Special Plots methods with diabetes disease data

Yida Bao, Zheran Wang, Jingping Guo, Philippe Gaillard, Auburn University

Abstract:

A graph has always been more intuitive than icy statistics. SAS gives us quite powerful graph capabilities. In this project, we use diabetes datasets as an example to do data visualization research. The detection of diabetes can generally be judged by several indicators- Glucose, Insulin, BMI , and so on. In general, our research contains three parts. First, we apply **SAS ®** procedure **PROC Gplot** to create a line diagram, which helps us to explore the inner structure between different factors. We will introduce two different methods to generate an overlay plot based on the binary detection results. Also, we will apply **SAS Enterprise Miner** to proceed with the principal component analysis, which helps us to bright the project. Later, we'll use the typical discrimination method, convert the dataset into several canonical variables, and generate the proper plot to express the result. At last, we'll summarize the result to establish information hierarchy, and tell the other researcher our understanding of the diabetes dataset.

Key words: diabetes ●canonical● sgplot ● gplot ● candic

Introduction:

Data visualization is considered to be an identical modern concept as visual communication by many subjects. In the general sense, communicating with data is at the intersection of science and art. It contains inherent diversity. Different people will come up with various ways to solve the same data visualization challenge. Also, software visualization methods involve multiple creative ideas. There is no single "correct" answer. On the contrary, there are many potential paths for effective data communication.

To communicate information clearly and effectively, data visualization uses statistical graphs, charts, infographics, and other tools. A dozen years ago, the challenge of data visualization was very difficult, and many researchers even need to draw the graph by hand. Nowadays, it is very simple to generate tables and graphs, but more information does not make communication easier but makes it more difficult for people to filter out the most important parts. The ideal situation of a data analyst is to draw a profound story with simple graphics.

The choice of tools becomes important. Most of the time, tools don't seem to need to be confined to certain software. In this paper, we use SAS 9.4 and SAS Enterprise Miner to show the data visualization process. SAS and SAS Enterprise Miner is an advanced analytics data mining tool intended to help users quickly develop descriptive and predictive models through a streamlined data mining process. We will show the charm of the GPLOT statement in different ways and will use a linear transformation of statistical methods to integrate high-dimensional data sets.

Data description:

In recent years, data visualization has been fully utilized in the field of medicine and pharmacology. Choosing the proper dataset has become the first challenge of data visualization.

In this paper, we choose pima indians diabetes database, a database that has been used as a benchmark in several studies. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. So, there is no variable about gender. There exist over 769 patients test result in the dataset, including 269 person label as diabetes and 500 person as a negative result. Based on different types of testing results, we got 9 variables.

Table 1 variable description

variable		description
Pregnancies	Categorical variable	Times for pregnancies
Glucose	Continuous variable	a simple sugar value
BloodPressure	Continuous variable	pressure of circulating blood on the walls of blood vessels
SkinThickness	Continuous variable	skinthickness
Insulin	Continuous variable	a peptide hormone produced by beta cells of the pancreatic islets;
BMI	Continuous variable	Body Mass Index
DiabetesPedigreeFunction	Continuous variable	a function which scores likelihood of diabetes based on family history
Age	Continuous variable	age
outcome	Binary variable	test result

(Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.)

Since this article is about data visualization, we will not use any machine learning statistical method to predict the probability of a patient getting sick or to predict different methods of misclassification. However, this does not prevent us from using a plot or graph to achieve the same purpose.

Proc gplot:

There are various types of charts for data visualization, but in fact, mastering a few of them can meet most needs. From my point of view, a well-designed plot often conveys information faster than well-designed tables. Graphs are an important method for displaying data, and the intuitive effect of graphs cannot be replaced by data tabulation. In this paper, we will focus on how to use **Proc Gplot** to create the overlay plot.

Before the GPlot statement, through the **AXIS statement**, we set the scale range of the coordinate axis, the color description label, the number of this scale in the middle of every two main scales and other attributes. There are many options in the AXIS statement, and the usage is also very rich. Here is a brief list of some commonly used options and usage

Table.2 AXIS statement

option	Role	Usage
Order=	Specify the value displayed on the main scale, which can be numeric and text.	Axis order=(1 to 6); Axis order=('China',USA'','Canada')
Major=	Specify the display properties of the main scale, such as color, height, thickness, number, etc.	Axis3 Major=(color=blue height=3)
Minor=	Specify the display attributes of the minor scale	Axis4 minor = (color=blue height=3 number=2)
Length=	specify the length of the coordinate axis	Axis5 Length= 7IN
Label=	specify the attributes of the coordinate axis label, such as color, height, font, etc	Axis6 LABEL= (color = red Height=5 'year')
Value=	Specify the display text and attributes of the main scale on the coordinate axis	Axis7 value= (color = red t=1 '74 in' t=2 color=green t=3 '5 year')

Besides, **Symbol statement** use to set up scatter symbol, color and other attributes

Table.3 Symbol statement

option	Role	Usage
Value= or V=	Specify the scatter symbol	x, STAR, SQUARE,DOT,etc
V=	specify the color of the scatter symbol	Red or blue , etc
Height= or H=	specify the size of the scatter symbol	N bigger than 0,the bigger value the bigger size of scatter symbol

At the same time, the SYMBOL statement also provides more options to control the line style, thickness, color and other attributes In this paper, we will apply **INTERROL =** to create connection plot. **INTERROL =None** means no connection diagram, it's the system default value. **INTERROL = Join**, which means that the data points are linked with straight lines in the order in which they appear in the data set, and **INTERROL =Spine** means that the data points

are connected in the order in which they appear in the data set using smooth interpolation curves, and the curve passes through each data point

The **SYMBOL statement** also provides more options to control the line style, thickness, color and other attributes

Table.4 AXIS line chart statement

option	Role	Usage
WIDTH= or W=	Specify the thickness of the line	The larger the value, the thicker the connection.
LINE= or L=	specify the line type of the plot	Each number represents a certain line type
CI=	specify the color of the line	Red, green, blue,etc
Color = or C=	specify the color of the line and scatter	Red, green, blue, etc

After we introduce the basic usage of **SYMBOL** and **AXIS Statement**, we'll apply these two statements into our overlay research. For comparison, it is very useful to draw multiple lines in the same graph. Besides, in time-series research, to compare the trend and size of the predicted and actual values, we also need to compare in the one graph.

Overlay method 1: Use the option **OVERLAY** to show that multiple graphs in the same plot statement are displayed in the same graph.

```
axis1 order = (0 to 17 by 1);
axis2 order = (20 to 120 by 5)
      minor=(color = blue height = 0.25 number = 1);
symbol1 value = dot cv = blue interpol =spline ci= blue;
symbol2 value = # cv= green interpol = spline ci=green line=4;
proc gplot data = diabetes2;
plot  Insulin_Mean*Pregnancies BloodPressure_Mean*Pregnancies/ overlay
haxis = axis1 vaxis = axis2;
run;
```

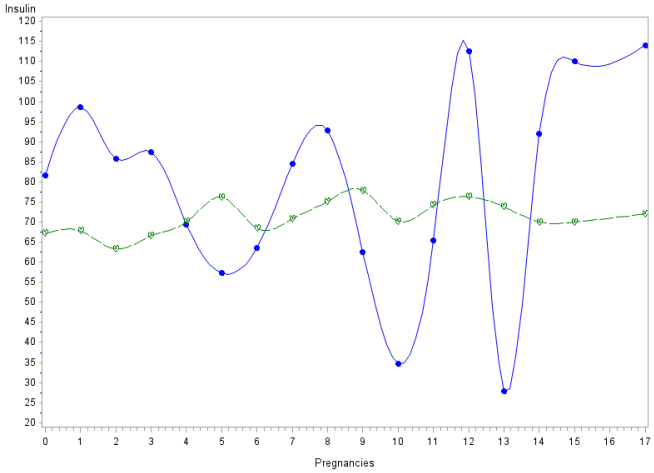


Table 5 method 1 for overlay

The X-axis is the number of pregnancy while the y-axis is the value of Insulin and Bloodpressure. We didn't scale the variable's value, so the graph looks unnatural for those two line plots. Since the two line plots are in the same graph, SAS stipulates that the SYMBOL1 statement sets the first line plot and the SYMBOL2 statement sets the second line plot. Besides, in order to make the graphics look smoother, we used **interpool = spine** instead of **interpool = join** statement

Overlay methods 2: We can use **Plot2 statement**.

```
axis1 order = (0 to 17 by 1);
axis2 order = (60 to 80 by 1)
      minor=(color = blue height = 0.25 number = 1);
axis3 major=(number = 8)
      minor=(number = 1);
symbol1 value = dot cv = blue interpol =spline ci= blue;
symbol2 value = diamond cv= green interpol = spline ci=green line=10;
proc gplot data = diabetes2;
plot  BloodPressure_Mean*Pregnancies/ legend haxis = axis1 vaxis = axis2;
plot2  Insulin_Mean*Pregnancies    / legend vaxis = axis3;
run;
```

From the code above, we need to mention two points,

1. **Plot statement** has to use before **plot2 statement.**
2. If we need to display all the graphics, we need to use the legend statement in both the plot and plot2 statements.

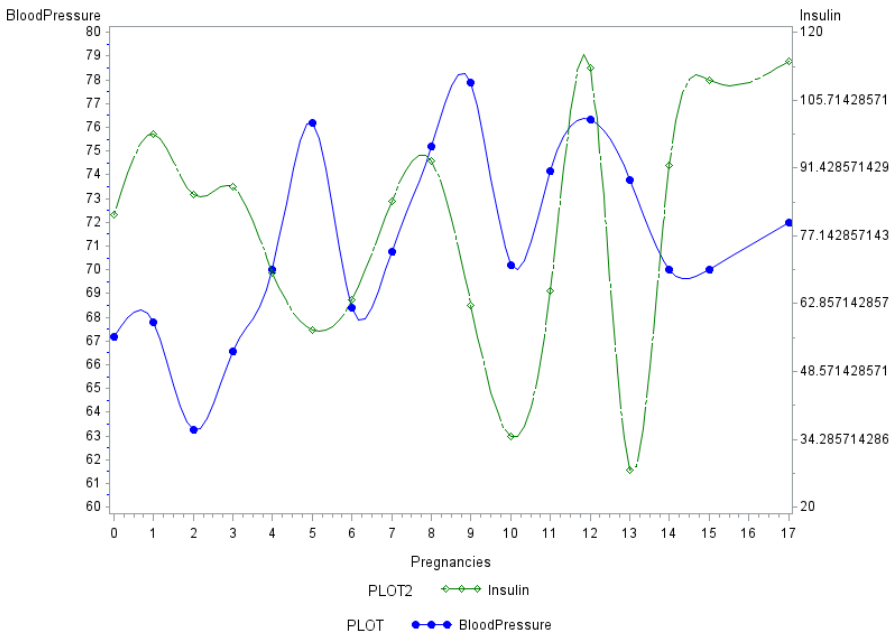


Table 6 method 2 for overlay

Use the plot 2 statement to set up a y- coordinate for the specified ordinate variable on the right side of the figure, so that different ordinate variables can use different ordinate coordinates. From the graph above, we can find BloodPressure represents the left y-axis, and the Insulin represents the right. We use VAXIS = statement to set the right y-axis symbol.

Proc Candisc:

In this section, we use Fisher discrimination to graph the plot. The idea of Fisher discrimination analysis is very simple: given a set of training samples, try to project the samples on a straight line, so that the projection points of similar samples are as close as possible, and the projection points of different samples are as far away as possible. When classifying the new sample, project it onto the same straight line, and then determine its category according to the location of the projection point of the new sample.

We need to make the distance between the projection points of the same category samples as close as possible, that is, minimize $J_1 = w^T S_w w$, and none category samples as far as possible, that is, maximize $J_2 = w^T S_b w$.

Construct the function $\max(J) = \frac{J_2}{J_1} = \frac{w^T S_b w}{w^T S_w w}$, which is the target of fisher linearly discriminates, that is, to find the best projection direction.

In high dimension case, Fisher discriminant analysis, using projection technology to reduce dimensionality, after dimensionality reduction, calculate the intra-group deviation (here can be compared to the random error in ANOVA), and calculate the inter-group deviation (here can be compared to the level of each factor in the ANOVA Deviation between groups), using the convex optimization method to find a straight line or hyperplane that minimizes the deviation within the group and maximizes the deviation between groups to segment different categories.

We can assume a linear discriminant function $U(X) = C^T X$, To find the mean value among K sample size $G_1, G_2, G_3, \ldots . G_k$, then we obtained K values can be used to calculate the spread Which is

$$\frac{\sum_{i=1}^{k}[E_i(C^T X) - \frac{1}{k}\sum_{i=1}^{k} E_i(C^T X)]^2}{\sum_{i=1}^{k} D_i(C^T X)}$$

After determining the linear discriminant function, the distance from the calculation result to each population is determined to be from the population with the smallest distance. When there are too many training sample data sets, they are often not well distinguished, we need to find a second or more discriminant function

In SAS, we can use **PROC CANDIS** statement to proceed with the Fisher discriminate method. **PROC CANDIS** linearly combines the numerical variables of the original data set with

discriminant analysis and obtains several canonical variables, which makes the distinction of the different categories of the original data set on these canonical variables the most obvious.

```
proc CANDISC data = diabetes2 NCAN = 2;
Class  outcome;
Var  Insulin Glucose Pregnancies Bloodpressure skinthickness BMI
DiabetesPedigreeFunction Age ;
run;
```

Table 7 Canonical variables result

Raw Canonical Coefficients			
Variable	**Label**	**Can1**	**Can2**
Pregnancies	Pregnancies	0.0938638298	0.2860866947
Glucose	Glucose	0.0269863520	-.0124274712
BloodPressure	BloodPressure	-.0106293929	-.0114761170
SkinThickness	SkinThickness	0.0007043468	-.0087433210
Insulin	Insulin	-.0008229296	0.0022128609
BMI	BMI	0.0603702056	0.0582040195
DiabetesPedigreeFunction	DiabetesPedigreeFunction	0.6711517147	-.4757337236
Age	Age	0.0119490869	-.0513149739

Then we got the Canonical variable result, the result is a linear combination of the original variables. These linear combinations can explain most of the differences between groups. since we use **NCAN = 2** in the statement, no matter what, we got two canonical variables, Can1, and Can2. Then we can use a scatter plot to represent the result based on a different outcome.

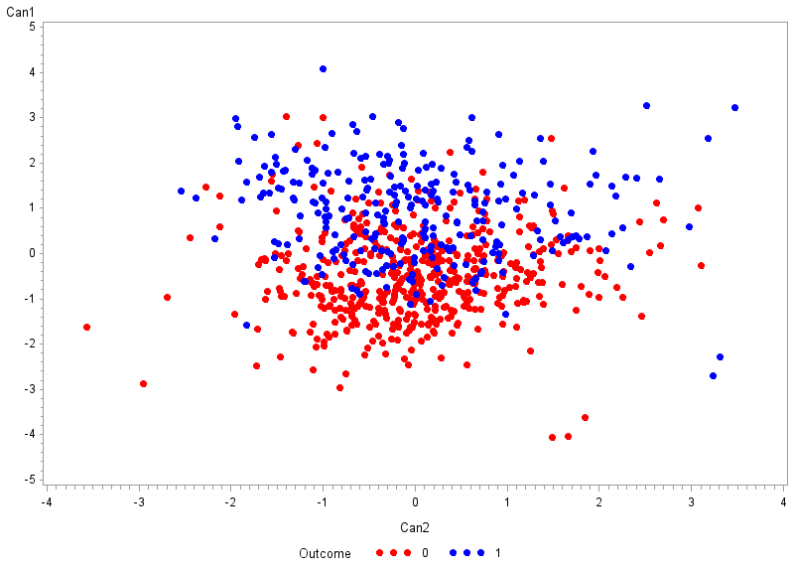


Table 8 Canonical variable scatter plot

From the plot above, we can easily find the red 0 which represents the non-diabetes patient

focus on the below part of the plot, while the blue 1 that represents the diabetes patient on the top of the plot. There may not be a very obvious dividing line, but the difference between different groups is significant.

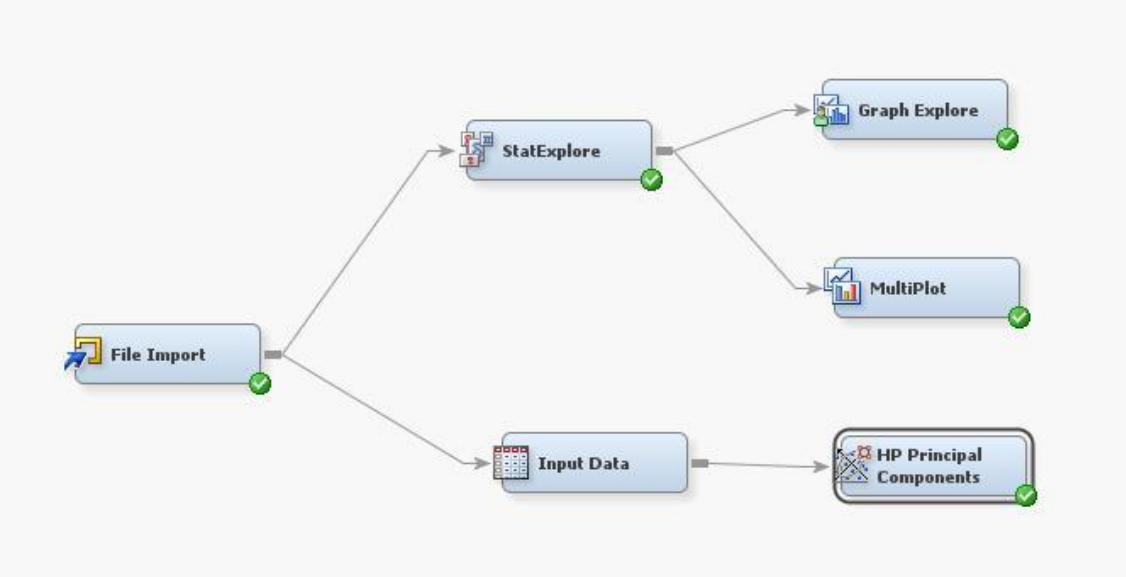
PCA procedure with SAS Enterprise Miner

In this part, we will use the **SAS Enterprise Miner** platform instead of **SAS 9.4** to deal with the diabetes dataset. SAS Enterprise Miner offers many features and functionalities for the business analysis to model the data. We will use **Principal component analysis** to proceed with the research.

PCA, just like the canonical variable method, can use linear combinations to explain most of the differences between groups. PCA is predominantly used as a dimensionality reduction technique in domains like facial recognition, computer vision, and image compression. It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc.

In **SAS Enterprise Miner**, the data mining process is driven by a process flow diagram.

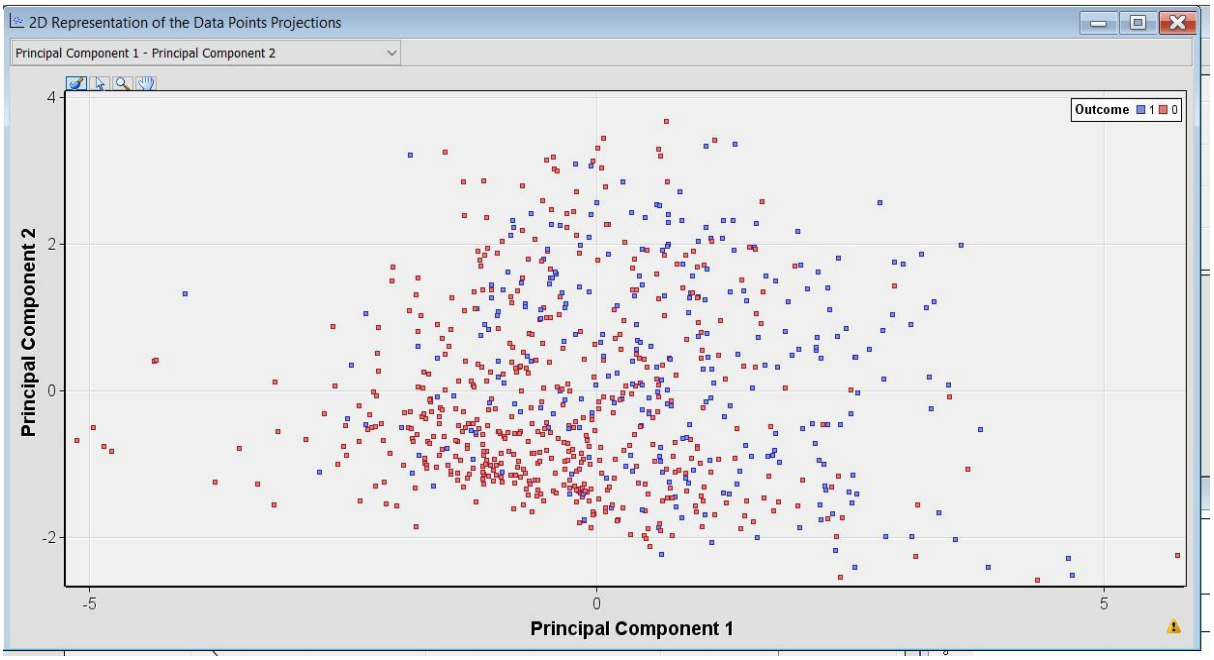
Table 9 Sas Enterprise Miner flow diagram



In SAS Enterprise Miner, we don't need code anymore unless we have the specific requirements. We use the proper node to finish the statistic analysis.

After the Principal component analysis, we got several components. We choose the first and second components to create the Scatter Plot.

Table10 PCA result



From the plot above, principal component1 represent x-axis, and principal component 2 represent the y-axis. We can easily find the red 0 which represents the non-diabetes patient focus on the down left part of the plot, while the blue 1 that represents the diabetes patient on the top of the plot. Also, there may not be a very obvious dividing line, but the difference between different groups is significant.

Conclusion:

In this paper, we use two different methods to create the overlay plot in SAS 9.4, then we use canonical variable and principal component analysis to find the difference between outcome ---binary variable. Since this is an attempt based on data visualization, we do not have strict mathematical or statistical results to apply for our results. But on the other hand, these graphs also clearly show whether the patient has diabetes or not.

It takes time to look at the data from different perspectives and decide how best to present it. Our visual application is far from perfect, and the tools do not understand our needs. To analyze and explain, we must combine those two very carefully and enjoy our graphing style.

Reference:

1. SAS programming in the pharmaceutical industry.

2. Getting Started with SAS ® Enterprise Miner 7.1

3. Bao, Yida, and Philippe Gaillard. "Summarizing some conventional methods to classify a binary target."

Contact Information:

Your comments and questions are valued and encouraged.

Contact the author at:

YIDA BAO, Math PHD Candidate, SAS certified advanced programmer

Department of Mathematics and Statistics, Auburn University

E-mail: yzb0010@auburn.edu

Zheran Rachel Wang, Math PHD student

Department of Mathematics and Statistics, Auburn University

E-mail: zzw0049@auburn.edu

Dr, Jingping Guo, Biology PHD

School of Fisheries and Allied Aquacultures, Auburn University

E-mail: jzg0078@auburn.edu

Dr. Philippe Gaillard, Associate Professor

Department of Mathematics and Statistics, Auburn University

Director of Statistical Consulting Center

E-mail: prg0007@auburn.edu