Goal

Implementing an anomaly detector using linear unsupervised learning and dimensionality reduction.

Data

The included data.csv contains a set of points corresponding to US Treasury rates for a consecutive 30 days – each row corresponds to several rates (known as the Treasury curve) published on a single day.

Model

The various treasury rates are highly correlated and follow a model:

$$r_t = \overline{r} + \Sigma \cdot z_t$$

Where r_t is a N-dim vector representing the rates at date t; $\overline{r}$ and Σ are parameters; and z_t is a vector of normally distributed latent variables (these are the hidden states that we use for dimensionality reduction).

When the sample is closely spaced (such as within a window of 30 days) then this model is stationary and $\overline{r}$ corresponds to the sample mean vector and $\Sigma^T \Sigma$ corresponds to the sample covariance matrix.

If we use eigen decomposition on the sample covariance:

$$\Sigma^T \Sigma = U^T \Lambda^2 U$$

then we have solved for Σ:

$$\Sigma = \Lambda U$$

Detecting Anomalies

We will use dimensionality reduction to find anomalies.

After solving for $\overline{r}$ and Σ for our sample, we will use only the first M columns of Σ and first M elements of z_t to model the rates:

$$r_t = \overline{r} + \Sigma_M \cdot [z_t]_M + \epsilon$$

Where ϵ is the residual left over by dropping the rest of the columns and latent variables.

Now, for each time, t, let us expand and rewrite this equation:

$$\begin{bmatrix} r_{t,1} - \overline{r}_1 \\ \vdots \\ r_{t,L} - \overline{r}_L \\ \vdots \\ r_{t,N} - \overline{r}_N \end{bmatrix} = \Sigma_M \cdot [z_t]_M + \epsilon$$

Where $r_{t,i}$ is the i-th rate at time t, and $\overline{r}_i$ is the mean of the i-th rates in the sample.

If we change the rate, $r_{t,L}$, into a variable, $\hat{r}_{t,L}$, then we can estimate the rate projected by the model and the rest of the rates for that particular day. This is done by transfering $\hat{r}_{t,L}$ to the side of our latent variables and solving it as a linear regression:

$$\begin{bmatrix} r_{t,1} - \overline{r}_1 \\ \vdots \\ -\overline{r}_L \\ \vdots \\ r_{t,N} - \overline{r}_N \end{bmatrix} = \begin{bmatrix} & \begin{bmatrix} 0 \\ \vdots \\ -1 \\ \vdots \\ 0 \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} [z_t]_M \\ \hat{r}_{t,L} \end{bmatrix} + \epsilon$$

Where we have extended Σ_M with an additional column of all zeroes except for a -1 at row L, and we have extended the vector $[z_t]_M$ with the additional variable $\hat{r}_{t,L}$. If we set:

$$y = \begin{bmatrix} r_{t,1} - \overline{r}_1 \\ \vdots \\ -\overline{r}_L \\ \vdots \\ r_{t,N} - \overline{r}_N \end{bmatrix}; \quad X = \begin{bmatrix} & \begin{bmatrix} 0 \\ \vdots \\ -1 \\ \vdots \\ 0 \end{bmatrix} \end{bmatrix}; \quad \beta = \begin{bmatrix} [z_t]_M \\ \hat{r}_{t,L} \end{bmatrix}$$

Then the least squares solution is found by:

$$\beta = (X^T X)^{-1} X^T y$$

Now we have a set of projected rates, $\hat{r}_{t,L}$, for each data point originally in the data set. The error terms: $r_{t,L} - \hat{r}_{t,L}$ should be normally distributed and any term > 4 standard deviations from 0 will indicate an anomaly. For Treasury rates, a good choice of the reduced dimensions, M, will be 2 or 3.

Task

Please implement this algorithm in Python to discover all the anomalous data in data.csv with comments and and analysis of results at the end. You may use any Python package you want. Please contact me directly (alon.kadashev@zesty.co) if you will have any questions or need to clarify anything in this document.