Description of algorithm

1. For each column impute missing values using cubic splines. Any missing values coming before the first non-missing value or after the last non-missing value are not imputed since the behavior of extrapolating cubic splines before the first spline or after the last spline can sometimes be erratic.

2. For each unique pair c_i and c_j of columns:

    a. Center c_i and c_j by subtracting the mean/average of each.
    b. Calculate the following distance metric based on crosscorelation (taken from the function function CCorDistance in R's Tdist package (author Usue Mori).

    $$D = \sqrt{((1 - CC(x, y, 0)^2)/\Sigma(1 - CC(x, y, k)^2)} \quad (1)$$

    where $CC(x, y, k)$ is the normalized cross-correlation between centered column x and centered column y at lag k.

3. Do agglomerative hierarchical clustering The distance metric between individual columns x and y is given in (1).
    For groups u and v of multiple columns the distance is calculated as the max of all distances between individual columns of u and v.

    $$d(u, v) = \max\_\big(D(u[i], v[j])\big)$$

    for any columns $u[i] \in u,\ v[j] \in v$.

4. The selected criterion for stopping the agglomeration of different groups is that in order for two groups to be joined the distance must be less than the 20 % quantile of all distances between any two unique columns in the data.