Partial Least Squares Regression Analysis of CHO Cell Batch Cultures

Anirudh Rao

Batch Culture Data

The dataset contains time course data for two CHO cell lines, Clone 1 and Clone 2, which are grown in two different media A and B. These cell lines have been engineered to produce an IgG antibody. The dataset contains data about the viable cell density, IgG titer, and the concentrations of amino acids, ammonia, potassium, glucose, and lactate in the cell culture media, which were measured over a course of 8 days.

Specific Rates

Using the given concentration data, we compute the specific growth rates, specific productivities, and specific rates of the substrates, for all cell lines on all days except Day 0.

$$\text{Specific growth rate} = \mu = \frac{\ln\left(\frac{x}{x_0}\right)}{\Delta t}\ \mathrm{day}^{-1}$$

$$\text{Specific productivity} = \mu \frac{\Delta[\text{Product}]}{\left(\frac{x+x_0}{2}\right)}\ \mathrm{pg\ cell^{-1}\ day^{-1}}$$

$$\text{Specific rate} = \mu \frac{\Delta[\text{Metabolite}]}{\left(\frac{x+x_0}{2}\right)}\ \mathrm{pmol\ cell^{-1}\ day^{-1}}$$

Here, $\left(\frac{x+x_0}{2}\right)$ is the average biomass concentration (in $\mathrm{cells\ mL^{-1}}$) during the time period.

As the given substrate concentration data was given in units of $\mathrm{g\,L^{-1}}$ or $\mathrm{mg\,L^{-1}}$, the molar masses of all substrates were used to convert the concentration in terms of $\mathrm{pmol\ mL^{-1}}$.

To understand the correlation between the various substrate rates with the growth rate and productivity, the absolute value of their Pearson correlation coefficients were computed. This has been visualised in Figure 1. Ammonia, glucose, lactate, glutamine, glutamate, glycine, and alanine seem to have good correlation with the growth rate.

On the other hand, the productivity has good correlation with the specific rates of ammonia, glucose, lactate, glutamine, glutamate, serine, alanine, glycine, tyrosine, lysine, leucine, and phenylalanine.

Partial Least Squares Regression

To predict the growth rate and productivity from the metabolite specific rates, two partial least squares regression models were built using Python's `sklearn` library. This is a statistical technique that combines linear regression, principal component analysis, and correlation analysis to model multiple response variables and analyse data with collinear variables.

The metabolite specific rates were set as the X variable and the growth rate and productivity were set as the y variable. The regression was performed using the default value of `n_components` = 2.

The partial least squares model achieved a root mean square error (RMSE) of 0.33 day^{-1} for the growth rate and 3.95 $\mathrm{pg\ cell^{-1}\ day^{-1}}$ for the productivity. The model achieved an R^2 score of 0.71 for the growth rate and 0.87 for the productivity, indicating good performance of the model. The predictions are visualised in Figure 2 and Figure 3.

To assess the importance of each metabolite in influencing the prediction of the growth rate and productivity, their variable importance in projection (VIP) scores were computed and tabulated in Table 1. These are visualised in Figure 4 and Figure 5. For growth rate prediction, glycine and glucose rates have the highest VIP scores, followed closely by ammonia, glutamine, lactate, and glutamate. For productivity prediction, glucose, glutamine, ammonia, lactate, and glutamate have the highest VIP scores. The metabolite rates with high VIP scores are the same for growth rate and productivity prediction, with the exception of glycine. These are also the metabolite rates that have good correlation with the growth rate and productivity.

Another method to assess the contribution of each metabolite rate to the final model prediction is by analysing their coefficients in the model. Metabolites with a higher coefficient, either positive or negative, have a higher contribution. These are tabulated in Table 2 and visualised in Figure 6 and Figure 7.

For growth rate prediction, the metabolite rates with the highest coefficients are glycine, aspartate, and phenylalanine. For productivity prediction, the metabolite rates with the highest coefficients are phenylalanine, isoleucine, and glutamate.

Phenylalanine appears to influence the prediction of both the growth rate and productivity. Glutamate has a high VIP score and a high coefficient value. However, in general, the metabolite rates deemed important by the VIP score and the coefficient values are different.

Conclusion

By analysing the specific rates of the batch culture data using PLS regression, we are able to predict the performance characteristics of the cell types in different media. We are also able to assess the importance of different metabolites in making these predictions.

Annexure

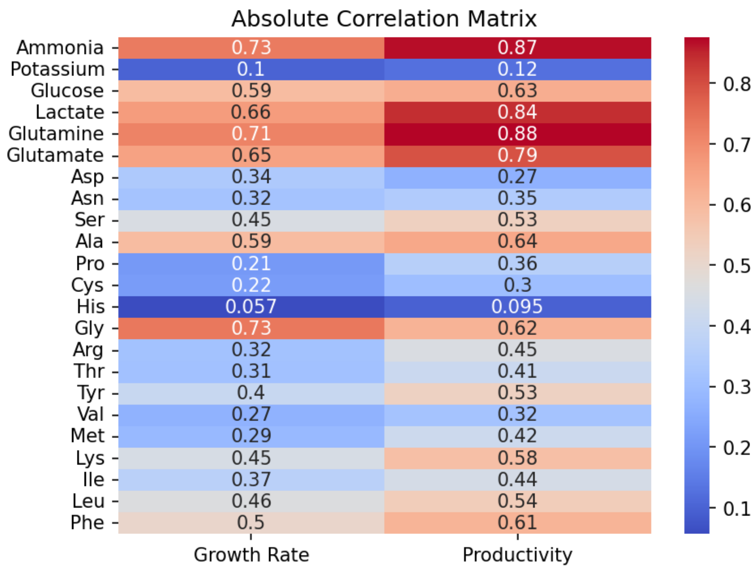

Figure 1: Correlation between rates

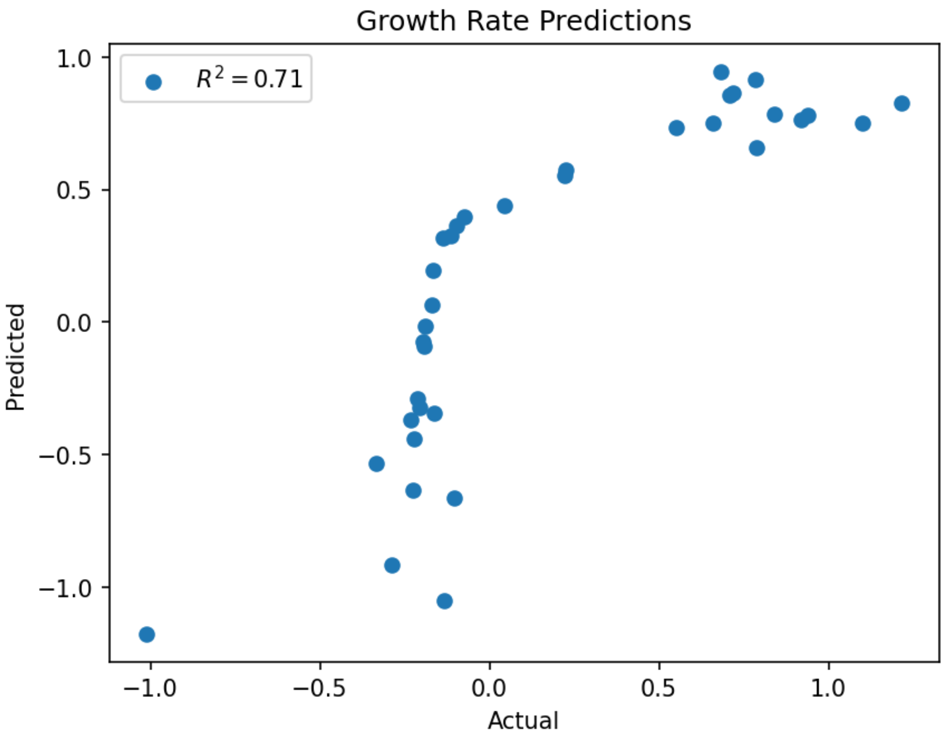

Figure 2: Growth rate prediction using PLS regression

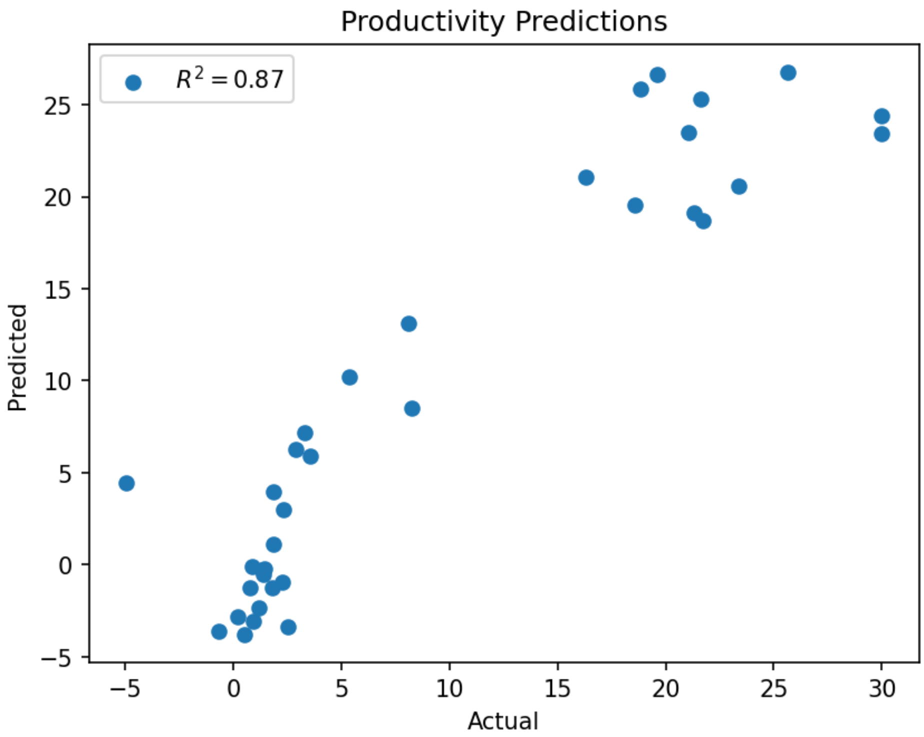

Figure 3: Productivity prediction using PLS regression

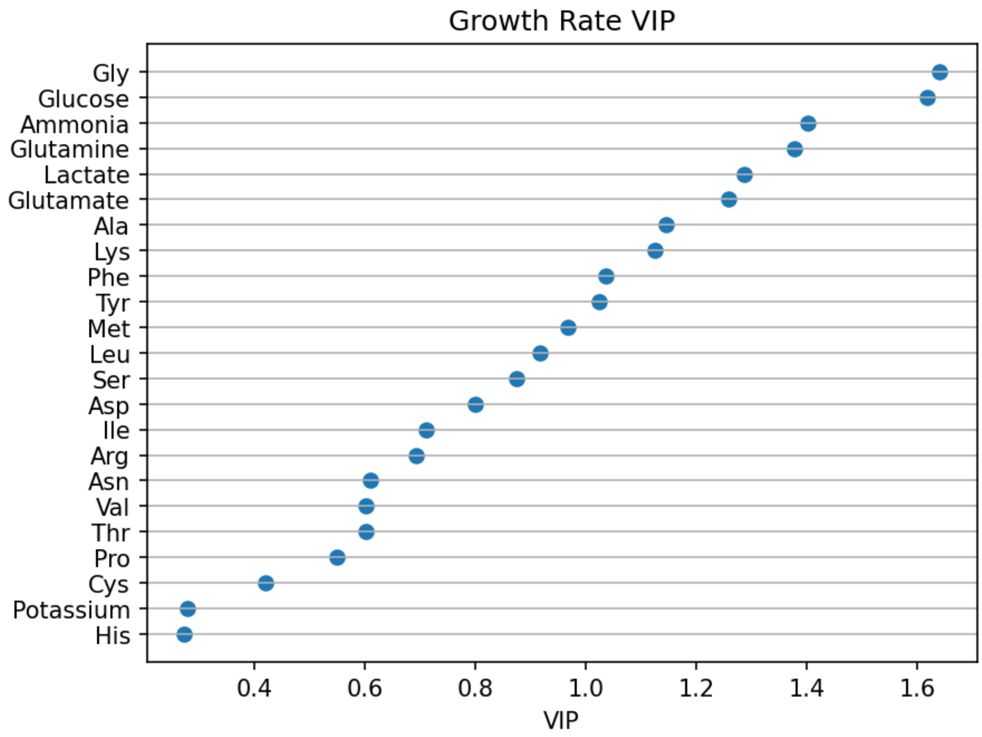

Figure 4: VIP score plot for growth rate prediction using PLS regression

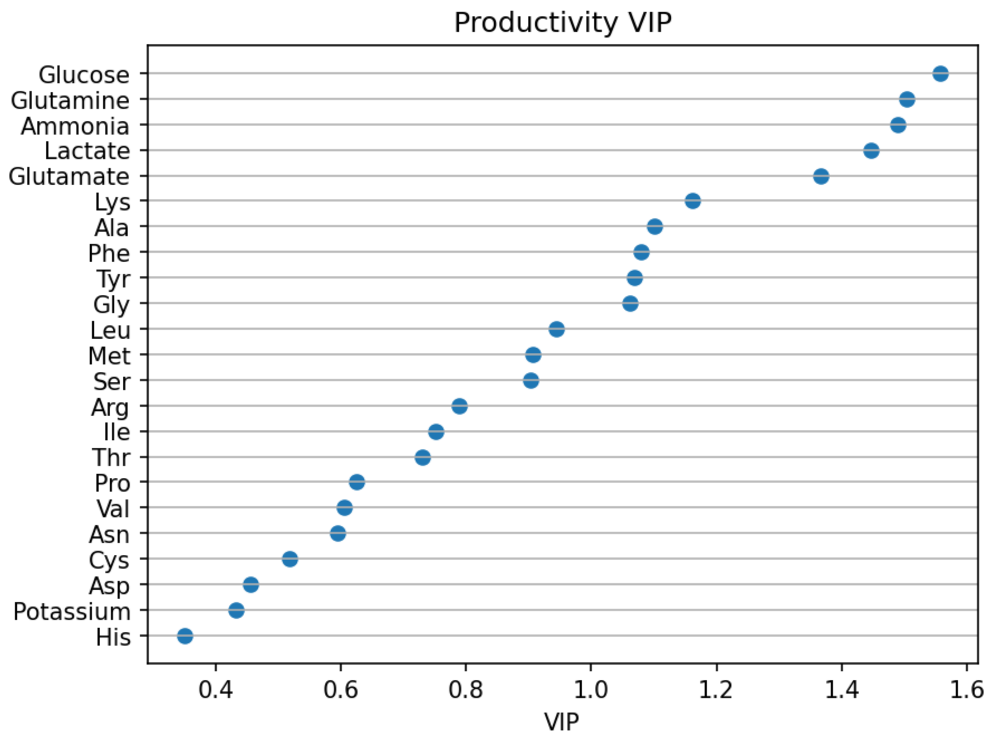

Figure 5: VIP score plot for productivity prediction using PLS regression

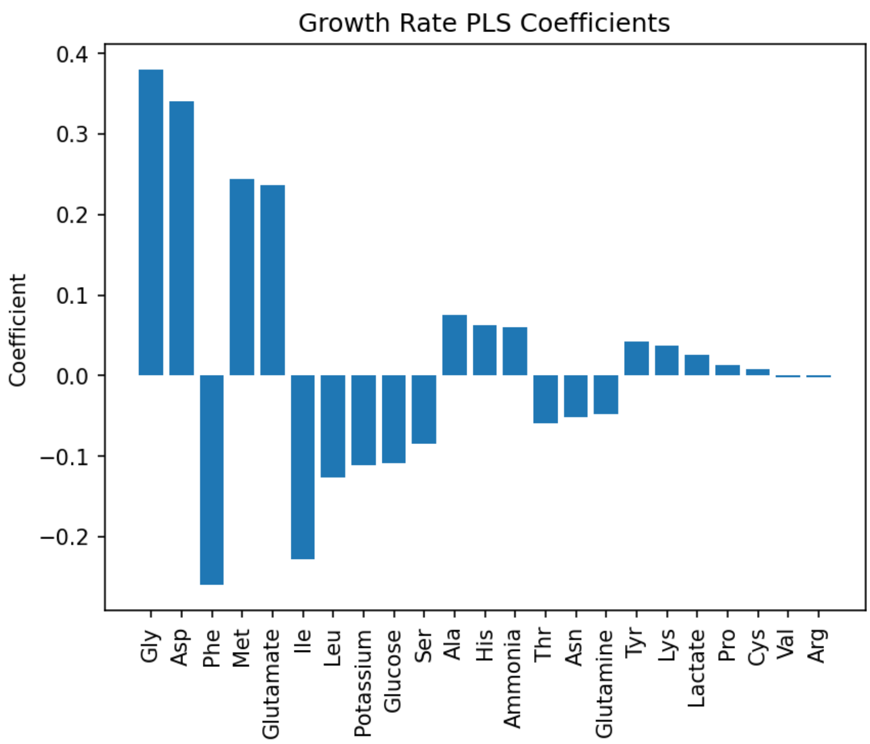

Figure 6: Coefficient plot for growth rate prediction

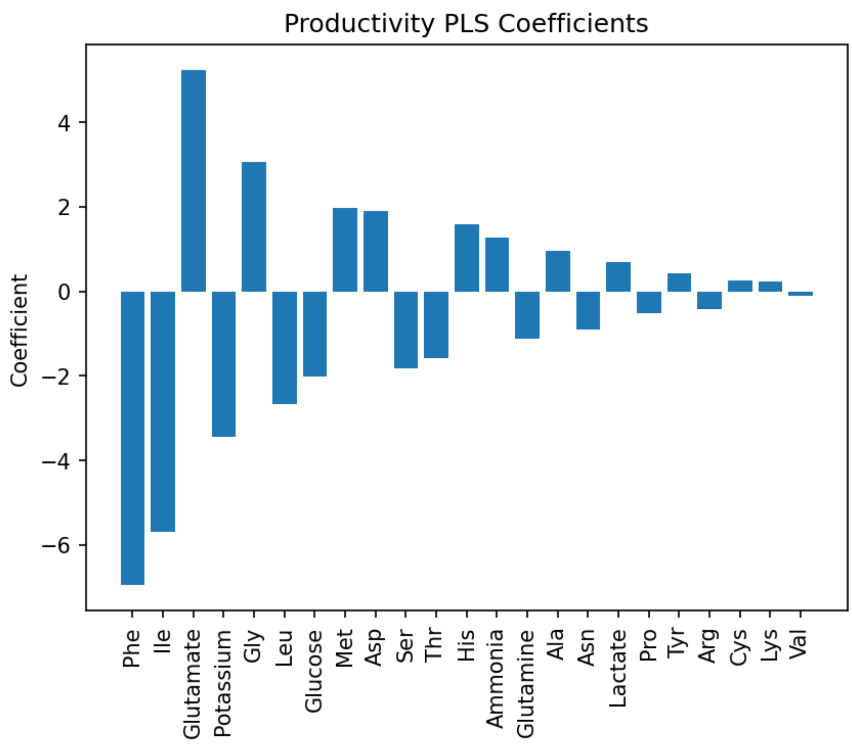

Figure 7: Coefficient plot for productivity prediction

Table 1: VIP scores

Component	Growth Rate VIP score	Productivity VIP score
Ammonia	0.273066	0.351259
Potassium	0.278871	0.432167
Glucose	0.419193	0.455952
Lactate	0.548391	0.517595
Glutamine	0.601241	0.595517
Glutamate	0.601922	0.604521
Asp	0.60923	0.624272
Asn	0.692067	0.730525
Ser	0.710445	0.751957
Ala	0.800599	0.788662
Pro	0.874741	0.903724
Cys	0.916202	0.90606
His	0.967564	0.943871
Gly	1.023368	1.062071
Arg	1.035367	1.068831
Thr	1.124328	1.079785
Tyr	1.145977	1.099962
Val	1.258841	1.160932
Met	1.286707	1.365584
Lys	1.378142	1.446991
Ile	1.401959	1.48931
Leu	1.618272	1.503671
Phe	1.640218	1.556573

Table 2: PLS model coefficients

Component	Growth Rate Coefficient	Productivity Coefficient
Ammonia	0.060384	1.281181
Potassium	-0.11133	-3.43928
Glucose	-0.10823	-2.00388
Lactate	0.025643	0.697375
Glutamine	-0.04717	-1.12726
Glutamate	0.236562	5.232483
Asp	0.340785	1.896357
Asn	-0.05119	-0.91422
Ser	-0.08438	-1.81194
Ala	0.075669	0.94671
Pro	0.013315	-0.51747
Cys	0.008275	0.260366
His	0.062625	1.575633
Gly	0.379649	3.066281

Arg	-0.00197	-0.42843
Thr	-0.05927	-1.58395
Tyr	0.04277	0.437646
Val	-0.00235	-0.10714
Met	0.243537	1.980084
Lys	0.036852	0.229718
Ile	-0.22859	-5.69582
Leu	-0.126	-2.66629
Phe	-0.25998	-6.95186
Intercept	0.169507	9.003848

All supplementary files associated with this project can be found here –

https://drive.google.com/drive/folders/1mQyDUxYDLId9xy8jDjf1-OqAsqJWa47d?usp=sharing