ELE427 - Testing Linear Sensors

Linear Regression, Accuracy, and Resolution.

Introduction:

In the first three lab experiments we will be concerned with the characteristics of linear sensors. The basic function of these sensors is to convert a mechanical property to an electrical signal. For instance, the potentiometer will take the angular position of its shaft as an input and convert this to voltage at its output. Linear sensors are those devices for which a plot of the output versus the input forms a straight line. Thus, for an ideal potentiometer, knowing the voltage output you could easily calculate the position input. This would make the potentiometer a very good sensor.

Unfortunately, real sensors suffer from a number of real problems that cause the input/output relationship to be less than ideal. Electronic noise, friction, bending of physical structures, and looseness in couplers and bearings all contribute to the non-ideal behavior observed in the lab. When you design a product that includes a sensor, you will need some way to know if the sensor you choose is going to be good enough. If you choose a cheap sensor, it may degrade the performance of your entire product. Conversely, if you choose a high quality sensor, the cost may make your product less competitive in the marketplace.

Accuracy is a quantitative measure of how closely a real sensor approaches the ideal linear model. The IEEE, ISA, and SAMA organizations define accuracy as having three components: linearity, repeatability, and hysteresis. Manufacturers will usually include some or all of these measurements in their descriptions of a sensor but it is wise to obtain a sample sensor and test it yourself.

The three cycle test is the accepted method of measuring accuracy. This test requires measuring the inputs and outputs of the sensor over the entire operating range of the device. Start with the input at one extreme and vary it until the other extreme is reached taking 10 data points along the way. Then return the input to the original position taking data at the same 10 points as you go. This makes one cycle of data, one end to the other and back. Repeat the same procedure to collect two more cycles of data. Use the same input points for all three cycles. It is useful to make a table for recording your data. Label the columns so you will know which data was taken when the input was increasing, and which data was taken when the input was decreasing.

Linear Regression:

Since we have assumed that the data we collected is from a linear sensor, it seems logical to ask, “What is the one line that this fits data best?”

More mathematically, we are looking for a line, \( y = mx + b \), that passes as close as possible to the N data points we collected, \((x_i, y_i)\), \(i = 1, 2, \ldots, N\). We know \(x_i, y_i\) and \(N\). We want to find the slope, \(m\), and the y-intercept, \(b\). In order to find these unknowns, we need to be more precise about what is meant by “as close as possible”. Considering an input \(x_i\), the line requires that the output should be \(m x_i + b\), but the actual data point is \(y_i\). Thus the vertical distance between the ideal line and the actual point is,
To get a measure of the distance between all the data points and the ideal line, we use the sum of the squares of the individual distances.

\[ E = \frac{1}{2} \sum_{i=1}^{N} [(mx_i + b) - y_i]^2 \]

The square is used to prevent negative differences from canceling positive differences giving an unwarranted impression of accuracy. In this equation \( E \) is referred to as “mean square error” between the data points and the ideal line.

### Example: 3 Cycle Test Data

<table>
<thead>
<tr>
<th>Input</th>
<th>Up</th>
<th>Down</th>
<th>Up</th>
<th>Down</th>
<th>Up</th>
<th>Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.06</td>
<td>0.49</td>
<td>0.06</td>
<td>1.00</td>
<td>0.39</td>
<td>0.50</td>
</tr>
<tr>
<td>80</td>
<td>1.11</td>
<td>1.51</td>
<td>1.26</td>
<td>1.69</td>
<td>1.39</td>
<td>1.65</td>
</tr>
<tr>
<td>160</td>
<td>1.35</td>
<td>2.27</td>
<td>1.77</td>
<td>2.60</td>
<td>2.45</td>
<td>2.34</td>
</tr>
<tr>
<td>240</td>
<td>3.44</td>
<td>3.48</td>
<td>2.88</td>
<td>3.52</td>
<td>3.14</td>
<td>3.70</td>
</tr>
<tr>
<td>320</td>
<td>3.79</td>
<td>5.16</td>
<td>4.52</td>
<td>3.53</td>
<td>4.24</td>
<td>3.94</td>
</tr>
<tr>
<td>400</td>
<td>5.44</td>
<td>5.26</td>
<td>4.91</td>
<td>5.90</td>
<td>4.78</td>
<td>5.68</td>
</tr>
<tr>
<td>480</td>
<td>5.86</td>
<td>5.91</td>
<td>6.17</td>
<td>6.40</td>
<td>5.96</td>
<td>6.55</td>
</tr>
<tr>
<td>560</td>
<td>7.29</td>
<td>6.59</td>
<td>6.62</td>
<td>7.12</td>
<td>7.48</td>
<td>7.34</td>
</tr>
<tr>
<td>640</td>
<td>8.34</td>
<td>8.17</td>
<td>7.97</td>
<td>7.78</td>
<td>8.36</td>
<td>8.45</td>
</tr>
<tr>
<td>800</td>
<td>10.26</td>
<td>10.04</td>
<td>10.02</td>
<td>11.02</td>
<td>9.72</td>
<td>10.40</td>
</tr>
</tbody>
</table>

Input-Encoder counts, Output-volts

Our goal is to find the parameters \( m \) and \( b \) that minimize the mean square error. To do this we will use the fact that the minimum of a function occurs when the first derivatives equal zero. Because \( E \) depends on both \( m \) and \( b \), we will take partial derivatives with respect to \( m \) and \( b \).

\[ \frac{\delta E}{\delta m} = \sum_{i=1}^{N} [(mx_i + b) - y_i] x_i \]

\[ \frac{\delta E}{\delta b} = \sum_{i=1}^{N} [(mx_i + b) - y_i] l \]

Setting the derivatives equal to zero and expanding, we obtain:
\[
\sum_{i=1}^{N} (mx_i^2 + bx_i - x_iy_i) = 0 \\
\sum_{i=1}^{N} (mx_i + b - y_i) = 0
\]

Reordering the terms, we recognize two simultaneous equations in two unknowns:

\[
m\sum x_i^2 + b\sum x_i = \sum x_iy_i \\
m\sum x_i + bN = \sum y_i
\]

or, in matrix format:

\[
\begin{bmatrix}
\sum x_i^2 & \sum x_i \\
\sum x_i & N
\end{bmatrix}
\begin{bmatrix}
m \\
b
\end{bmatrix}
=
\begin{bmatrix}
\sum x_iy_i \\
\sum y_i
\end{bmatrix}
\]

Using determinants (Cramer’s rule) you can solve for \( m \) and \( b \).

\[
m = \frac{\left| \begin{array}{cc}
\sum x_iy_i & \sum x_i \\
\sum x_i & N
\end{array} \right|}{\left| \begin{array}{cc}
\sum x_i^2 & \sum x_i \\
\sum x_i & N
\end{array} \right|} \\
b = \frac{\left| \begin{array}{cc}
\sum x_i^2 & \sum x_iy_i \\
\sum x_i & \sum y_i
\end{array} \right|}{\left| \begin{array}{cc}
\sum x_i^2 & \sum x_i \\
\sum x_i & N
\end{array} \right|}
\]

Although this looks very messy, the calculations are not that bad (especially if you use MATLAB). Let’s consider the example data set shown in the previous table.

\[
\sum x_i = 26400 \\
\sum x_i^2 = 14784000 \\
\sum x_iy_i = 190640 \\
\sum y_i = 347 \\
N = 66
\]

Note that the sum of \( x_i \) is not just the sum of the 11 input values shown on the table. It is six times the sum of the table values because each \( x \) value shown actually corresponds to six \( y \) values. Using these values we find that,

\[
m = 0.0123 \quad \text{and} \quad b = 0.3485.
\]

This “best-fit line” and the actual lab data are shown in Figure 1.
Linearity:

Once we have found the best-fit line, it is now possible to consider a quantitative measure of how closely the sensor approaches the ideal linear behavior. Specifically, we want to obtain a single number that characterizes how far the data lies from the line. There are several ways that this is commonly done.

One method is called the worst case linearity. In this case we are interested in the data point that lies furthest from the ideal line. First, you need to calculate where the ideal line is for each of the input data points \(x_i\) that you used when you took the data. Then find the difference between these ideal outputs and the actual lab measurements. The worst case linearity is just the largest (magnitude) of these differences. In our example the worst case linearity occurs on the third measurement of the first cycle when the input is 160 counts and the output is 1.35 volts. The ideal line passes through 2.3165 = 0.0123(160) + 0.3485, and so the difference is 0.9665 volts. However, linearity is never reported in units of volts. Instead, it is reported as a percentage of the full-scale operating range of the sensor. The extremes of the ideal line are at 0.3485 volts and 10.1885 volts (0.0123 x 800 + 0.3485) so the full-scale is 9.84 (10.1885- 0.3485) volts. Thus the worst case linearity is 0.9665/9.84 = 0.0982 or 9.82%.

A second common method of calculating linearity is the mean square linearity. As the name suggests, this number is obtained by squaring the individual differences between the ideal line and the data points, and then averaging the squares. Again, the result is reported as a percentage of the full-scale output of the sensor. Calculating this quantity for the example data set we find that the mean of the squares is 0.1339 and so the mean square linearity is 0.1339/9.84 = 0.0136 or 1.36%.
Repeatability:
Recall that the second part of accuracy is repeatability. This refers to the ability of the sensor to yield the same output any time it receives the same input. One of the reasons for taking three cycles of data is to test whether you get the same output each time you apply the same input. Similar to linearity there are several methods by which repeatability is calculated.

Worst case repeatability refers to the largest difference between two outputs taken with the same input and going in the same direction. This means you are interested in the largest difference between the three outputs you obtained for at each input point as the input was increased or decreased. We are not interested in the difference between outputs that were taken with the input going in different directions, nor do we care where the ideal line lies. This is just a measure of whether the sensor will repeat its previous outputs when you give it the same inputs.

For our example data set the largest difference between two outputs with the same input in the same direction occurs between the first and second cycle in the decreasing direction when the input was 320. During cycle 1 the output was 5.16 volts, during cycle 2 the output was 3.53 volts for a difference of 1.63 volts. Like linearity, repeatability is reported as a percentage of the full-scale output of the sensor. Thus the worst case repeatability is 1.63/9.81 = 0.166 or 16.6%.

A mean square repeatability could be found by taking the differences between all the points obtained with the same input in the same direction, squaring each individual difference, and then averaging all the squares. Using this method, our data set has an average of squares of 0.2535 which yields a mean square repeatability of 2.58%.

Hysteresis:
The third measure of accuracy has to do with the behavior of the sensor when the same input is approached from opposite directions. Here we are interested in the difference between outputs that occurred as the input was increasing and outputs that occurred as the input was decreasing. The location of the ideal line doesn't matter.

Worst case hysteresis is the largest difference between two outputs taken with the same input and going in opposite directions. This occurs between the first cycle increasing and the first cycle decreasing when the input was 320 counts. The outputs were 3.79 volts and 5.16 volts respectively for a difference of 1.37 volts or 13.9% of full-scale. Like the other components of accuracy, hysteresis is always reported as a percentage of the full-scale output of the sensor.

We could calculate a mean square hysteresis by the same method used to get mean square repeatability. A more illuminating procedure might be to find 2 best fit lines. One using only data taken with the input increasing, and the other using only data taken with the input decreasing. Performing this calculation on our example data set we find that,

\[ m_{up} = 0.0124, \quad b_{up} = 0.1363, \quad \text{while,} \quad m_{down} = 0.0121, \quad b_{down} = 0.5602. \]

Figure 2 shows the actual data and these two best fit lines. It can be seen that there is a distinct difference in the outputs depending on which direction the input was moving.
Resolution:

Although a sensor might be very accurate with respect to linearity, repeatability, and hysteresis, it may still not be suitable for some applications. Resolution has to do with the ability of a sensor to discern fine detail. We define resolution as the smallest change in input that can always be detected at the output. Consider an optical encoder that has 1000 windows. Such an encoder will generate 1000 counts every time it turns 360 degrees. This means that if you turn the encoder $360/1000 = 0.36$ degrees or more you will always generate a count at the output. So, we would say that the resolution of this encoder is 0.36 degrees. Notice that resolution is an absolute quantity, not a relative percentage. It will always have the same units as the input to the sensor.

A potentiometer that is set up as a voltage divider is another common situation. In this case, any movement of the input shaft will change the output voltage. This device seems to have an infinite resolution! However, for the potentiometer to be of any use the output voltage must be measured. So the limit of resolution for the potentiometer system is defined by how exactly you can measure the output voltage. If your voltmeter can measure a one millivolt change, and the potentiometer voltage changes by one volt per 32 degrees, then we would say that the resolution of the potentiometer/voltmeter system is $(10^{-3} \text{ volts})(36 \text{ degrees/volt})=0.036$ degrees.

Last update:
C. Turtle,
August, 2000