1 Introduction

In this project we propose a method for detecting sensor anomalies in time-series data based on the difference between the predicted and actual output of a sensor. We use a linear regression model to predict the output of a sensor based on the outputs of sensors on other channels. We additionally validate our model on fault-free data to determine how well our model correlates with actual sensor output. Then, we can run our model in real time and can detect sensor anomalies when the correlation between the actual and expected sensor data is far from the expected correlation.

2 Assumptions

We treat the data as if it were a point cloud, meaning that at each time step, t, the values of all channels at time t as a vector are a point. We treat all points as if they are independent and are drawn from the same distribution.

3 Problem Setting

Say that we have n channels, $c_1, \ldots, c_n$, and that $C_i(t)$ is the value of all channels but i at a given time t.

For each channel c_i we have a map $f_i : \mathbb{R}^{n-1} \to \mathbb{R}$ that, given the state of other channels at a given point in time, predicts the value of channel i. For each f_i, we also have the metric ρ_i, where

$$\rho_i = \frac{\mathrm{Cov}_t(f_i(C_i(t)), c_i(t))}{\sqrt{\mathrm{Var}_t(f_i(C_i(t)))\,\mathrm{Var}_t(c_i(t))}},$$

which is the Pearson correlation coefficient between the predicted channel, given by $f_i(C_i(t))$ for a singe point in time, and the actual channel c_i. Note that $|\rho_i|$ may not be large for some channels, meaning that our function f_i does a poor job fitting this channel. However, our model in detecting anomalies will account for the fact that many fits may be imperfect.

Then, we will use our model and some additional statistics to check for anomalies in real-time. At a high level, our model will calculate the correlation between how our model predicts what the sample will be versus what the sample actually is. For a window of length k, meaning entries from $t_0 - k + 1, \ldots t_0$, where t_0 is the current time, we will find r_i, the test statistic, given by

$$r_i = \frac{\mathrm{Cov}_{t_0-k<t\leq t_0}(f_i(C_i(t)), c_i(t))}{\sqrt{\mathrm{Var}_{t_0-k<t\leq t_0}(f_i(C_i(t)))\,\mathrm{Var}_{t_0-k<t\leq t_0}(c_i(t))}},$$

We will show how to calculate this test statistic efficiently (in $\Theta(1)$ time amortized per window) and will show how to reject a sample as an anomaly using probabilistic methods.

For the rest of the discussion, we will fix some t_0 and some k. For notation purposes, let $X_j = f_i(C_i(t_0 - k + j))$ and $Y_j = c_i(t_0 - k + j)$ for $1 \leq j \leq k$, meaning that X_j and Y_j are the jth entries of the predicted and actual samples of the time series. Then, let A and B be the normalized version of X and Y, meaning that

$$A_i = \frac{X_i - \bar{X}}{\sigma_X}, \qquad B_i = \frac{X_i - \bar{Y}}{\sigma_Y},$$

such that

$$\mathbb{E}[A] = \mathbb{E}[B] = 0, \qquad \mathrm{Var}(A) = \mathrm{Var}(B) = 1.$$

It is important to note that this normalization is really just a ”trick” to simplify the calculations. We additionally note that it does not matter whether we use the sample variance (multiplied by $\frac{k}{k-1}$ or not as long as we are consistent and also use the sample covariance, as both result multiplying the top and bottom of the expression for r_i by the same value. Now, with A and B, we have that the expression for r_i is

$$r_i = \frac{\mathrm{Cov}(A, B)}{\mathrm{Var}(A)\,\mathrm{Var}(B)},$$

but since the $\mathrm{Var}(A) = \mathrm{Var}(B) = 1$, this is just

$$r_i = \mathrm{Cov}(A, B).$$

Then, by the definition of the covariance and since we know the values of A and B this is just

$$r_i = \mathrm{Cov}(A, B) = \mathbb{E}[AB] = \frac{1}{k} \sum_{i=1}^{k} A_i B_i,$$

which, interestingly, is the cosine-distance if we treat these as scalars (an aside: since $\mathrm{Var}[A] = \mathrm{Var}[B] = 1$, this says $\frac{1}{k} \|A\|_2^2 = \frac{1}{k} \|B\|_2^2 = 1$, meaning that $\|A\|_2 = \|B\|_2 = \sqrt{k}$, so the cosine distance $\frac{A \cdot B}{\|A\|_2 \|B\|_2}$ is $\sum_{i=1}^{k} A_i B_i / \sqrt{k * k} = \frac{1}{k} \sum_{i=1}^{k} A_i B_i = r_i$.).

4 Detecting Anomalies

Now, we want to run a statistical test to determine if the window ending at t_0 represents an anomaly. Then, we use the two hypotheses:

$$H_0 : \mu_{r_i} = \rho_i, \qquad H_a : \mu_{r_i} \neq \rho_i.$$

The null hypothesis, H_0, represents that the correlation between this window of the data is the same as the correlation that we would expect. This then represents that the data is as expected and that there is no anomaly. H_a represents some sort of anomaly in the window. Our goal will then be to calculate the probability p that this window has correlation r_i given that H_0 is true, and if p is "low enough", meaning $p < \alpha$ for an α we will define later, then we can say with confidence that there is an anomaly.

Then, to calculate this probability, we want to know the distribution of r_i. We can think of A_i and B_i themselves as being random variables, so $A_i B_i$ is a random variable, and by our assumptions these are independent and identically distributed. Then, since r_i is a sum of independent indentically distributed random variables, the Central Limit Theorem applies, which says that in the limit (as $k \to \infty$), that r_i is normally distributed. This gives us a good approximation for the distribution of r_i. Then, since we assume that $\mu = \rho$, the only parameter we need estimate is the standard deviation of r_i. We have that this is

$$\mathrm{Var}(r_i) = \mathrm{Var}(\frac{1}{k} \sum_{i=1}^{k} A_i B_i) = \frac{1}{k^2} \sum_{i=1}^{k} \mathrm{Var}(A_i B_i),$$

by linearity, but since $A_i B_i$ are all identically distributed, their variances are the same, so we can write this as

$$= \frac{1}{k} \mathrm{Var}(AB) = \mathbb{E}((AB - r_i)^2) = \frac{1}{k} \left(\mathbb{E}((AB)^2) - r_i^2 \right),$$

by a well-know formula for variance. To calculate $\mathbb{E}((AB)^2)$, we just do

$$\mathbb{E}((AB)^2) = \frac{1}{k} \sum_{i=1}^{k} (A_i B_i)^2.$$

Additionally, to get the sample variance, we multiply by $\frac{k}{k-1}$, so the full formula for $S_{r_i}^2$ is

$$\mathrm{Var}(r_i) = \frac{1}{k-1} \left(\sum_{i=1}^{k} (A_i B_i)^2 - r_i^2 \right).$$

Then, this means that

$$\frac{r_i - \rho}{\sqrt{S_{r_i}^2}}$$

is distributed approximately normally, which means that given some other standard normal random variable N that the probability of drawing r_i randomly under the assumption that $\mu_{r_i} = \rho$ is

$$p \approx P\left(|N| \geq \left|\frac{r_i - \rho}{\sqrt{S^2_{r_i}}}\right|\right) = 2P\left(N \leq -\left|\frac{r_i - \rho}{\sqrt{S^2_{r_i}}}\right|\right),$$

which we can calculate using existing normal CDF functions.

5 Finding r_i efficiently

We can see from above that we can calculate r_i in $\Theta(k)$ time, where k is the the size of the window, which isn't bad. However, if we want the anomaly detection system to work in real-time, we want to calculate r_i faster than this. We will show a method for calculating r_i in amortized $\Theta(1)$ time and using $\Theta(k)$ memory, assuming that we are calculating r_i for every window of size k.

The method works by using two different queues that store the last k values and additionally by maintaining a number of different accumulators that store some value that changes as we move from left to right. Each queue stores the last k values of X_i and Y_i. We have to spend $\Theta(k)$ time initially populating each of the queues but when we move from the window ending at t_0 to the one ending at $t_0 + 1$, we extract the oldest element from each of the queues, say X_0 and Y_0, and replace it with the newest element, say X_k and Y_k. Then, we use the old values to subtract off the accumulators and the new values to add to them to keep accurate. We'll denote $\sum(Z)$ to be the accumulator $\sum_{i=1}^{k} Z_i$. For example, $\sum(X) = \sum_{i=1}^{k} X_i$. We maintain the following accumulators:

$$\sum(X), \sum(Y), \sum(X^2), \sum(Y^2), \sum(XY), \sum((XY)^2), \sum(X^2Y), \sum(XY^2).$$

For calculating r_i, we have

$$\begin{aligned} r_i = \operatorname{Cov}(A, B) &= \operatorname{Cov}\left(\frac{X - \bar{X}}{\sigma_X}, \frac{Y - \bar{Y}}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X \sigma_Y} \operatorname{Cov}(X - \bar{X}, Y - \bar{Y}) \\ &= \frac{1}{\sigma_X \sigma_Y} (\mathbb{E}[XY] - \bar{X}\bar{Y}). \end{aligned}$$

We will write down how to calculate all of the variables in this expression in terms of our accumulators.

$$\begin{aligned} \sigma_X = \sqrt{\mathbb{E}[(X - \mathbb{E}[X]^2)]} = \sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2} &= \sqrt{\frac{1}{k}\sum_{i=1}^{k} X^2 - \left(\frac{1}{k}\sum_{i=1}^{k} X\right)^2} \\ &= \frac{1}{k}\sqrt{k\sum(X^2) - \sum(X)^2} \end{aligned}$$

$$\begin{aligned} \sigma_Y = \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y]^2)]} = \sqrt{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2} &= \sqrt{\frac{1}{k}\sum_{i=1}^{k} Y^2 - \left(\frac{1}{n}\sum_{i=1}^{k} Y\right)^2} \\ &= \frac{1}{k}\sqrt{\sum(Y^2) - k\sum(Y)^2} \end{aligned}$$

$$\mathbb{E}[XY] = \frac{1}{k}\sum_{k=1}^{k} XY = \frac{1}{k}\sum(XY)$$

$$\bar{X} = \mathbb{E}[X] = \frac{1}{k}\sum_{i=1}^{k} X = \frac{1}{k}\sum(X)$$

$$\bar{Y} = \mathbb{E}[Y] = \frac{1}{k}\sum_{i=1}^{k} Y = \frac{1}{k}\sum(Y)$$

Now, we will show how to calculate Var(r). We know from above that this is

$$\operatorname{Var}(r) = \frac{1}{k}\left(\mathbb{E}((AB)^2) - r^2\right),$$

so since we know r (we just calculated it) we only need to find $\mathbb{E}((AB)^2)$. Inserting the definitions for A and B this is

$$\begin{aligned}
\mathbb{E}((AB)^2) &= \mathbb{E}\left(\left(\frac{X-\bar{X}}{\sigma_X}\cdot\frac{Y-\bar{Y}}{\sigma_Y}\right)^2\right)\\
&= \frac{1}{\sigma_X^2\sigma_Y^2}\,\mathbb{E}\left((X-\bar{X})^2(Y-\bar{Y})^2\right)\\
&= \frac{1}{\sigma_X^2\sigma_Y^2}\,\mathbb{E}\left((X^2-2X\bar{X}+\bar{X}^2)(Y^2-2Y\bar{Y}+\bar{Y}^2)\right)\\
&= \frac{1}{\sigma_X^2\sigma_Y^2}\,\mathbb{E}\left(X^2(Y^2-2Y\bar{Y}+\bar{Y}^2)-2X\bar{X}(Y^2-2Y\bar{Y}+\bar{Y}^2)+\bar{X}^2(Y^2-2Y\bar{Y}+\bar{Y}^2)\right)\\
&= \frac{1}{\sigma_X^2\sigma_Y^2}\,\mathbb{E}\left(X^2Y^2-2X^2Y\bar{Y}+X^2\bar{Y}^2-2X\bar{X}Y^2+4X\bar{X}Y\bar{Y}-2X\bar{X}\bar{Y}^2+\bar{X}^2Y^2-2\bar{X}^2Y\bar{Y}+\bar{X}^2\bar{Y}^2)\right)\\
&= \frac{1}{\sigma_X^2\sigma_Y^2}\left(\mathbb{E}(X^2Y^2)-2\,\mathbb{E}(X^2Y)\bar{Y}+\mathbb{E}(X^2)\bar{Y}^2-2\,\mathbb{E}(XY^2)\bar{X}\right.\\
&\quad\left.+4\,\mathbb{E}(XY)\bar{X}\bar{Y}-2\,\mathbb{E}(X)\bar{X}\bar{Y}^2+\bar{X}^2\,\mathbb{E}(Y^2)-2\,\mathbb{E}(Y)\bar{X}^2\bar{Y}+\bar{X}^2\bar{Y}^2\right).
\end{aligned}$$

Then, noting that $\mathbb{E}[X]=\bar{X}$ and $\mathbb{E}[Y]=\bar{Y}$, we can simply this further and some terms cancel out, getting

$$\begin{aligned}
&= \frac{1}{\sigma_X^2\sigma_Y^2}\left(\mathbb{E}(X^2Y^2)-2\bar{Y}\,\mathbb{E}(X^2Y)+\bar{Y}^2\,\mathbb{E}(X^2)-2\bar{X}\,\mathbb{E}(XY^2)\right.\\
&\quad\left.+4\bar{X}\bar{Y}\,\mathbb{E}(XY)-2\bar{X}^2\bar{Y}^2+\bar{X}^2\,\mathbb{E}(Y^2)-2\bar{X}^2\bar{Y}^2+\bar{X}^2\bar{Y}^2\right).\\
&= \frac{1}{\sigma_X^2\sigma_Y^2}\left(\mathbb{E}(X^2Y^2)+\bar{X}^2\,\mathbb{E}(Y^2)+\bar{Y}^2\,\mathbb{E}(X^2)-2\bar{X}\,\mathbb{E}(XY^2)-2\bar{Y}\,\mathbb{E}(X^2Y)-3\bar{X}^2\bar{Y}^2+4\bar{X}\bar{Y}\,\mathbb{E}(XY)\right).
\end{aligned}$$

Then, we have already described how to find $\sigma_X, \sigma_Y, \bar{X}, \bar{Y}$, and $\mathbb{E}[XY]$. in the previous calculation of r_i. We will show how to find the other variables in terms of the accumulators below:

$$\mathbb{E}[X^2Y^2] = \mathbb{E}[(XY^2) = \frac{1}{k}\sum i = 1^k (X_iY_i)^2 = \frac{1}{k}\sum((XY)^2)$$

$$\mathbb{E}[Y^2] = \frac{1}{k}\sum_{i=1}^{k} Y_i^2 = \frac{1}{k}\sum(Y^2)$$

$$\mathbb{E}[X^2] = \frac{1}{k}\sum_{i=1}^{k} X_i^2 = \frac{1}{k}\sum(X^2)$$

$$\mathbb{E}[XY^2] = \frac{1}{k}\sum_{i=1}^{k} X_iY_i^2 = \frac{1}{k}\sum(XY^2)$$

$$\mathbb{E}[X^2Y] = \frac{1}{k}\sum_{i=1}^{k} X_i^2Y_i = \frac{1}{k}\sum(X^2Y)$$

6 Choosing α

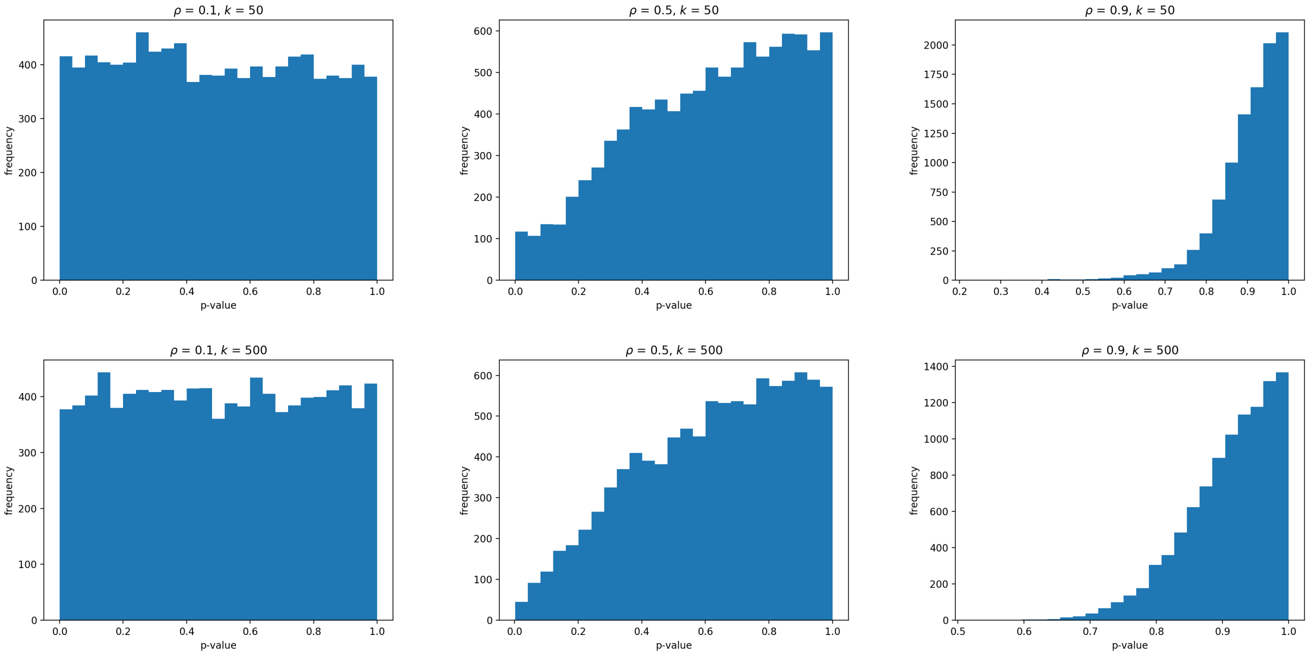
When we run our statistical test, we will reject H_0 if $p < \alpha$, meaning that we detect an anomaly. However, since we will our anomaly detection algorithm will be running many times, if we set $\alpha = 0.05$, if the algorithm run 100 times and there are no malfunctioning sensors, we will still report on average 5 anomalies. Then, we see that we need to make α much lower. We will set α such that, given every sensor is working (meaning H_0 is true) the probability that there will be no reported anomalies is α_0. Then, if the probability of a single test reporting an anomaly is α, the probability it does not report one is $1 - \alpha$, so the probability that given N tests we return no anomalies is $(1 - \alpha)^N$. Then, the probability of at least one anomaly being reported is $1 - (1 - \alpha)^N$, so we want to have $1 - (1 - \alpha)^N < \alpha_0$, so

$$
\begin{gathered}
1 - \alpha_0 < (1 - \alpha)^N \\
\sqrt[n]{1 - \alpha_0} < 1 - \alpha \\
\alpha < 1 - \sqrt[n]{1 - \alpha_0}.
\end{gathered}
$$

Then, for example, if we want $\alpha_0 = 0.05$, so if we have $N = 200,000$ tests and want $\alpha_0 = 0.05$, then $\alpha = 2.564 \times 10^{-7}$.

7 Evaluating the Approximation

Here we will experimentally evaluate how well the approximation of assuming r_i is normally distributed works. With $10,000$ trials in each test, we randomly generated k data points of standard normal X and Ys which have correlation coefficient ρ, and used X as the predicted sample and Y as the actual sample. We ran the test in each trial and calculated a p-value. The distributions of the p-values in the cases we tested are shown below.



If the approximation worked well, we would expect that p-values would be uniformly distributed. We see that this is the case for $\rho = 0.1$. Unfortunately, for high ρ, we see that the approximation breaks down. We see that in cases where the two series are highly correlated that the chance of a low p-value is very low. However, while this isn't optimal, it also isn't terrible as it just means that the false-positive rate of detecting anomalies will be lower.