Determination of Polynomial Degree in the Regression of Drug Combinations

Boqian Wang, Xianting Ding, and Fei-Yue Wang, Fellow, IEEE

Abstract—Studies on drug combinations are becoming more and more popular in the past few decades, with the development of computer and algorithms. One of the most common methods in optimizing drug combinations is regression of a polynomial model based on certain number of experimental observations. In this paper, we study how to determine the degree of polynomials in different circumstances of drug combination optimization. Using cross-validation, we have found that in most cases, a high degree results in failures of accurate prediction, named overfitting. An anti-noise test has also revealed that polynomial model with high degree tends to be less resistant to random errors in the observations.

Index Terms—Cross-validation, drug combination, polynomial regression, polynomial degree, overfitting.

I. INTRODUCTION

Two significantly important bottle necks in developing drug combinations are: 1) lack of knowledge on the complex interactions between multiple drugs [11]–[14]; 2) a huge searching space generated with multiple drugs at multiple dose levels (10 drugs at 3 dose levels would easily end up 310 combinations) [5]–[8]. In face of these difficulties, platform techniques that allow for rapid identification of drug-drug interactions and effective drug combinations are challenging. Optimization of the drug combination can be described as a function extreme problem: the drug dosage can be considered as the independent variable \( x_i \) of a function, while the drug efficacy can be treated as the dependent variable \( y \). In reality, the responses function between \( y \) and \( x_i \) can be an extremely complex form, denoted as:

\[
  y = f(x_1, x_2, \ldots, x_i).
\]

However, with polynomial regression, we could decode the function with a series of polynomials, based on the calculation of a certain quantity of results of the experiments [9], [10]. The theoretical basis of this method is Taylor expansion. Taylor had proven that if a function is smooth enough at point \( a \) (\( k \) can be any integer greater than 0), the function around point \( a \) can be expressed as a series of polynomials called Taylor series, or a Taylor polynomial, for instance:

\[
  f(x_1, x_2, \ldots, x_i) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_n=0}^{\infty} \frac{(x_1-a_1)^{k_1} \cdots (x_n-a_n)^{k_n}}{k_1! \cdots k_n!} \frac{\partial^{k_1+\cdots+k_n}f}{\partial x_1^{k_1} \cdots \partial x_n^{k_n}}(a_1, \ldots, a_n).
\]

These polynomials can be simplified by opening the brackets:

\[
  f(x_1, x_2, \ldots, x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_1^2 + \beta_{n+2} x_2^2 + \cdots + \beta_{2n+1} x_1 x_2 + \beta_{2n+2} x_1 x_3 + \cdots.
\]

If \( x_i \) is made between [0, 1], the higher the degree of the polynomial is, the smaller its value will be, so polynomials with higher degrees can be abandoned to fulfill the regression process [11]–[13].

Since all the species that are still thriving in this era have gone through millions of years' evolution, the living objects shall adapt well to the dynamic change in their living surroundings within a reasonable range. These facts indicate that when changes occur for the factors such as temperature, humidity, oxygen level and drug stimulation in a narrow range, biological systems such as cells, animals and human beings rarely see a cliffy or prickles function, otherwise the species would not be robust enough to survive. In another word, biological systems responded to drug combination stimuli shall satisfy the prerequisite of the Taylor expansions mentioned above.

II. DETERMINATION OF THE POLYNOMIAL DEGREE

Usually, in drug combination optimization, it is believed that a high degree will enable the polynomial model to preserve more details when simulating the real dose-effect function, making the regression more accurate. A simple example (Fig. 1), we simulate a single drugs dose-effect function as:

\[
  y = \frac{1}{1 + \left( \frac{x}{0.5} \right)^3}
\]

according to Hill curves (Hill coefficient = 3, \( D_0 = 0.5 \)) [3], [4], [14]. With 5 well distributed data points, we regress them at 1–4 degrees \( D \) separately. With the increase of the degree
of the model, the correlation coefficient $R$ gets closer to 1 and the function of the model seems to be more similar to the single drug dose-effect function.

![Dose-effect function](image)

Fig. 1. Regression of a theoretical single dose-effect curve.

However, this actually may be an $R$-value trap: in most cases, higher degree polynomial regression is neither necessary nor reliable. The main reason is a phenomenon called overfitting [15]. When overfitting happens, a statistical model starts to capture random errors or noises instead of the underlying relationship between dependent variables and independent variables. When the model is excessively complex with too many parameters relative to the number of observations, a violation to Occam’s razor occurs. The most common way to check overfitting is cross-validation. For the example in Fig. 1, the simplest type of cross-validation is to leave one observation out at a time and repeat the regression to see which model has the smallest sum of squared error ($SSE$) [16].

$$SSE = \sum \frac{(Y' - Y)^2}{n}. \quad (5)$$

$SSE$ is greatly dependent on the observations, which are the sample points in this case. By calculating the data in Fig. 1, the result is 0.026, 0.075, 0.055 and 0.129 separately for degrees 1–4 regression, indicating that a $D = 1$ polynomial regression is actually most efficient and reliable. In the context of drug combination optimization, which degree is the best model depends greatly on the actual situation. The main factors to be considered are the number of drugs and the number of tested combinations. A common method to avoid overfitting is to ensure the number of observations much larger than the number of coefficients to be regressed. In Table I, we have enumerated the number of coefficients in different cases. The number rises dramatically when the polynomial degree increases.

### III. Real Cases of Optimizing Drug Combinations

#### A. Optimizing TCM Extracts

To demonstrate the influence of polynomial degree, here we discuss a real case of optimizing the combination of 8 compounds extracted from traditional Chinese medicine (TCM) formula Sijunzi Soup, which consists of ginseng, atractylode, glycyrrhiza and poria. The combinations of these extracts are designed for optimal effect of killing lung cancer cells and meanwhile protecting the normal lung tissue cells from the damage of chemotherapeutic drugs so that they could be ideal adjuvant drugs in the treatment of lung cancers.

<table>
<thead>
<tr>
<th>Drug number</th>
<th>$D = 1$</th>
<th>$D = 2$</th>
<th>$D = 3$</th>
<th>$D = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>10</td>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>15</td>
<td>35</td>
<td>64</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>21</td>
<td>56</td>
<td>106</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>28</td>
<td>84</td>
<td>165</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>36</td>
<td>120</td>
<td>246</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>45</td>
<td>165</td>
<td>355</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>55</td>
<td>220</td>
<td>499</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>66</td>
<td>286</td>
<td>686</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>78</td>
<td>364</td>
<td>925</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>91</td>
<td>455</td>
<td>1226</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>105</td>
<td>560</td>
<td>1600</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>120</td>
<td>680</td>
<td>2059</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>136</td>
<td>816</td>
<td>2616</td>
</tr>
</tbody>
</table>

For these purpose, the drug combinations are tested on two cell lines, lung cancer cell A549 and lung normal tissue cell MRC-5, and each drug has 7 doses including 0. For each cell line, 120 combinations are tested for the regression. The regression results (Figs. 2 and 3) show that the $R$ values get remarkably greater with the increase of polynomial degree on both cell lines, reaching nearly 1.0 at degree of three ($D = 3$). The histograms on the right show the distribution of the residues of each set of data, which is the deviation between predicted value and experimental value. With the rise of degree, the residues decrease and the distribution also get better. Through cross-validation, $D = 2$ has the smallest $SSE$.

With the $D = 2$ model, we then computed several potential drug combinations that could both suppress A549 cell proliferation and attain MRC-5 cell viability. Verification experiments were then performed accordingly. The results showed that about 50% of the combinations showed experimental readouts close to predicted value, while others have relatively greater deviation. We could attribute such deviation to the errors in the experimental data points used for regression, which lowers the fitness of the model. Since the effects of TCM extracts are relatively mild, errors become inevitable and non-ignorable.

#### B. Optimizing Chemotherapeutic Drugs

To further test the $SSE$ model, we applied another set of data that studied the effect of eight chemotherapeutic drugs on three lung cancer cell lines, A546, H226 and H460. The purpose of this experiment is to find optimal combinations of frequently used chemotherapeutic drugs that can effectively kill multiple types of lung cancer cells at the same time, which makes them the potential clinical formulas for cancer patients carrying more than one type of tumor cells. For each cell line,
Fig. 2. Regression results compared with experimental results of
A549 cells treated by TCM drugs at different degrees.

Each drug was assigned 3 doses, and totally 59 combinations
were tested. The regression is done at 1−3 degrees. Higher
degrees have apparently better fitness (R value). However, for
A549 and H460, SSE values are smallest at D = 2 (Figs. 4–6).

Compared with the combination of TCM drugs discussed
before, chemotherapeutic drugs often deliver more intense
effects. Thus the fitness in Figs. 2 and 3 is overall better than
that in Figs. 4–6.

IV. ANTI-NOISE PROPERTIES OF
THE POLYNOMIAL MODELS

As discussed in II, in overfitting, the model captures random
errors or noises instead of the function relationship. Therefore,
a noise resistance test is also carried out on each model, in
addition to the cross-validation. Normally distributed random
errors are added to both the inputs (dose) and the outputs
(effect) of the experiments, by 5%, 10%, 20% and 50%
separately. The regression is repeated by 100 times with these
random errors and SSE is calculated by (5), similar to the

cross-validation (Y is the experimental value and Y′ is the
predicted value). The results show that the SSE of higher
degree polynomial regression raises much faster with the
increase of the random error magnitude (Figs. 7 and 8). With
errors no more than 10%, D = 2 has the minimum SSE,
but when it comes to errors greater than about 15%, D = 1
becomes the most reliable model. Usually, systematic errors
within 10%−15% are acceptable in cell experiments or some
other kinds of biological experiments.

Moreover, it can be observed that Figs. 7 and 8 have two
slightly different patterns: with 5% errors, the SSE of
D = 3 in Fig. 7 is much larger than in Fig. 8. This is due to the
reason mentioned before that the effects of TCM extracts
are relatively mild, causing more errors are recorded in the
experimental data.

In addition to SSE, we also calculated correlation coef-
ficients R between model predicted value and input value
with and without manually added random errors separately
(Figs. 9–12). In Figs. 9–12, the R value varies inversely with
the magnitude of random errors, except that in Figs. 9 and 10,
Fig. 4. Regression results compared with experimental results of A549 cells treated by chemotherapeutic drugs at different degrees.

Fig. 5. Regression results compared with experimental results of H226 cells treated by chemotherapeutic drugs at different degrees.


de D = 3 curves stay still as a line of R = 1, even if the random errors reach 50% of the real value. On the contrary, when we compare model predicted value and input value without errors in Figs. 11 and 12, R value of D = 3 drops most significantly, even close to 0 at 50% errors, suggesting that the predicted value and the input value are almost irrelevant to each other. The contrast between Figs. 9–10 and Figs. 11–12 verifies the R-value trap mentioned before, as in real cases of experiments, the real values without errors are definitely not known to us.

Due to the fact that errors are inevitable in biological experiments, the D = 2 polynomial model is most recommended as it keeps the balance between the anti-noise ability and the accuracy of prediction.

V. METHODS

A. Cell Experiments

In all the cell experiments, the viability of cells is tested by CCK-8 kit, a dye whose absorption at 450nm is proportional to the cell viability.

B. Regression

All the polynomial regression process in this research is completed by MATLAB, using function nlinfit. This function uses the Levenberg-Marquardt nonlinear least squares algorithm and an iterative reweighted least squares algorithm separately for non-robust and robust estimation [17]–[19].

VI. DISCUSSION

For all the examples, R reaches 1 at D = 3. This is because at D = 3, there are totally 165 coefficients to be regressed, but the number of data points is less than it, which means as long as there is no contradictory data points, there must exists a set of coefficients that fit all the data points, making R equals to 1. However, by calculating the SSE value, we now know that R = 1 is not always reliable. In this research, we choose the number of polynomials by enumerating all the polynomials at a certain degree. For example, 8 drugs result in 45 coefficients for D = 2 and 165 coefficients for D = 3. However, a desired model may only require part of the significant coefficients, for instance 60 or 120 coefficients. To achieve 60 or 120 coefficients, we need to abandon some of the polynomials.
Fig. 6. Regression results compared with experimental results of H460 cells treated by chemotherapeutic drugs at different degrees.

We have basically 3 types of cubic polynomials, $x^3$, $x^2x_1$, and $x_1x_2x_3$, and the values of the polynomials should be at the same magnitude due to the same degree. To scientifically drop those insignificant coefficients shall benefit the problem of overfitting.

VII. CONCLUSION

Based on these instances, the determination of the polynomial degree in drug combination optimization can be concluded that a quadratic ($D = 2$) model is most appropriate regressing limited number of observations, especially for combinations with more than 6 drugs. Although the complexity of the quadratic model does not well match the real dose-effect function of drug combinations, a higher degree greatly increases the number of coefficients in the model, thus often causes the problem of overfitting and less noise-resistance. To confidently determine the polynomial degree, a cross-validation test is recommended.
Fig. 9. The correlation coefficient $R$ between model predicted value and regression input value in TCM drugs data.

Fig. 10. The correlation coefficient $R$ between model predicted value and regression input value in chemotherapeutic drugs data.

Fig. 11. The correlation coefficient $R$ between model predicted value and real value without errors in TCM drugs data.

Fig. 12. The correlation coefficient $R$ between model predicted value and real value without errors in chemotherapeutic drugs data.
REFERENCES


Boqian Wang graduated from Zhiyuan College, Shanghai Jiao Tong University (SJTU), China, in 2016, with the B.S degree. He is currently a master student at the School of Biomedical Engineering, SJTU, China. His research interests include drug combination optimization and CyTOF data analysis.

Fei-Yue Wang received his Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, New York in 1990. He joined the University of Arizona in 1990 and became a Professor and Director of the Robotics and Automation Lab (RAL) and Program in Advanced Research for Complex Systems (PARCS). In 1999, he founded the Intelligent Control and Systems Engineering Center at the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Overseas Chinese Talents Program from the State Planning Council and “100 Talent Program” from CAS, and in 2002, was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory of Management and Control for Complex Systems. Dr. Wang’s current research focuses on methods and applications for parallel systems, social computing, and knowledge automation. He was the Founding Editor-in-Chief of the International Journal of Intelligent Control and Systems (1995-2000), Founding EiC of IEEE ITS Magazine (2006-2007), EiC of IEEE Intelligent Systems (2009-2012), and EiC of IEEE Transactions on ITS (2009-2016). Currently he is EiC of China’s Journal of Command and Control. Since 1997, he has served as General or Program Chair of more than 20 IEEE, INFORMS, ACM, ASME conferences. He was the President of IEEE ITS Society (2005-2007), Chinese Association for Science and Technology (CAST, USA) in 2005, the American Zhu Kezhen Education Foundation (2007-2008), and the Vice President of the ACM China Council (2010-2011). Since 2008, he is the Vice President and Secretary General of Chinese Association of Automation. Dr. Wang is elected Fellow of IEEE, INCOSE, IFAC, ASME, and AAAS. In 2007, he received the 2nd Class National Prize in Natural Sciences of China and awarded the Founding Scientist by ACM for his work in intelligent control and social computing. He received IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, and IEEE SMC Norbert Wiener Award in 2014.

Xianting Ding is associate professor at the School of Biomedical Engineering, Institute for Personalized Medicine, Shanghai Jiao Tong University. He received his Ph.D. degree from the Department of Mechanical Engineering at University of California, Los Angeles (UCLA) in 2012. His research interests focus on developing personalized therapy and precision medicine, including: 1) developing bio-sensors for early detection of cancer, infectious disease, metabolic diseases, age-related diseases, cardiovascular diseases, based on cellular electrochemical, impedance, mechanical and photonic signaling; 2) developing personalized treatment; optimizing drug combinations; studying drug-drug interactions; building up models for bio-complex systems, based on Feedback System Control (FSC), Microfluidics and Bio-MEMS; 3) traditional Chinese medicine (TCM) modernization, extraction, purification, and re-combination for treating ischemia, osteoporosis and depression.

He participated in the early investigation and development for 3 international research centers, including Institute for Cell Mimetic Space Exploration (CMISE, funded by NASA), Center for Cell Control (CCC, funded by NIH) and Institute for Personalized Medicine (IPM, funded by Chinese Central Organization Department). He is now leading 16 combinatorial-drug-optimization related international projects. He is the editorial board member for Journal of the Association for Laboratory Automation (JALA), the reviewer for 7 journals, published 26 peer reviewed journal papers and filed 11 patents. Corresponding author of this paper.