathbf{e}_k$ is a vector that is 0 in all entries except k, which is 1 (the k^{th} canonical vector). Using our previous identity, we know that

$$\sum_{k=1}^{D} \mathbf{e}_k^T \Sigma \mathbf{e}_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{D} \mathbf{e}_k^T \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^T \mathbf{e}_k = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{x}}_i^T \mathbf{I}\mathbf{I}^T \bar{\mathbf{x}}_i = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_i = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{D} \bar{\mathbf{x}}_{i,k}^2,$$

which incidentally is the equation for the total variance of all of the dimensions. So if we want to know what portions of the total variance is explained by the first K eigenvectors, we can compute this as

$$\text{explained variance} = \frac{\sum_{k=1}^{K} \mathbf{u}_k^T \Sigma \mathbf{u}_k}{\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{D} \bar{\mathbf{x}}_{i,k}^2}.$$

The derivation above will also be useful for a problem in homework 4. But for choosing K, what we might do is choose a target explained variance (e.g. 90%), and pick the K that makes the above ratio greater than our target.

3 Singular Value Decomposition

Computing a full eigenvalue decomposition of the covariance matrix Σ can be used to obtain the K largest eigenvectors and perform PCA. But in practice, a more efficient way to perform PCA is to use something called singular value decomposition (SVD), which does not require even forming the full covariance Σ. This can be particularly useful if the dimensionality of $\mathbf{x}$ is extremely large, making an eigenvalue decomposition expensive. For example, $\mathbf{x}$ might represent an image with millions of pixels, or a 3D mesh with millions of vertices.

SVD generalizes the idea of eigenvalues to non-square matrices, and is based on the idea that any $N \times D$ matrix $\mathbf{X}$ can be decomposed according to

$$\underbrace{\mathbf{X}}_{N\times D} = \underbrace{\mathbf{W}}_{N\times N}\underbrace{\mathbf{S}}_{N\times D}\underbrace{\mathbf{V}^T}_{D\times D}.$$

Here, $\mathbf{W}$ and $\mathbf{V}$ are both orthonormal, and $\mathbf{S}$ is a diagonal matrix, with the off-diagonal entries all zero. If $N \neq D$, which is true in general, $\mathbf{S}$ is not square, so it will be padded with 0 on the bottom or on the right. The diagonal entries of $\mathbf{S}$ are called the singular values $\sigma_k \geq 0$, and there are $\min(N, D)$ such values.

In the case of PCA, we'll choose $\mathbf{X}$ to be a matrix where each row is a datapoint $\bar{\mathbf{x}}_i$, such that, and the whole matrix is divide by $\sqrt{N}$ (we'll see why in a second!):

$$\mathbf{X} = \frac{1}{\sqrt{N}}\begin{bmatrix} \bar{\mathbf{x}}_1^T \\ \bar{\mathbf{x}}_2^T \\ \cdot \\ \cdot \\ \bar{\mathbf{x}}_N^T \end{bmatrix}$$

In that case, $\mathbf{X}^T\mathbf{X} = \frac{1}{N}\sum_{i=1}^N \bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T$ by the definition of matrix multiplication, and we have

$$\Sigma = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^T\mathbf{W}^T\mathbf{W}\mathbf{S}\mathbf{V}^T.$$

Recall, however, that both $\mathbf{W}$ and $\mathbf{V}$ are orthonormal, so we can multiply both sides by $\mathbf{V}$ on the right side and simplify to get:

$$\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \mathbf{V}\mathbf{S}^T\mathbf{W}^T\mathbf{W}\mathbf{S}\mathbf{V}^T \\
\mathbf{X}^T\mathbf{X}\mathbf{V} &= \mathbf{V}\mathbf{S}^T\mathbf{W}^T\mathbf{W}\mathbf{S}\mathbf{V}^T\mathbf{V} \\
\mathbf{X}^T\mathbf{X}\mathbf{V} &= \mathbf{V}\mathbf{S}^T\mathbf{W}^T\mathbf{W}\mathbf{S} \\
\mathbf{X}^T\mathbf{X}\mathbf{V} &= \mathbf{V}\mathbf{S}^T\mathbf{S}
\end{aligned}$$

Size $\mathbf{S}$ is diagonal with entries σ_k, $\mathbf{S}^T\mathbf{S}$ is also diagonal with entries σ_k^2, which we'll denote as Λ. Substituting in $\Sigma = \frac{1}{N}\mathbf{X}^T\mathbf{X}$, we get

$$\Sigma\mathbf{V} = \mathbf{V}\Lambda \Rightarrow \Sigma\mathbf{V}_k = \mathbf{V}_k\sigma_k^2.$$

We see therefore that the columns of $\mathbf{V}$ are the eigenvectors of Σ, and the eigenvalues correspond to the squared singular values. So instead of performing a full eigenvalue decomposition on Σ, we can simply take the columns of $\mathbf{V}$ that correspond to the largest singular values σ_k.