EE16B - Spring'20 - Lecture 9A Notes[1]

Murat Arcak

17 March 2020

Learning

System Identification Continued

In many applications the matrices A and B in the state model

$$\vec{x}(t+1) = A\vec{x}(t) + B\vec{u}(t) \tag{1}$$

are not known exactly and change with operating conditions. The goal in system identification is to learn these matrices by observing the input sequence and the resulting state sequence.

Last time we considered the scalar system:

$$x(t+1) = \lambda x(t) + bu(t) + e(t)$$

where e is a disturbance term. From $t = 1$ to $t = \ell$ we have

$$\begin{aligned} x(1) &= \lambda x(0) + bu(0) + e(0) \\ x(2) &= \lambda x(1) + bu(1) + e(2) \\ &\vdots \\ x(\ell) &= \lambda x(\ell-1) + bu(\ell-1) + e(\ell) \end{aligned} \tag{2}$$

which we rewrite in the following standard form for Least Squares:

$$\underbrace{\begin{bmatrix} x(0) & u(0) \\ x(1) & u(1) \\ \vdots & \cdots \\ x(\ell-1) & u(\ell-1) \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} \lambda \\ b \end{bmatrix}}_{\vec{p}} + \underbrace{\begin{bmatrix} e(0) \\ e(1) \\ \vdots \\ e(\ell-1) \end{bmatrix}}_{\vec{e}} = \underbrace{\begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(\ell) \end{bmatrix}}_{\vec{y}}. \tag{3}$$

Thus, when D^TD is invertible, we obtain the estimates $\hat{\lambda}$, $\hat{b}$ from

$$\hat{\vec{p}} = \begin{bmatrix} \hat{\lambda} \\ \hat{b} \end{bmatrix} = (D^TD)^{-1}D^T\vec{y}.$$

Now let's return to the vector case (1), with disturbance $\vec{e}(t)$ added to the right-hand side. The equations below are analogous to (2):

$$\begin{aligned} \vec{x}(1) &= A\vec{x}(0) + B\vec{u}(0) + \vec{e}(0) \\ \vec{x}(2) &= A\vec{x}(1) + B\vec{u}(1) + \vec{e}(1) \\ &\vdots \\ \vec{x}(\ell) &= A\vec{x}(\ell-1) + B\vec{u}(\ell-1) + \vec{e}(\ell-1). \end{aligned} \tag{4}$$

If we transpose these equations, we get:

$$\begin{aligned}\vec{x}(1)^T &= \vec{x}(0)^T A^T + \vec{u}(0)^T B^T + \vec{e}(0)^T \\ \vec{x}(2)^T &= \vec{x}(1)^T A^T + \vec{u}(1)^T B^T + \vec{e}(1)^T \\ &\vdots \\ \vec{x}(\ell)^T &= \vec{x}(\ell-1)^T A^T + \vec{u}(\ell-1)^T B^T + \vec{e}(\ell-1)^T,\end{aligned} \tag{5}$$

which we can rewrite in a form similar to (3):

$$\underbrace{\begin{bmatrix} \vec{x}(0)^T & \vec{u}(0)^T \\ \vec{x}(1)^T & \vec{u}(1)^T \\ \vdots & \cdots \\ \vec{x}(\ell-1)^T & \vec{u}(\ell-1)^T \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} A^T \\ B^T \end{bmatrix}}_{[\vec{p}_1 \cdots \vec{p}_n]} + \underbrace{\begin{bmatrix} \vec{e}(0)^T \\ \vec{e}(1)^T \\ \vdots \\ \vec{e}(\ell-1)^T \end{bmatrix}}_{[\vec{e}_1 \cdots \vec{e}_n]} = \underbrace{\begin{bmatrix} \vec{x}(1)^T \\ \vec{x}(2)^T \\ \vdots \\ \vec{x}(\ell)^T \end{bmatrix}}_{[\vec{y}_1 \cdots \vec{y}_n]}. \tag{6}$$

Note that the unknowns and measurements are now contained in matrices with n columns:

$$\begin{bmatrix} A^T \\ B^T \end{bmatrix} =: \begin{bmatrix} \vec{p}_1 & \cdots & \vec{p}_n \end{bmatrix} \quad \begin{bmatrix} \vec{x}(1)^T \\ \vec{y}(x)^T \\ \vdots \\ \vec{x}(\ell)^T \end{bmatrix} = \begin{bmatrix} x_1(1) & \cdots & x_n(1) \\ x_1(2) & \cdots & x_n(2) \\ \vdots & & \vdots \\ x_1(\ell) & \cdots & x_n(\ell) \end{bmatrix} =: \begin{bmatrix} \vec{y}_1 & \cdots & \vec{y}_n \end{bmatrix}.$$

Thus, we can separate (6) into n separate equations

$$D\vec{p}_i + \vec{e}_i = \vec{y}_i, \quad i = 1, 2, \ldots, n$$

and apply Least Squares to each one independently:

$$\hat{\vec{p}}_i = (D^T D)^{-1} D^T \vec{y}_i.$$

Note that $\vec{y}_i$ here is a column consisting of measurements of the ith state variable collected from $t = 1$ to $t = \ell$, and $\hat{\vec{p}}_i$ is our estimate for the ith column of A^T concatenated with the ith column of B^T.

Singular Value Decomposition (SVD)

SVD separates a rank-r matrix $A \in \mathbb{R}^{m \times n}$ into a sum of r rank-1 matrices, each written as a column times row. Specifically, we can find:

1) orthonormal vectors $\vec{u}_1, \ldots, \vec{u}_r \in \mathbb{R}^m$,

2) orthonormal vectors $\vec{v}_1, \ldots, \vec{v}_r \in \mathbb{R}^n$,

3) real, positive numbers $\sigma_1, \ldots, \sigma_r$ such that

$$\boxed{A = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \cdots + \sigma_r \vec{u}_r \vec{v}_r^T.} \tag{7}$$

The numbers $\sigma_1, \ldots, \sigma_r$ are called *singular values* and, by convention, we order them from the largest to smallest:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

In its original form A has mn entries to be stored. In the SVD form each of the r terms is the product of a column of m entries with a row of n entries; therefore we need $r(m+n)$ numbers to store. This is an advantage when r is small relative to m and n, that is $r(m+n) \ll mn$.

In a typical application the exact rank r may not be particularly small, but we may find that the first few singular values, say $\sigma_1, \ldots, \sigma_{\hat{r}}$, are much bigger than the rest, $\sigma_{\hat{r}+1}, \ldots, \sigma_r$. Then it is reasonable to discard the small singular values and approximate A as

$$A \approx \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \cdots + \sigma_{\hat{r}} \vec{u}_{\hat{r}} \vec{v}_{\hat{r}}^T \tag{8}$$

which has rank $= \hat{r}$, thus $\hat{r}(m+n) \ll mn$ numbers to store.

Besides enabling data compression, SVD allows us to extract important features of a data set as illustrated in the next example.

Example (Netflix): Suppose we have a $m \times n$ matrix that contains the ratings of m viewers for n movies. A truncated SVD as suggested above not only saves memory; it also gives insight into the preferences of each viewer. For example we can interpret each rank-1 matrix $\sigma_i \vec{u}_i \vec{v}_i^T$ to be due to a particular attribute, *e.g.*, comedy, action, sci-fi, or romance content. Then σ_i determines how strongly the ratings depend on the ith attribute, the entries of $\vec{v}_i^T$ score each movie with respect to this attribute, and the entries of $\vec{u}_i$ evaluate how much each viewer cares about this particular attribute. Then truncating the SVD as in (8) amounts to identifying a few key attributes that underlie the ratings. This is useful, for example, in making movie recommendations as you will see in a homework problem.

Finding a SVD

To find a SVD for A we use either the $n \times n$ matrix $A^T A$ or the $m \times m$ matrix AA^T. We will see later that these matrices have only *real eigenvalues*, r of which are positive and the remaining zero, and a complete set of *orthonormal eigenvectors*. For now we take this as a fact and outline the following procedure to find a SVD using $A^T A$:

1. Find the eigenvalues λ_i of $A^T A$ and order them from the largest to smallest, so that $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and $\lambda_{r+1} = \cdots = \lambda_n = 0$.

2. Find orthonormal eigenvectors $\vec{v}_i$, so that

$$A^T A \, \vec{v}_i = \lambda_i \vec{v}_i \quad i = 1, \ldots, r. \tag{9}$$

3. Let $\sigma_i = \sqrt{\lambda_i}$ and obtain $\vec{u}_i$ from

$$A\vec{v}_i = \sigma_i \vec{u}_i \quad i = 1, \ldots, r. \tag{10}$$

We will provide a justification for this procedure in the next lecture. For now we provide an example:

Example: Let

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}.$$

Since this matrix is rank-1 it is not difficult to write it as a column times row, but we will instead practice the general procedure above. Note that

$$A^T A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 12 \end{bmatrix}$$

and the eigenvalues of $A^T A$ are obtained from:

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda - 3 & -6 \\ -6 & \lambda - 12 \end{bmatrix} = \lambda^2 - 15\lambda = \lambda(\lambda - 15) = 0.$$

Therefore, $\lambda_1 = 15$ and $\lambda_2 = 0$. Next we find an eigenvector $\vec{v}_1$ from

$$\begin{bmatrix} \lambda_1 - 3 & -6 \\ -6 & \lambda_1 - 12 \end{bmatrix} \vec{v}_1 = \begin{bmatrix} 12 & -6 \\ -6 & 3 \end{bmatrix} \vec{v}_1 = 0,$$

with length normalized to one:

$$\vec{v}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

We compute the singular value from $\sigma_1 = \sqrt{\lambda_1} = \sqrt{15}$, and $\vec{u}_1$ from (10):

$$\vec{u}_1 = \frac{1}{\sigma_1} A\vec{v}_1 = \frac{1}{\sqrt{15}} \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \frac{1}{\sqrt{15}} \frac{1}{\sqrt{5}} \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Thus we have obtained the SVD:

$$A = \sigma_1 \vec{u}_1 \vec{v}_1^T = \sqrt{15} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix}.$$