Cross Validated

What norm of the reconstruction error is minimized by the low-rank approximation matrix obtained with PCA?

Asked 9 years, 2 months ago Modified 3 years, 6 months ago Viewed 14k times

28

Given a PCA (or SVD) approximation of matrix X with a matrix $\hat{X}$, we know that $\hat{X}$ is the best low-rank approximation of X.

Is this according to the induced $||\cdot||_2$ norm (i.e. the largest eigenvalue norm) or according to the Frobenius $||\cdot||_F$ norm ?

pca svd matrix-decomposition

Share Cite

Improve this question Follow

edited Sep 28, 2015 at 13:13
amoeba
104k 35 311 338

asked Dec 30, 2014 at 14:12
Donbeo
3,129 5 35 50

1 Answer

Sorted by: Highest score (default)

39

Single word answer: Both.

Let's start with defining the norms. For a matrix X, operator 2-norm is defined as

$$\|X\|_2 = \sup \frac{\|Xv\|_2}{\|v\|_2} = \max(s_i)$$

and Frobenius norm as

$$\|X\|_F = \sqrt{\sum_{ij} X_{ij}^2} = \sqrt{\mathrm{tr}(X^\top X)} = \sqrt{\sum s_i^2},$$

where s_i are singular values of X, i.e. diagonal elements of S in the singular value decomposition $X = USV^\top$.

PCA is given by the same singular value decomposition when the data are centered. US are principal components, V are principal axes, i.e. eigenvectors of the covariance matrix, and the reconstruction of X with only the k principal components corresponding to the k largest singular values is given by $X_k = U_k S_k V_k^\top$.

The **Eckart-Young theorem** says that X_k is the matrix minimizing the norm of the reconstruction error $\|X - A\|$ among all matrices A of rank k. This is true for both, Frobenius norm and the operator 2-norm. As pointed out by @cardinal in the comments, it was first proved by Schmidt (of Gram-Schmidt fame) in 1907 for the Frobenius case. It was later rediscovered by Eckart and Young in 1936 and is now mostly associated with their names. Mirsky generalized the theorem in 1958 to all norms that are invariant under unitary transformations, and this includes the operator 2-norm.

This theorem is sometimes called Eckart-Young-Mirsky theorem. Stewart (1993) calls it Schmidt approximation theorem. I have even seen it called Schmidt-Eckart-Young-Mirsky theorem.

- Eckart and Young, 1936, The approximation of one matrix by another of lower rank
- Mirsky, 1958, Symmetric gauge functions and unitarily invariant norms
- Stewart, 1993, On the early history of the singular value decomposition

Proof for the operator 2-norm

Let X be of full rank n. As A is of rank k, its null space has $n - k$ dimensions. The space spanned by the $k + 1$ right singular vectors of X corresponding to the largest singular values has $k + 1$ dimensions. So these two spaces must intersect. Let w be a unit vector from the intersection. Then we get:

$$\|X - A\|_2^2 \geq \|(X - A)w\|_2^2 = \|Xw\|_2^2 = \sum_{i=1}^{k+1} s_i^2 (v_i^\top w)^2 \geq s_{k+1}^2 = \|X - X_k\|_2^2,$$

QED.

Proof for the Frobenius norm

We want to find matrix A of rank k that minimizes $\|X - A\|_F^2$. We can factorize $A = BW^\top$, where W has k orthonormal columns. Minimizing $\|X - BW^\top\|^2$ for fixed W is a regression problem with solution $B = XW$. Plugging it in, we see that we now need to minimize

$$\begin{aligned}\|X - XWW^\top\|^2 = \|X\|^2 - \|XWW^\top\|^2 &= \mathrm{const} - \mathrm{tr}(WW^\top X^\top XWW^\top) \\ &= \mathrm{const} - \mathrm{const} \cdot \mathrm{tr}(W^\top \Sigma W),\end{aligned}$$

where Σ is the covariance matrix of X, i.e. $\Sigma = X^\top X/(n-1)$. This means that reconstruction error is minimized by taking as columns of W some k orthonormal vectors maximizing the total variance of the projection.

It is well-known that these are first k eigenvectors of the covariance matrix. Indeed, if $X = USV^\top$, then $\Sigma = VS^2V^\top/(n-1) = V\Lambda V^\top$. Writing $R = V^\top W$ which also has orthonormal columns, we get

$$\mathrm{tr}(W^\top \Sigma W) = \mathrm{tr}(R^\top \Lambda R) = \sum_i \lambda_i \sum_j R_{ij}^2 \le \sum_{i=1}^k \lambda_k,$$

with maximum achieved when $W = V_k$. The theorem then follows immediately.

See the following three related threads:

- What is the objective function of PCA?
- Why does PCA maximize total variance of the projection?
- PCA objective function: what is the connection between maximizing variance and minimizing error?

Earlier attempt of a proof for Frobenius norm

This proof I found somewhere online but it is wrong (contains a gap), as explained by @cardinal in the comments.

Frobenius norm is invariant under unitary transformations, because they do not change the singular values. So we get:

$$\|X - A\|_F = \|USV^\top - A\| = \|S - U^\top AV\| = \|S - B\|,$$

where $B = U^\top A V$. Continuing:

$$\|X - A\|_F = \sum_{ij}(S_{ij} - B_{ij})^2 = \sum_i (s_i - B_{ii})^2 + \sum_{i \neq j} B_{ij}^2.$$

This is minimized when all off-diagonal elements of B are zero and all k diagonal terms cancel out the k largest singular values s_i **[gap here: this is not obvious]**, i.e. $B_{\text{optimal}} = S_k$ and hence $A_{\text{optimal}} = U_k S_k V_k^\top$.

Share Cite Improve this answer

Follow

edited Aug 24, 2020 at 16:38

Luke Burns

103 2

answered Dec 30, 2014 at 16:06

amoeba

104k 35 311 338

3 The proof in the case of the Frobeniius norm is not correct (or at least complete) since the argument here does not preclude the possibility that a matrix of the same rank could cancel out some of the other diagonal terms while having "small" off-diagonals. To see the gap more clearly note that holding the diagonals constant and "zeroing" the off-diagonals can often *increase* the rank of the matrix in question! – cardinal Dec 30, 2014 at 18:59

1 Note also that the SVD was known to Beltrami (at least in a quite general, though special case) and Jordan as early as 1874. – cardinal Dec 30, 2014 at 19:01

@cardinal: Hmmmm, I am not sure I see the gap. If B cancels out some other diagonal terms in S instead of k largest ones and has some nonzero off-diagonal terms instead, then both sums, $\sum_i (s_i - B_{ii})^2$ and $\sum_{i \neq j} B_{ij}^2$, are going to increase. So it will only increase the reconstruction error. No? Still, I tried to find another proof for Frobenius norm in the literature, and have read that it should somehow follow easily from the operator norm case. But so far I don't see how it should follow... – amoeba Dec 31, 2014 at 0:27

3 I *do* like G. W. Stewart (1993), On the early history of the singular value decomposition, *SIAM Review*, vol. 35, no. 4, 551-566 and, given your prior demonstrated interest in historical matters, I think you will too. Unfortunately, I think Stewart is unintentionally overly dismissive of the elegance of Schmidt's 1907 proof. Hidden within it is a regression interpretation that Stewart overlooks and which is really quite pretty. There is another proof that follows the initial diagonalization approach you take, but which requires some extra work to fill the gap. (cont.) – cardinal Dec 31, 2014 at 2:05

2 @cardinal: Yes, you are right, now I see the gap too. Thanks a lot for the Stewart paper, that was a very interesting read. I see that Stewart presents Schmidt's and Weyl's proofs, but both of them look more complicated than what I would like to copy here (and so far I have not had the time to study them carefully). I am surprised: I expected this to be a very simple result, but it seems it is less trivial than I thought. In particular, I would not have expected that the Frobenius case is so much more complicated than the operator norm one. I will edit the post now. Happy New Year! – amoeba Jan 2, 2015 at 23:58