Dissolved Oxygen Gradient as an Indicator in Cell Differentiation Using Monte Carlo Simulated Curve-Fitting

Kalpana Baheti, Graduate Student, Georgia Tech

ABSTRACT

In this academic article we study the relationship between the gradient of dissolved oxygen (DO) and the final concentration of cardiomyocyte (CM) cell content in the culture. The DO decrease per duration, as per clinical theory, acts as an analytical indicator of the final CM concentration in the culture. The work in this paper assesses the data from the 5th day of DO gradients along the ten day cell differentiation process and its relationship with the CM content on the tenth day of differentiation. The takeaway of this work covers three points. One, when using monte carlo simulations, posteriors are to be proposed and that warrants experimentation in order to arrive at a suitable estimated posterior to fit into; this also requires a fair bit of imagination of data that may not be there today but might be tomorrow. Two, for a posterior distribution that is hypothesized, a substantial part of fine-tuning the posterior curve rests on the hyperparameters, which is where the PyMC simulations will come in handy. And last, with our candidate posteriors, we will assess the errors and settle on one conclusive relationship on DO gradient that will serve as a baseline indicator of CM content.

INTRODUCTION

Context -

The heart is the least regenerative organ of the body and when infarction impacts the myocardium (native cardiac muscle) tissue, instead of new cardiomyocyte cells replacing damaged cells, it is patched up with fibrotic scar tissue. Recent years have shown substantial progress in clinical research with human pluripotent cells hPSC and stimulating their differentiation into developed cardiomyocyte (CM) cells and consequently engineered tissue. However, to be realistically used on patients in a scaled and trustworthy manner, the reproducibility of this method needs to be ensured and mass production with use of bioreactors would need to be studied and regulated.

One of the major analytical indicators of end CM content of the cell differentiation process is dissolved oxygen usage, since cell differentiation utilizes and needs this oxygen while some of it may also interact with the rest of the bioreactor setup.

With the establishment of a robust relationship between DO gradients and CM content, it would be possible for adjusting other components of the setup to maintain a certain level of DO, or otherwise, the DO gradients could also be monitored and plugged into this ready relation well before the final day and the general estimate of eventual CM content may be determined and timely changes may be performed on the bioreactor setup within the remaining days as possible.

Problem Statement -

In this article, we try to determine the closest mathematical relationship that associates the DO gradient % with CM concentration %. We will address this in the following steps -

1. Constraining the domain (and consequently the codomain).
2. Excluding all major families of curve-fits that under no circumstances will fit.
3. Testing each of the remaining fits exhaustively.
4. Per each candidate fit, utilize bayesian priors and likelihoods to tune hyperparameters.
5. Run deduced function on test data, evaluate error on both train and test sets.

After exhaustively testing relationships with the help of PyMC, we will then decide which relationship appears the most reliable.

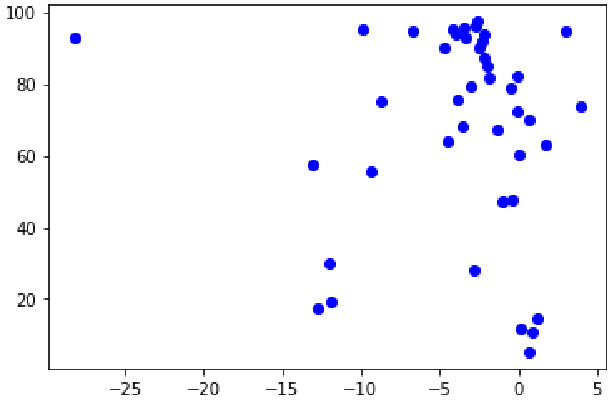
DATA VECTORS

The training set contains 42 points whereas the test set contains 18 points. There are two columns in the dataset, one for the factor; dissolved oxygen gradient, and one for the dependent which is the CM content on the tenth day. This is the scatter plot of the training and test set -

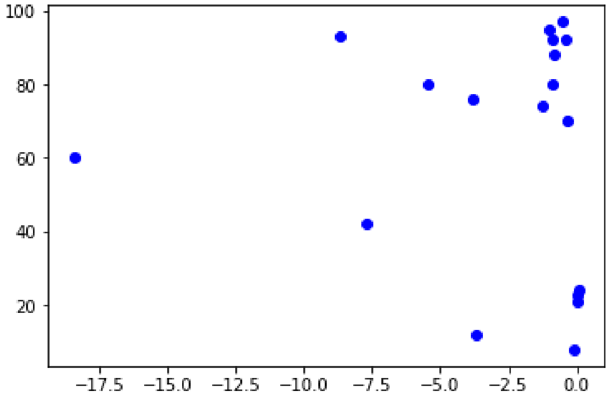
Legend -

X = DO Gradient
Y = CM Content

Training Set –

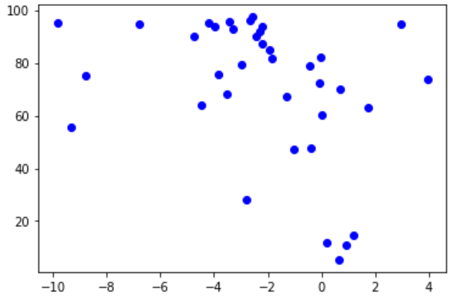


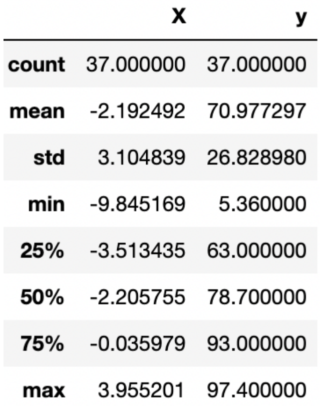
Testing Set –



After removal of outliers –

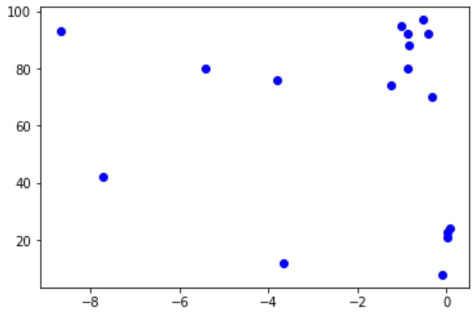
Training Data Complete Information –

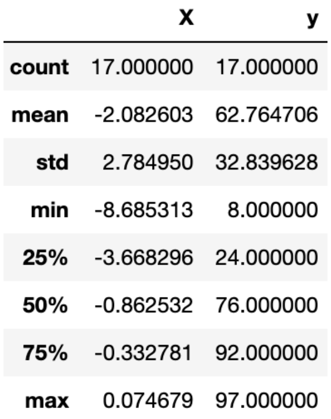




	X	y
count	37.000000	37.000000
mean	-2.192492	70.977297
std	3.104839	26.828980
min	-9.845169	5.360000
25%	-3.513435	63.000000
50%	-2.205755	78.700000
75%	-0.035979	93.000000
max	3.955201	97.400000

Testing Data Complete Information –

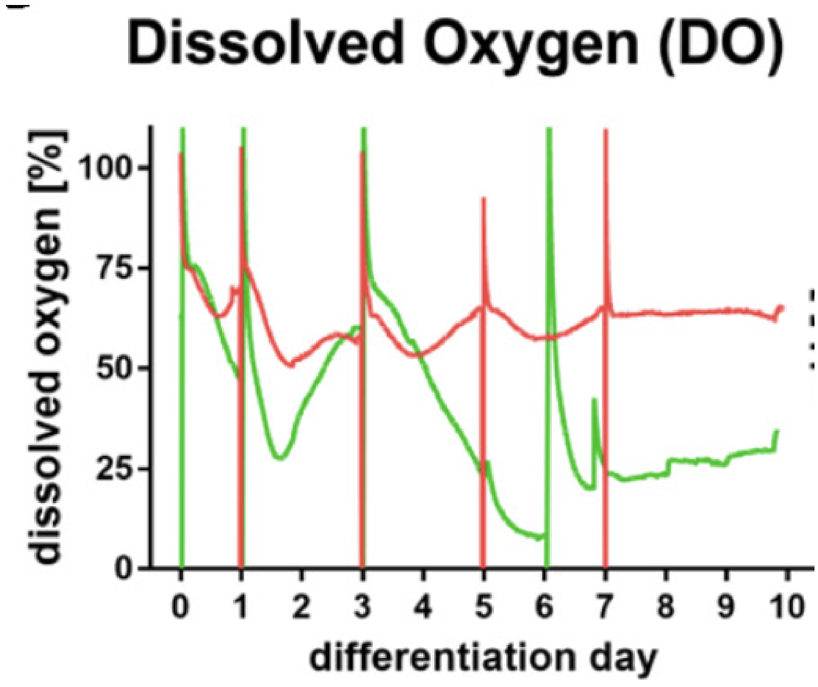




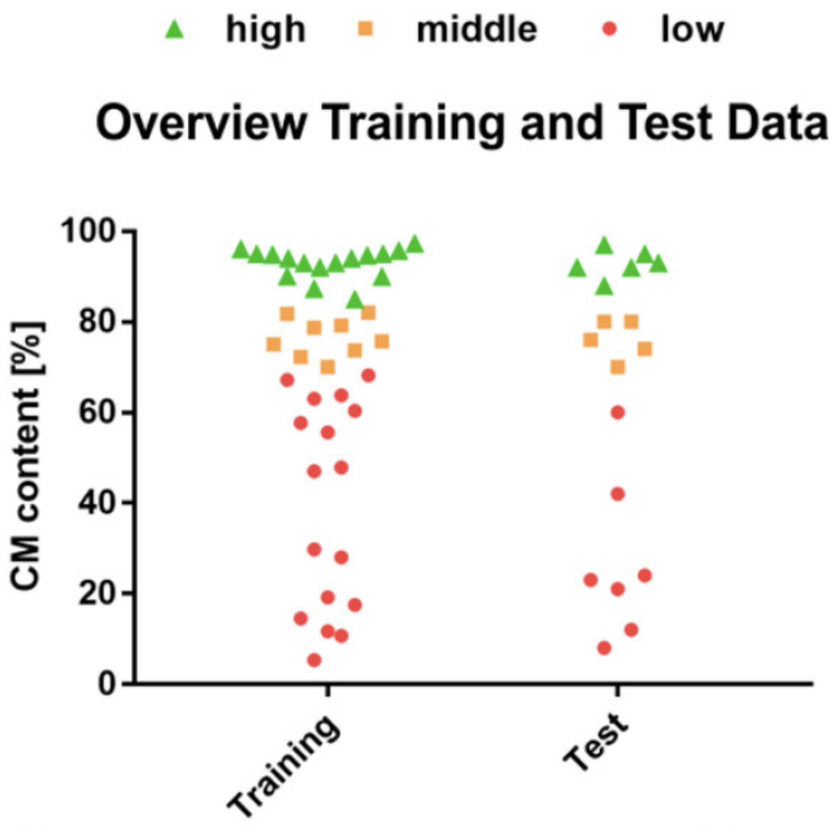
	x	y
count	17.000000	17.000000
mean	-2.082603	62.764706
std	2.784950	32.839628
min	-8.685313	8.000000
25%	-3.668296	24.000000
50%	-0.862532	76.000000
75%	-0.332781	92.000000
max	0.074679	97.000000

The dataset was extracted from the following source – https://github.com/CremaschiLab/Cardiac_Differentiation_Modeling

The gradients of DO as depicted on the graph with absolute DO content per day –



The green line indicates all the cases where end CM content was high (desirable case), and the red line indicates cases where the CM content was low. The CM content follows this quality graph –

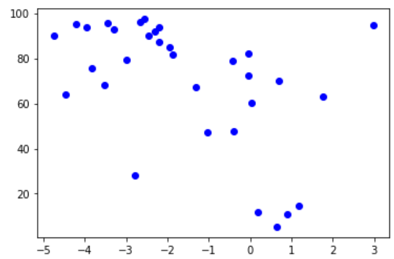


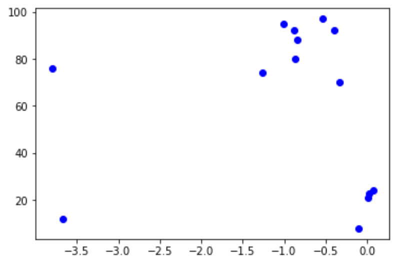
POSTERIOR HYPOTHESES AND HYPERPARAMETER TUNING VIA PyMC

Constraining the domain –

We will do this in two phases;

1. One that is cleaned of outliers – which has already been done..
2. After adjusting to 90% HDI of the factor distribution (train and test respectively) –





All continuous posterior curve-fits ***not*** further pursued and excluded -

On preliminary analysis, the following were ruled out for major deviance, sign of factor values, and also based on visualizing the general distribution of data points - Gamma, InvGamma, Exponential, StudentT, Pareto, ChiSquare, LogNormal, Laplace, Logarithmic, and Weibull.

The set of curves that made for promising in-depth trial candidates were -

Linear, Degree-2 Polynomial, Degree-3 Polynomial, Normal, Cauchy

Degree-5 polynomials showed promise too, but overfit on the densest region while performing very poorly on the rest of the data pushing up the RMSE. And skewed normal which might have been a promising candidate was not tested in this study.

Posterior curve and hyperparameter tuning -

Each of the posterior curves were fit on the training set by fine-tuning the hyperparameters of those curves. The hyperparameters were informative but that was not automated but a custom tertiary optimization procedure was followed to narrow in on the best performing hyperparameters. The procedure was as follows -

Step 1: Input starting value of mean and standard deviation of hyperparameters.
Step 2: Determine general direction of decrease in RMSE per hyperparameter in the first trial.
Step 3: A step per hyperparameter per mean/standard deviation is input and all discrete combinations are run in the improving direction.
Step 4: Set of hyperparameters is returned that satisfies a threshold of less than 100 in RMSE.

Then with these hyperparameters, the following PyMC structure is followed to simulate monte carlo bayesian sampling followed by bayesian fitting -

```
with pm.Model() as m:

        #Model Data
```

```
X_data = pm.Data("X_data", x_train)
y_data = pm.Data("y_data", y_train/150)

#Hyperparameter Tuning
hyperparameter_1 = pm.Normal(h1, derived_mu1, derived_sigma1)
hyperparameter_2 = pm.Normal(h2, derived_mu2, derived_sigma2)
.
 .
.
hyperparameter_n = = pm.Normal(hn, derived_mun, sigma=derived_sigman)

#Posterior Curve Selection
y_pred = proposal_posterior_function( X_data, hyperparameters {1,…,n} )
y_vals = pm.Deterministic('y_pred', y_pred)

#Curve-Fitting
likelihood = pm.Normal('likelihood', mu=y_vals, sigma=0.01, observed=y_data)

#Monte Carlo Sampling Run
trace = pm.sample(2000)
```

For the following experiments, consider -

Y: CM content on differentiation day 10
X: DO gradient on differentiation day 5 for different chemical setups

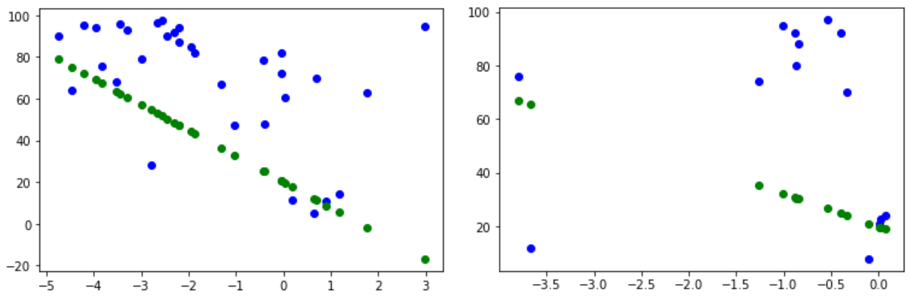
1. **Linear Curve-Fit -**

   Y = **-12.368 (X)** + **20**

   Training RMSE: **40.36737059**

   Testing RMSE: **45.96555175**

   Training Fit and Testing Projected Respectively (Green Lines):



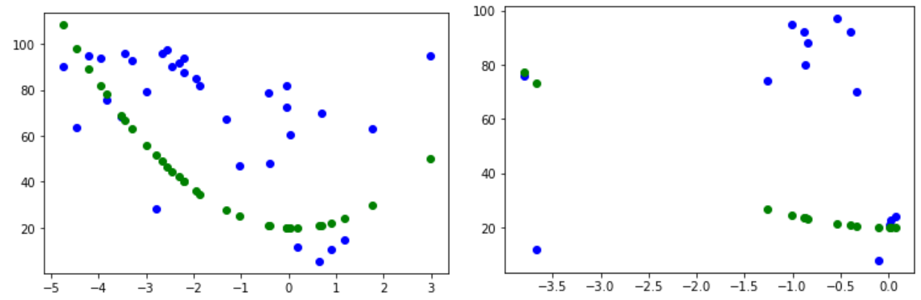
2. **Degree-2 Curve-Fit -**

$\mathbf{Y = 3.714\,(X)^2 - 0.926\,(X) + 20}$

Training RMSE: **37.05303234**

Testing RMSE: **50.90952825**

Training Fit and Testing Projected Respectively (Green Lines):



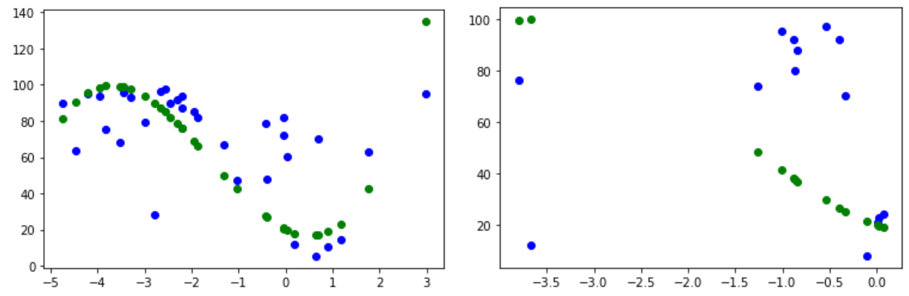
3. **Degree-3 Curve-Fit –**

$\mathbf{Y = 2.234\,(X)^3 + 10.628\,(X)^2 - 12.812\,(X) + 20}$

Training RMSE: **27.67429955**

Testing RMSE: **46.40122019**

Training Fit and Testing Projected Respectively (Green Lines):



4. **Normal Curve-Fit –**

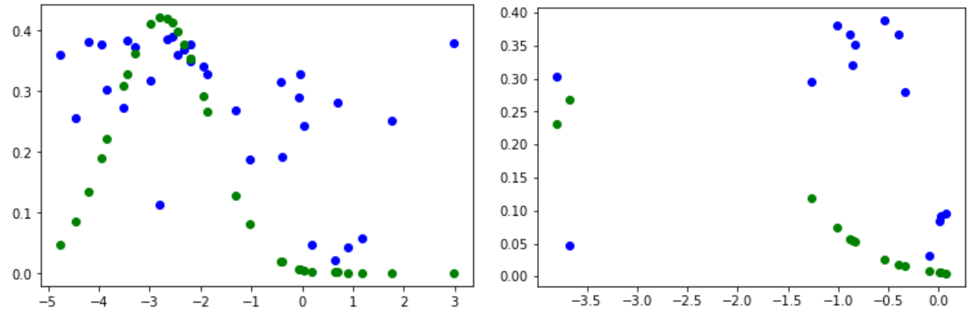
$\mu = \mathbf{-2.766}$
$\sigma = \mathbf{1.057}$

$\mathbf{Y = 250\ (\sqrt{\sigma^2/2\pi}\,)\ exp\ \{\ (-\sigma^2/2)\ (X - \mu\ )^2\ \}}$

Training RMSE: **74.84163393**

Testing RMSE: **69.33281511**

Training Fit and Testing Projected Respectively (Green Lines):



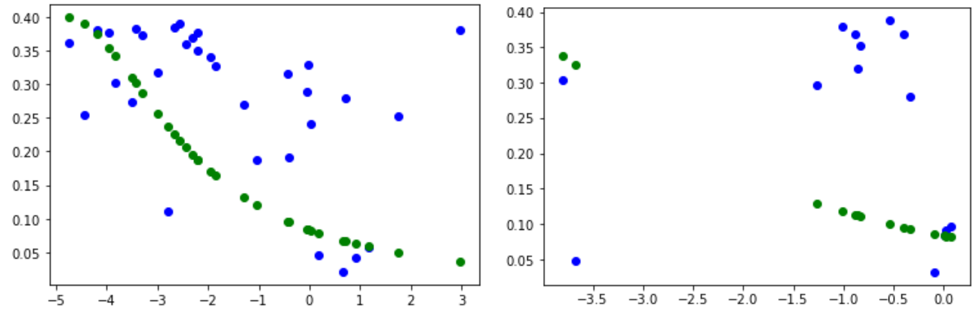
5. Cauchy Curve-Fit -

α = **-4.863**
β = **2.501**
$\mathbf{Y = 250\ [\ \{\ ((\ X - \alpha\ )\ /\ \beta\ )^2 + 1\ \}\ \beta\ ]^{-1}}$

Training RMSE: **74.84389859**

Testing RMSE: **69.27700142**

Training Fit and Testing Projected Respectively (Green Lines):



TRAIN ERROR VS. TEST ERROR

Tabular Results -

Posterior Selection	**Train Error**	**Test Error**
Linear	40.36737059	45.96555175
Quadratic	37.05303234	50.90952825
Cubic	27.67429955	46.40122019
Normal	74.84163393	69.33281511
Cauchy	74.84389859	69.27700142

CONCLUSIONS AND CURVE SELECTION

Conclusion I:

Based on these observations, the best posterior curve selection based on training performance is the cubic curve - $Y = 2.234 (X)^3 + 10.628 (X)^2 - 12.812 (X) + 20$ - with a training error minimum of 27.67. However, there is a decrease in Bayesian predictive performance of the test set.

Conclusion II:

If the cubic polynomial is to be established more firmly, the next series of experiments set up would roughly require trials focused with a DO gradient simulated between -1.5 % and 0.5 % to offset the disturbance caused on the test set.

Conclusion III:

If the recommended trails given in conclusion III do not weigh in on the cubic curve, it means the observed target values show higher value of CM content than the cubic curve foresees.

Hence, the next closest relationship between the two for timely analytical indication would be the linear relationship.

Conclusion IV:

After the cubic relationship, the linear fit has captured the posterior curve the best way possible, since the test and train set Bayesian predictive performances are the most mutually consistent, as well comparatively better in terms in RMSE than the other curves. The normal and cauchy curves have performed almost equally poorly on both test and train sets. And the quadratic curve has performed less than satisfactory on both tests and train sets and suffers large test-train performance discrepancy.

These conclusions in combination help either narrow down the hypothesized relationship between DO gradient % and CM % to a cubic polynomial relationship, or warrant further experiments to disprove this relation, baseline at the linear relationship and then start to reconsider excluded posterior curves.

REFERENCES

1. *Williams B, Löbel W, Finklea F, Halloin C, Ritzenhoff K, Manstein F, Mohammadi S, Hashemi M, Zweigerdt R, Lipke E and Cremaschi S (2020) Prediction of Human Induced Pluripotent Stem Cell Cardiac Differentiation Outcome by Multifactorial Process Modeling Front. Bioeng. Biotechnol. 8:851. doi: 10.3389/fbioe.2020.00851*
2. *Dataset Source: https://github.com/CremaschiLab/Cardiac_Differentiation_Modeling*
3. *PyMC Modeling Pragmatics: https://areding.github.io/6420-pymc/intro.html*
4. *PyMC Documentation and Distribution Support: https://docs.pymc.io/en/v3/api/distributions/continuous.html*
5. *Theoretical Knowledge: ISyE 6420 Lectures and Notes, TA and Staff Support on Ed-Discussion.*
6. *Statistical Theory Reference: A Modern Introduction to Probability and Statistics By F.M.Dekking, C.Kraaikamp, H.P.Lopuhaa, and L.E.Meester.*
7. *Optimization Algorithm: ISyE 6669 Practical Coursework and Teaching by TAs.*
8. *Software Experiment Support: PyMC, Python 3, Matplotlib, Aesara, Scipy, Numpy, Pandas, CSV, Arviz*