Multivariate Analysis of NMR Data to Assess Embryo Implantation

Rebecca Hager

Department of Mathematics and Biostatistics, State University of New York at Buffalo

Introduction

During *in vitro* fertilization (IVF), an embryo grows in and interacts with a media before the embryo is transferred to the woman. Scientists think changes in the concentrations of compounds (e.g., amino acids) in this media can be analyzed to predict whether an embryo will successfully implant in a woman. IVF is an expensive procedure, and a failure could cause a woman both financial and emotional distress. We are trying to increase the success rate of IVF by identifying a good embryo to implant by analyzing the media around it before transfer. This non-invasive procedure could save money and reduce multiple births by eliminating the need to transfer more than one embryo. Multiple births, in addition to not always being desirable, can cause both mother and children additional medical problems. Decreasing the incidence of multiple births by implanting fewer embryos that are more likely to be successful is ideal.

We analyzed the media using Nuclear Magnetic Resonance (NMR) spectroscopy which produces thousands of variables. NMR yields a spectrum of the concentrations of compounds in the media. We then used various statistical methods to deal with these many variables in order to compare the spectra in successful and unsuccessful embryos to see if there is a difference that can be used to predict implantation status for new embryos.

We have 80 media samples, 61 of which are associated with an unsuccessful implantation while 19 are associated with a success. Each of these samples has a control media for comparison. We subtracted the spectrum of the control media from the spectrum of each media to determine the differences arising during the embryo's growth. We use 1 (and red) to represent an unsuccessful implantation and 2 (and blue) to represent a successful implantation.

Cross Validation

Cross Validation (CV) is a common method used to make sure that we do not over-fit our model to our sample data and report an overoptimistic accuracy of the model. Over-fitting our sample data will not produce accurate results when making predictions on new observations. We use double K-fold cross validation with K=10. The data is split into 10 sets and we model 9 of the sections using some technique. Then we use the model we created from the 9 sections, called the training set, and apply it to the one section we left out, or the test set, to see how it performs. We repeat this 9 more times so we leave each section out once. Then we choose the best model based on how it performed when applied to the training set.

We use cross validation repeatedly on the data to evaluate the stability and accuracy of the model. We normally report a measure of model accuracy such as mean square error (MSE). Double cross validation first does cross validation to choose the optimal parameter or model and then does an outer loop of cross validation to see how that model performs.

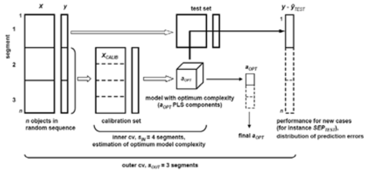

x
y
segment
1
2
3
n
test set
$y - \hat{y}_{TEST}$
X_{CALIB}
a_{OPT}
model with optimum complexity
(a_{OPT} PLS components)
n objects in
random sequence
calibration set
inner cv, s_{IN} = 4 segments,
estimation of optimum model complexity
final a_{OPT}
performance for new cases
(for instance SEP_{TEST}),
distribution of prediction errors
outer cv, s_{OUT} = 3 segments

Figure 1: Demonstrates double cross validation.

References

Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics.* Taylor & Francis - CRC Press, Boca Raton, FL, 2009.

(Picture above from page 117 of Varmuza book)

K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an easily understood classification method. KNN predicts the class membership (successful or unsuccessful implantation in our case) by looking at the nearest k neighbors in the variable space using a chosen distance measurement.

In our research, we use the common Euclidean distance. For a new observation, we can use our sample data and look at the k-nearest neighbors' implantation statuses. If the majority of the neighbors were successful in implantation, then we would predict that our new observation will also be successful and vice versa. The optimal number of neighbors to use, k, is chosen using cross validation. We found that it is optimal to look at the 3 nearest neighbors for our *in vitro* data. For illustration purposes, we plotted the data in the space spanned by the first 2 principal components. It does not appear that this method is accurately distinguishing the two groups.

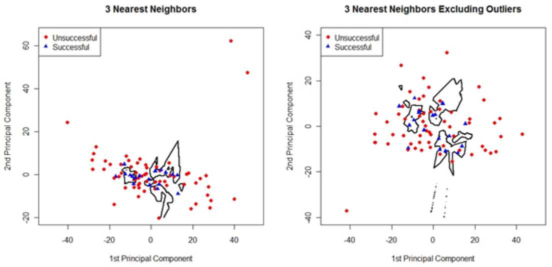


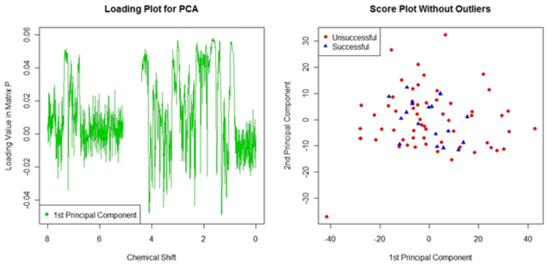
Figure 2 (Left): Results from looking at the 3 nearest neighbors using all of the data. New observations inside the black lines would be classified as successful while outside would be classified as unsuccessful.
Figure 3 (Right): Results from looking at 3 nearest neighbors excluding 2 outliers.

Principal Component Analysis

Principal Component Analysis (PCA) is designed to get the most information out of the x-data in fewer variables. This does not take the y-variable into account. PCA creates principal component vectors which are linear combinations of the original variables, orthogonal to one another, and uncorrelated. These principal components are chosen to maximize their variance. PCA deconstructs the **X** matrix in the following way where **P** is called the loading matrix and **T** is the score matrix.

$$\mathbf{T=XP}$$

P defines new dimensions in **X**-space and **T** has the coordinates of the data within the new coordinate space. We then graph the principal components from matrix **T** to see a separation of the data by group. We can also do regression on the principal components to predict the outcome for new observations. As we can see from the score plot below, there doesn't seem to be a separation between the unsuccessful and successful implantation groups in the first two principal components.



Figures 4 (Left): The loading plot shows what variables are responsible for the most variation in **X**.
Figure 5 (Right): Shows the score plot in the first 2 principal components. No separation between the groups is evident.

Partial Least Squares Regression

Partial Least Squares Regression (PLS) is a powerful statistical method that is used to model a response variable, y, from a large number of highly correlated predictor variables, x_1 through $x_{m,}$ or in matrix form

$$\mathbf{y=Xb+e}$$

Our **y** is a vector of 1's and 2's with 1 indicating an unsuccessful implantation and 2 indicating a successful implantation. **X** is a matrix of the NMR spectral data from the media for all the samples. Since there are thousands of x-variables, a regular regression model may be unstable and any signal may be swamped by the random noise in the system. We use PLS to distill the most important x-information into a few variables, t_1 through t_a, where a is usually not more than 5, modeling

$$\mathbf{X=TP'+E}$$

The optimal number of components used is determined by predictive ability, which must be estimated using cross validation to avoid over fitting. The variables in **T** are chosen to contain most of the variability (information) in **X** that is associated with the response, **y**. We can now successfully model **y** based on many fewer variables where **d** is the matrix of regression coefficients and **h** is the residual matrix.

$$\mathbf{y=Td+h}$$

Using cross validation, we chose the optimal number of components to be 3. PLS regression does not seem to give accurate predictions for the successful and unsuccessful implantation groups as can be seen below in the boxplot.

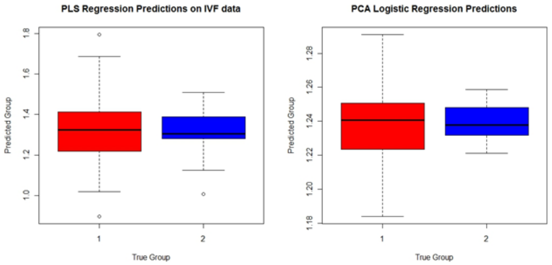


Figure 6 (Left): Shows the predicted values that PLS regression gives for the unsuccessful and successful implantations. There does not seem to be a difference between the two groups.
Figure 7 (Right): Shows predicted values from PCA Logistic Regression for the two groups.

Conclusions

Using K-Nearest Neighbor, PCA, PLS Regression and various other statistical techniques, we have not found a reliable association between this spectral data and implantation status. It could be that the NMR spectroscopy is not sensitive enough to measure the real association. Additionally, our data set had several external variables that were not constant throughout the data collection, such as type of media used. We also have a relatively small sample size of 80 observations coming from 39 subjects. More controlled studies with more subjects should be conducted to support or refute our findings.

		Predicted					
		KNN		PCA		PLS	
		1	2	1	2	1	2
Truth	1	55	6	15	46	23	38
	2	12	7	0	19	4	15
Specificity		0.902		0.246		0.377	
Sensitivity		0.368		1.000		0.789	

Table 1: Summary of performance of KNN, PCA, and PLS. Specificity is probability of correctly predicting an unsuccessful implantation and sensitivity is probability of correctly predicting a successful implantation.

Acknowledgements

I would like to thank my mentor, Dr. Andrews as well as CURCA, Dr. Hemmer, Dr. Szyperski, Jia Hua, Dr. Sukumaran, and Dr. Shen.