ENGINEERING
Chemical Engineering

McMaster University

Fuzzy Clustering based PLS Modelling for Microgel Property Prediction

Anish Anand Pophale

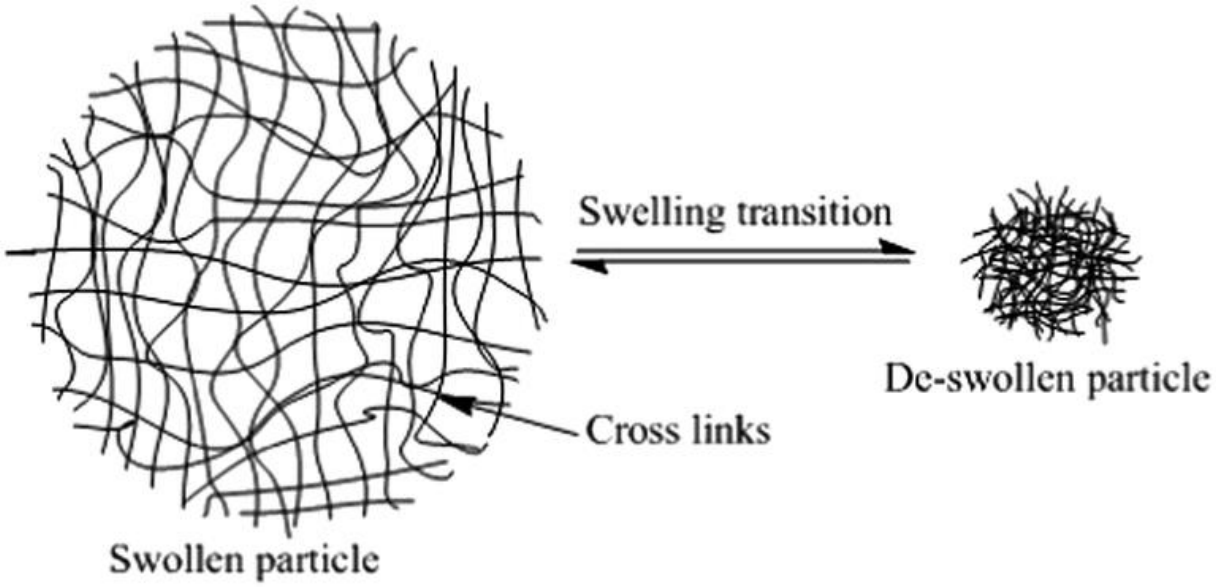
Supervisors
Dr. Prashant Mhaskar
Seyed Saeid Tayebi

Summer 2023

What are microgels?

- Microgels are colloidal gel particles, which are 3D cross linked polymer networks suspended in a solvent.
- They are responsive to external stimuli such as changes in pH, Temperature, Light, Magnetic field because of specific interactions within the gel network and between the gel and its environment.
- Some of the application of microgels are in Drug delivery, Enhanced Oil Recovery and Cosmetics.

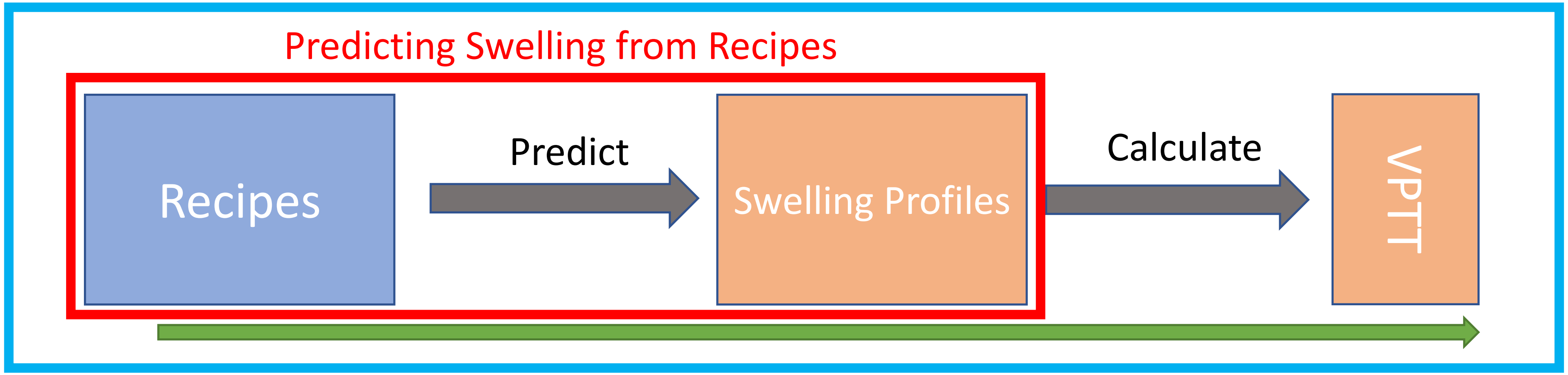
- For this project, we have used pH and temperature multi responsive microgels designed for drug delivery applications.
- The Volume Phase Transition Temperature (VPTT) is the temperature at which the microgel undergoes a significant change in its volume and transitions between the swollen / de-swollen phase.



Why a data driven approach?

- Controlling the Microgel Transition Temperature through the polymerization recipe can expand its application significantly for drug delivery purposes.
- But developing a first principles based model relating the recipe to the swelling profile or VPTT has several challenges associated with it such as the complex dynamics of swelling, parameters for the model being difficult to measure and multiple, interacting physical and chemical factors that influence the VPTT.
- A data driven approach using a latent variable modelling method such as PLS modelling can be used to relate the recipe to the transition behaviour without getting involved in the complex dynamics.
- Similar techniques have been used in such complex areas, for product design based on the available formulation or estimating potential formulation while seeking a desired product.

Modelling Approach (Previous Work)

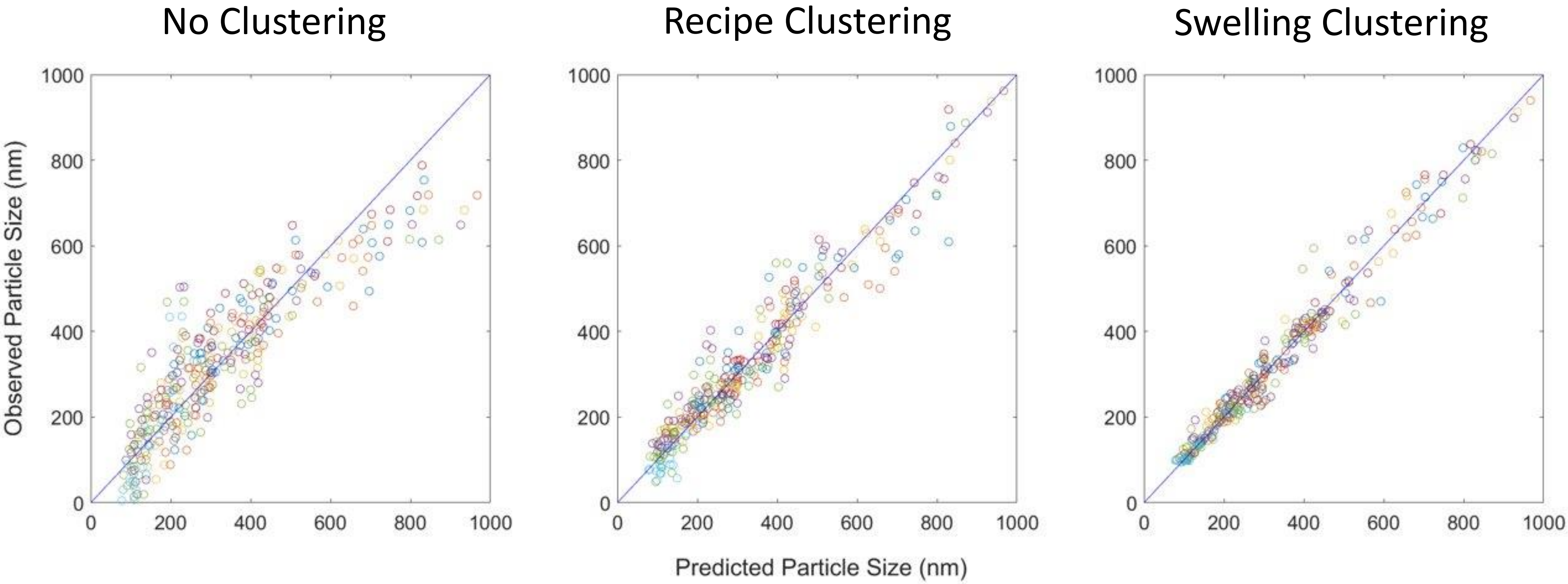


- Using a Partial Least Squared model, we first predict the Swelling Profiles of the microgels at pH 4 and pH 10 based on their recipe
- We have a dataset of the recipes and swelling profiles which was developed through experiments, using which the model is trained.
- From the swelling profile, by fitting a sigmoid curve, the VPTT is calculated
- In the previous work it was suggested that rather than using a single linear PLS model, a clustering based PLS method can provide better results by building separate PLS models for subsets of the data.
- Such an approach can provide better predictions as the models are built using similar observations in the data.

Results (Previous Work)

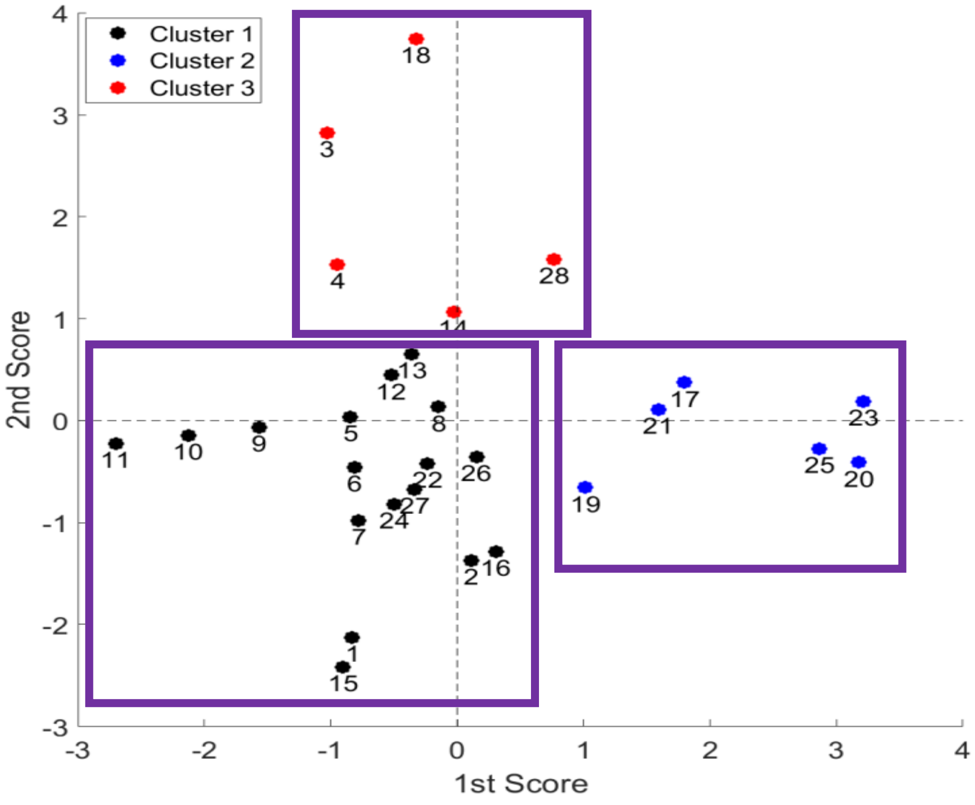
ENGINEERING
Chemical Engineering

McMaster University



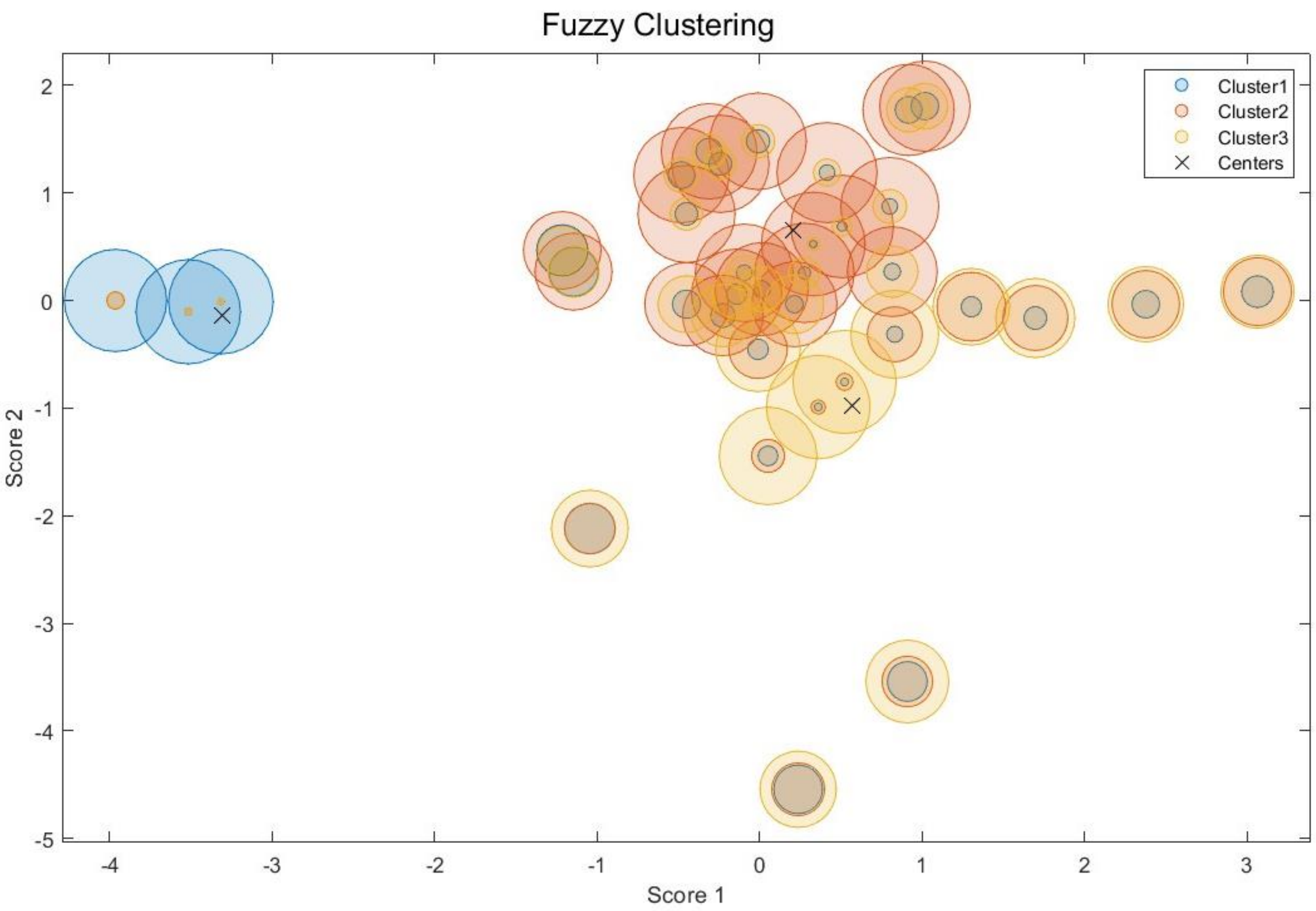
K Means PLS Approach (Previous Work)

- Clustering samples using a K Means approach so that each sample belongs to only one of the clusters.
- Separate PLS models are developed for each of the clusters.
- For predictions, the cluster in which is sample belongs to is identified and the respective PLS model is used.
- Advantage:
  - We are building the models on similar observations which can give improved predictions.
- Disadvantage:
  - We are ignoring relations between input and output that apply to all clusters but are only capturable only in some of the clusters, if any.



Potential Solution

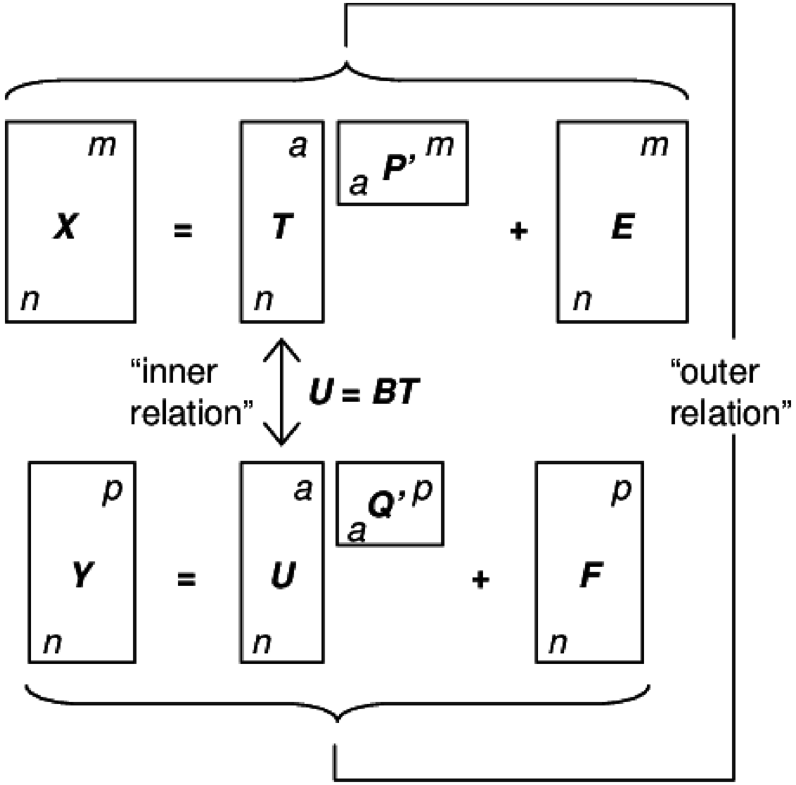
- Employing a fuzzy clustering approach can be the solution.
- In case of fuzzy clustering, a sample belongs to all of the clusters with a membership function corresponding to each of the clusters.
- Since each of the clusters contain all the observations with some membership function, for each of the PLS model we need to consider all the samples and, in some way, incorporate the membership functions which is the main challenge.
- The performance of such an approach should be compared with Linear PLS, the K Means PLS models and existing methods using fuzzy clustering in the literature.



Partial Least Squares Model

ENGINEERING
Chemical Engineering
McMaster University

- A PLS model decomposes the dependent and independent variables to a set of scores and loadings in a lower dimension in a way which maximizes the corelation between the scores assuming a linear relation between them.
- PLS is used in situations when the data has colinearity or there are more predictors than observations, where simple methods such as OLS do not work.



Outer Relation

$$X = TP' + E = \sum t_h p_h' + E$$

$$\mathrm{Y} = UQ' + \mathrm{F} = \sum u_h q_h' + \mathrm{F}$$

Inner Relation

$$\hat{u}_h = b_h t_h$$

Fuzzy Clustering Based PLS (Literature)

Inner Fuzzy PLS:

- Modifies the inner PLS relation in the NIPALS Algorithm as set of linear models which are weighted by the membership functions.
- This method is called as Fuzzy PLS (FPLS) in literature and has been used for non linear modelling.

$$\hat{u}_h = \sum_{i=1}^{C} b_h(i)\, \mu_i t_h$$

Outer Fuzzy PLS 2:

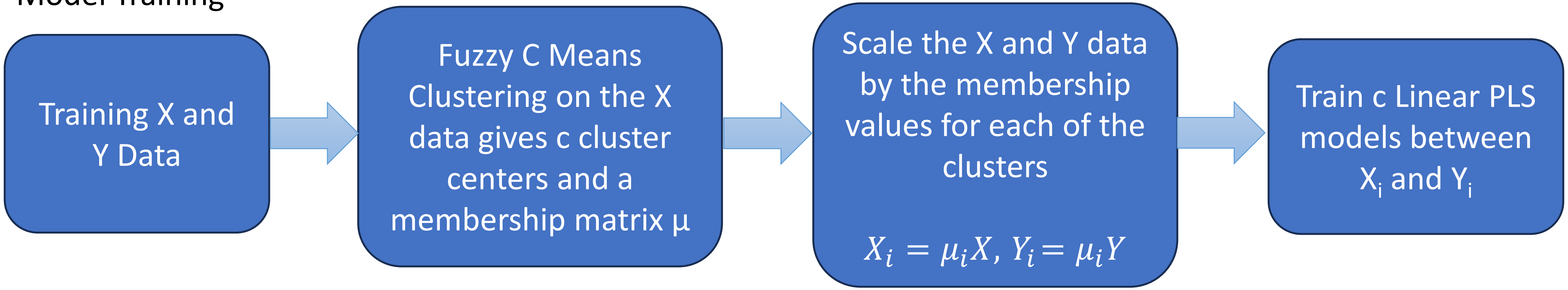
- This approach is a combination of the Outer Fuzzy PLS and K Means PLS
- Fuzzy clustering is implemented after which each point is assigned to the cluster in which it has maximum membership, each point now belonging to only one cluster
- For predictions, a weighted sum of the outputs from all the PLS models is used, where the weights are the membership function of that point in all the clusters

Fuzzy Clustering Based PLS (Our Suggestion)

Outer Fuzzy PLS:

- Changes the outer relation in the PLS method by weighting the X and Y matrices by the membership functions for each of the cluster.
- This is an extension to the K Means PLS model if it is formulated in terms of a membership function.

Model Training



Training X and Y Data

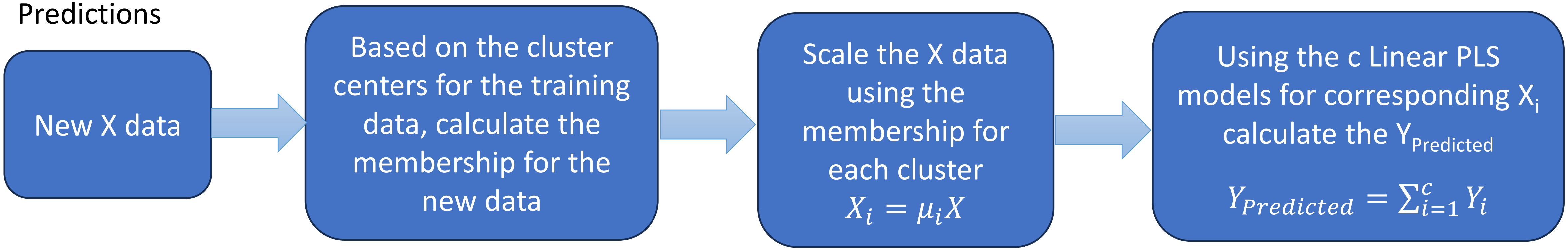
Fuzzy C Means Clustering on the X data gives c cluster centers and a membership matrix μ

Scale the X and Y data by the membership values for each of the clusters

$X_i = \mu_i X,\ Y_i = \mu_i Y$

Train c Linear PLS models between X_i and Y_i

Predictions



New X data

Based on the cluster centers for the training data, calculate the membership for the new data

Scale the X data using the membership for each cluster

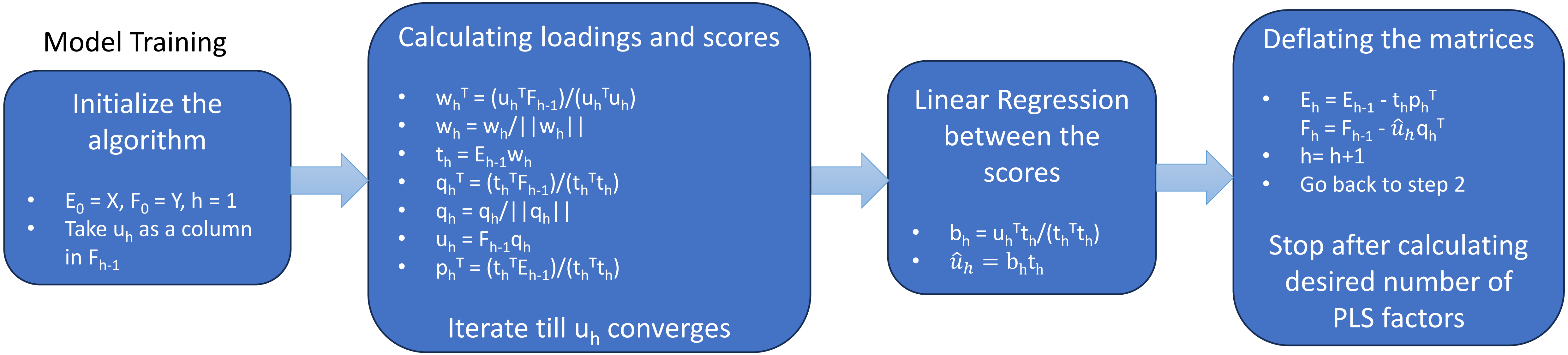
$X_i = \mu_i X$

Using the c Linear PLS models for corresponding X_i calculate the $Y_{Predicted}$

$Y_{Predicted} = \sum_{i=1}^{c} Y_i$

NIPALS Algorithm

Model Training



Initialize the algorithm

- E_0 = X, F_0 = Y, h = 1
- Take u_h as a column in F_{h-1}

Calculating loadings and scores

- $w_h^T = (u_h^T F_{h-1})/(u_h^T u_h)$
- $w_h = w_h/||w_h||$
- $t_h = E_{h-1} w_h$
- $q_h^T = (t_h^T F_{h-1})/(t_h^T t_h)$
- $q_h = q_h/||q_h||$
- $u_h = F_{h-1} q_h$
- $p_h^T = (t_h^T E_{h-1})/(t_h^T t_h)$

Iterate till u_h converges

Linear Regression between the scores

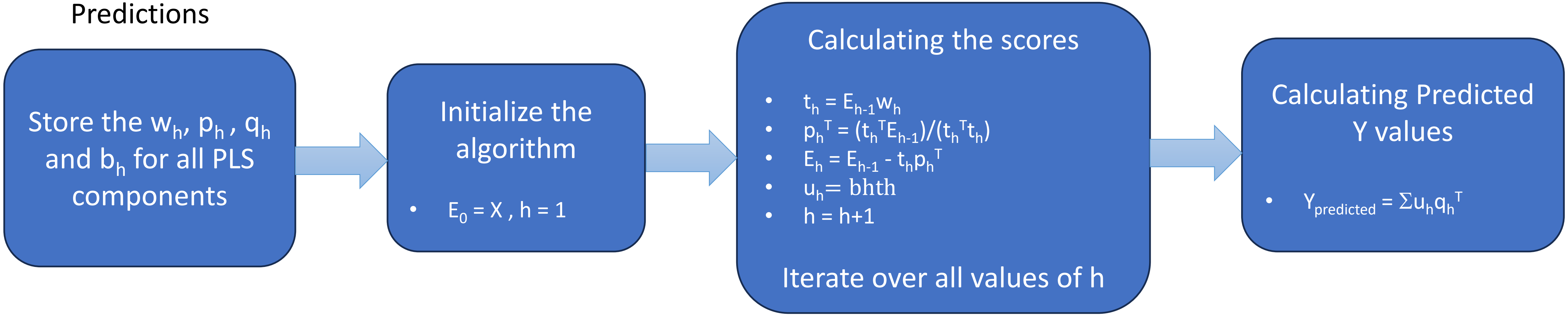
- $b_h = u_h^T t_h/(t_h^T t_h)$
- $\hat{u}_h = b_h t_h$

Deflating the matrices

- $E_h = E_{h-1} - t_h p_h^T$
  $F_h = F_{h-1} - \hat{u}_h q_h^T$
- h= h+1
- Go back to step 2

Stop after calculating desired number of PLS factors

Predictions



Store the w_h, p_h , q_h and b_h for all PLS components

Initialize the algorithm

- E_0 = X , h = 1

Calculating the scores

- $t_h = E_{h-1} w_h$
- $p_h^T = (t_h^T E_{h-1})/(t_h^T t_h)$
- $E_h = E_{h-1} - t_h p_h^T$
- u_h= bhth
- h = h+1

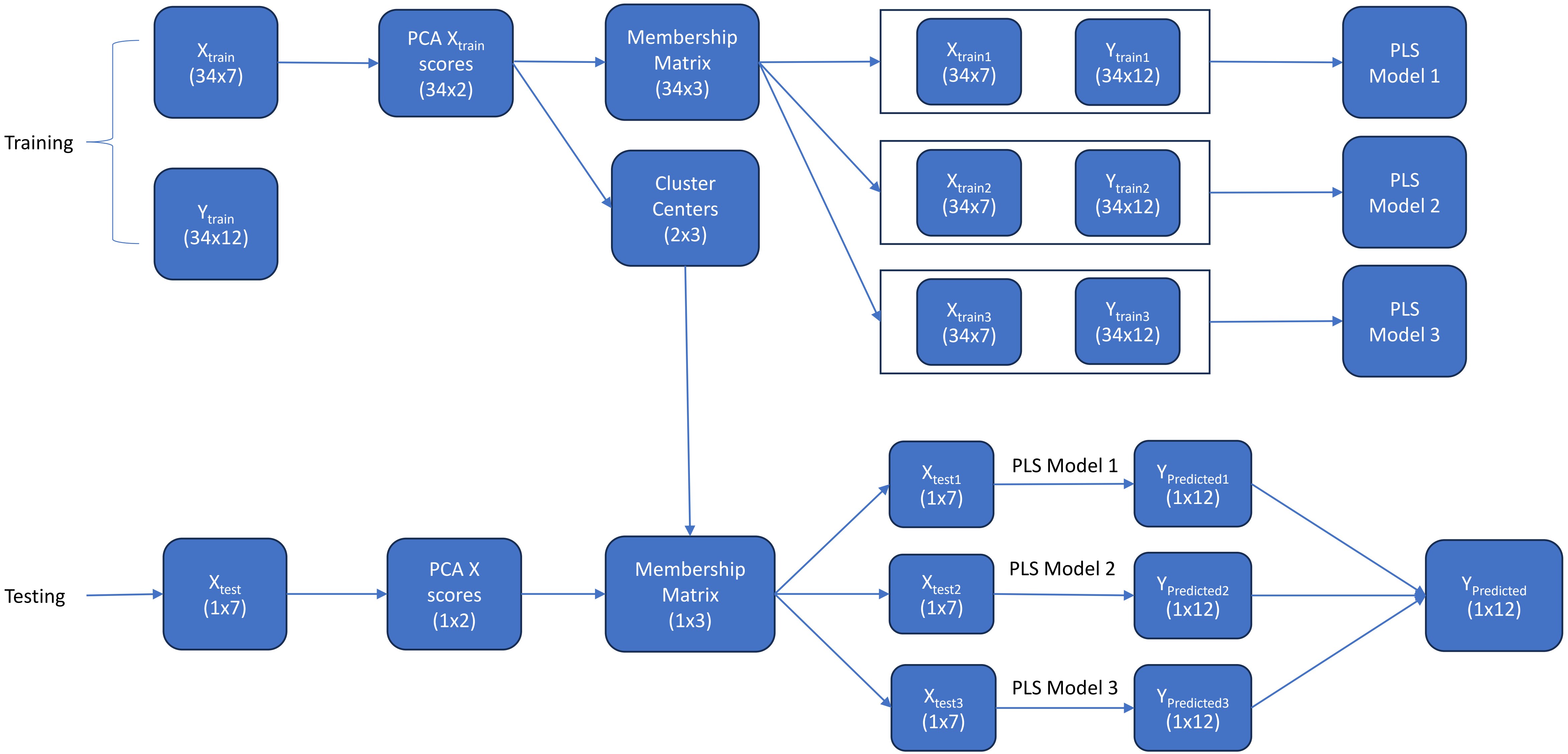
Iterate over all values of h

Calculating Predicted Y values

- $Y_{predicted} = \Sigma u_h q_h^T$

Fuzzy Clustering Based PLS (Our Suggestion)

ENGINEERING Chemical Engineering | McMaster University



Microgel Dataset

Dataset -

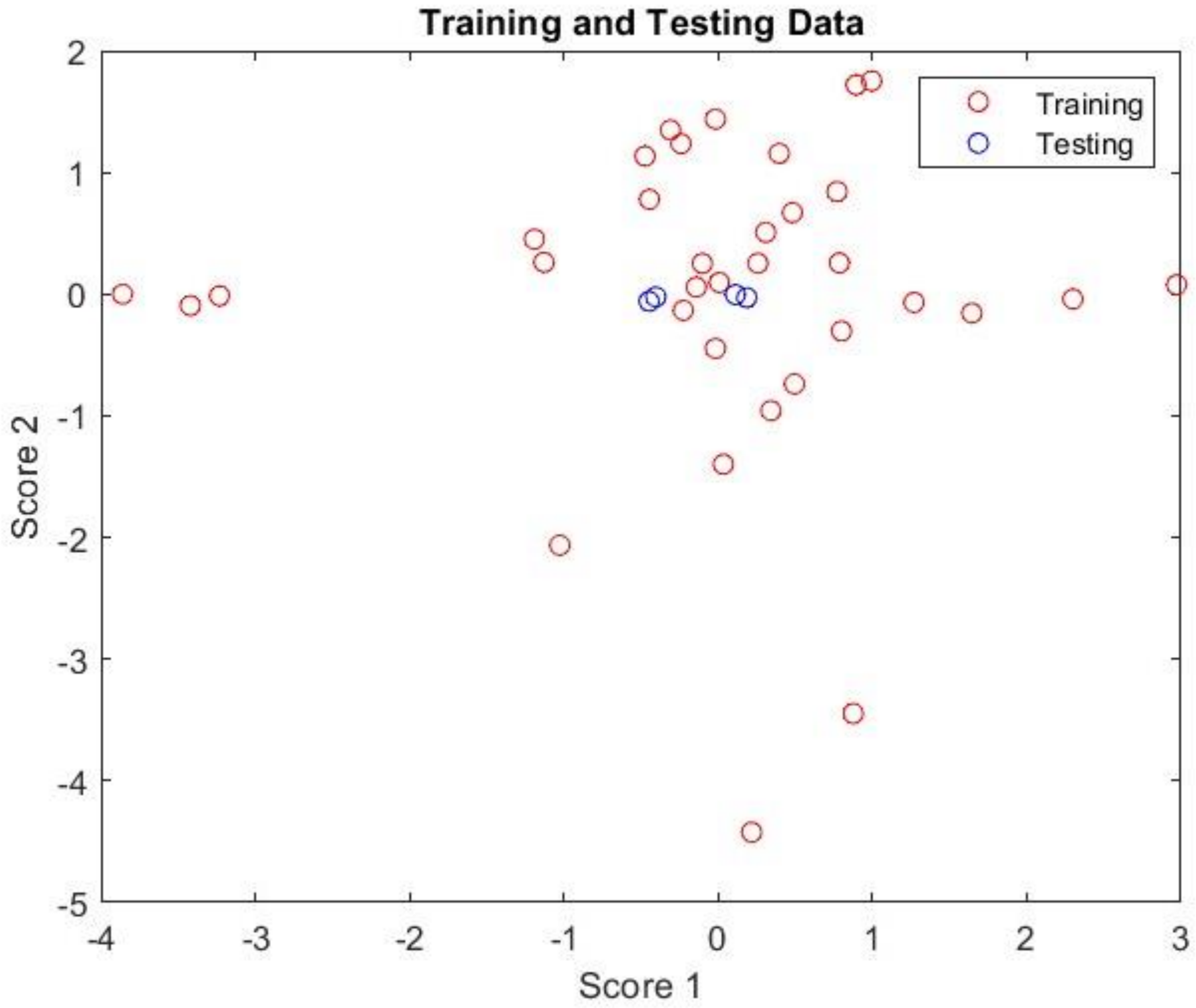
- 34 observations – Training
- 4 observations - Testing
- X = 7 variables
- Y = 12 variables

Parameters -

- Number of PLS Factors = h
- Number of Clusters = c
- Fuzziness Parameter = f
    - The objective function of FCM clustering is to minimize J = $\sum(\sum(u_{ij})^f * ||x_i - c_j||^2)$, f > 1. As f increases, the overlap between clusters increases.

Models -

- Linear PLS
- Outer Fuzzy PLS (Our suggestion)
- Outer Fuzzy PLS 2 (From literature)
- Inner Fuzzy PLS (From literature)
- K Means PLS (Saeid's First Phase)



Testing Methods

- **JackKnife Approach** -
  - Exclude one observation from the set for testing and develop the model using the remaining which is repeated for all the observations.
  - In this way, each time the predictions are made by the model for an unseen sample which is not included in the training set.

- **Testing Data Set** -
  - Split the data set into train and test sets, develop the model on the training set and use the test set for prediction.
  - We have used 34 samples of the dataset for training and tried to predict 4 new samples which Saeid has recently synthesized in the lab.

- For the JackKnife Approach, we go through different combinations of the parameters and report the parameters which give the best predictions.
- These parameters are used for predictions on the test set for comparison of the models

Results: Swelling Profile

JackKnife Approach

Type	R^2	MSE	h	c	f
Linear	0.6653	7430	5	-	-
Outer Fuzzy	0.69835	6720.415	5	2	2
Inner Fuzzy	0.66051	7489.425	5	2	1.4
K Means	0.66912	7385.598	7	2	-
Outer Fuzzy 2	0.67189	7565.56	4	3	2

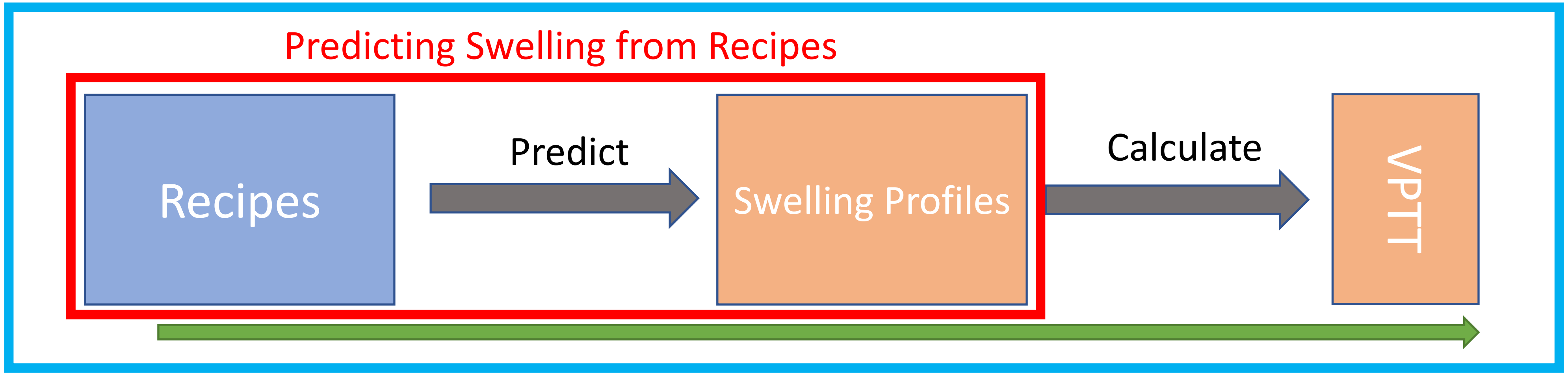
Testing

Type	R^2	MSE	h	c	f
Linear	0.69874	5684.7	5	-	-
Outer Fuzzy	0.70147	5716.4	5	2	2
Inner Fuzzy	0.69423	5940.7	5	2	1.4
K Means	0.69715	5511.3	7	2	-
Outer Fuzzy 2	0.39106	7057.5	4	3	2

Calculating the VPTT

ENGINEERING
Chemical Engineering

McMaster University



- From the swelling profile, the VPTT is found by fitting a sigmoid curve.
- For a given swelling profile, we get the VPTT and the size of the microgel at a pH of 4 and 10, the VPTT output block hence contains 4 variables.
- The VPTT block is calculated for the actual and the predicted swelling profile and the error between these two is the prediction error.
- To compare the prediction errors for different approaches, we combine the error in the 4 output variables as a weighted sum of the errors in each of the variables.
- The variables are weighted by the inverse of the range of the values of the corresponding variable in the VPTT for the true swelling profile.

Results: VPTT

ENGINEERING
Chemical Engineering
McMaster University

JackKnife Approach

Type	RMSE	h	c	f
Linear	1.2592	4	-	-
Outer Fuzzy	1.02676	3	5	2
Inner Fuzzy	1.12892	3	5	1.2
K Means	1.2019	3	2	-
Outer Fuzzy 2	1.41810	3	2	1.2

Testing

Type	RMSE	h	c	f
Linear	8.5145	3	-	-
Outer Fuzzy	8.4422	3	5	2
Inner Fuzzy	8.6813	3	5	1.2
K Means	8.3581	3	2	-
Outer Fuzzy 2	8.3981	3	2	1.2

Conclusions and Future Work

- Till now, the results which we have got show that our suggested method outperforms other methods in the Jackknife Approach.
- But on the Testing Dataset, the results for our method are similar to the other methods used, the results being dependent on the selected testing data which could be because of the size of the data and the train test split.
- The method to combine the VPTT errors which has been used can be used only to compare the results for different PLS methods on a given X data
- Considering different training and testing splits of the dataset and use the JackKnife + Testing approach for each of them and then compare the results.
- Change the VPTT combined error definition such that it can be used to compare the results for different inputs to the models.

THANK YOU