

STA 437/2005:
Methods for Multivariate Data
Week 1: Introduction and Preliminaries

Piotr Zwiernik

University of Toronto

Table of contents

1. What is this class about?
2. Administrative details
3. Introduction to multivariate statistics
4. Multivariate data visualization
5. Notation and preliminaries (blackboard)

Multivariate datasets

- ▶ Modern datasets are high-dimensional and highly unstructured.
- ▶ This means that we simultaneously observe many correlated features.
- ▶ Modelling them independently is both *wasteful*.
- ▶ The problem is that our intuition works well only in 1D-3D.

This course

- ▶ Discuss some classical and more modern methods for multivariate datasets.
- ▶ Build math toolbox that gives deeper understanding for these methods.
- ▶ Aimed at advanced undergrad and master level graduate students.
- ▶ 2 hrs lecture + 1 hour tutorial to get more hands-on experience.
- ▶ We will use **a lot of** real analysis, probability, and linear algebra.
 - ▶ To large extend I am assuming you are comfortable with matrix algebra.

Course Information

Course Website: <https://pzwiernik.github.io/sta437/>

Main source of information is the course webpage; check regularly!

We will also use **Quercus** for **announcements & grades etc.**

- ▶ You received an announcement on Sunday.

We will use **Piazza** for **discussions.**

- ▶ Sign up via quercus or:

<https://piazza.com/utoronto.ca/winter2025/sta4372005>

- ▶ Your grade **does not depend on your participation on Piazza.**

Course Information

- ▶ This course have two *identical* sections:
 - ▶ Section 1: Wed 9-11am (tutorial Fri 9-10am)
 - ▶ Section 2: Wed 1-3pm (tutorial Fri 1-2pm)
- ▶ You are welcome to attend either one of the sections.
- ▶ Instructor office hours are on Tuesdays 13:30-15:00 (UY 9033).
- ▶ Questions during lectures/tutorials are always welcome!
- ▶ The website contains the tentative schedule for the course and the dates of all the course mark items (the final exam will be confirmed later).

Course Information

- ▶ While cell phones and other electronics are not prohibited, talking, recording or taking pictures in class is prohibited without the consent of your instructor.
- ▶ The lecture notes will cover everything presented in class. Let me know about typos you notice and/or any suggestions you might have.
- ▶ For accessibility services: If you require additional academic accommodations, please contact UofT Accessibility Services as soon as possible, studentlife.utoronto.ca/as. No last minute arrangements will be considered.

Requirements and Marking

- ▶ Two midterms (conceptual/theoretical)
 - ▶ 7 February and March 14 (tentative)
 - ▶ 1 hour tests
 - ▶ Worth 20% of course mark each
- ▶ Final projects
 - ▶ More details will be posted in the reading week.
 - ▶ 1-2 person groups
 - ▶ 20% of course mark
- ▶ Final Exam (conceptual/theoretical)
 - ▶ ~ 2-3 hours
 - ▶ Date and time TBA
 - ▶ Worth 40% of course mark
- ▶ **Everybody must take the final exam! No exceptions.**

Multivariate data

- ▶ Many datasets consist of several variables measured on the same set of subjects: patients, samples, users, or organisms.
- ▶ Goal: investigation of connections or associations between the different variables measured.
- ▶ Usually the data are reported in a tabular data structure with one row for each subject and one column for each variable.

Example

Studying the expression of 25,000 gene (columns) on many samples (rows) of patient-derived cells, we notice that many of the genes act together, either that they are positively correlated or that they are anti-correlated. We would miss a lot of important information if we were to only study each gene separately.

Example 1: Decathlon

The columns are the 10 disciplines in decathlon. The rows are the competitors:

```
> data("olympic", package = "ade4")
> athletes = setNames(olympic$tab,
+   c("m100", "long", "weight", "high", "m400", "m110", "disc", "pole", "javel", "m1500"))
> head(athletes)
```

	m100	long	weight	high	m400	m110	disc	pole	javel	m1500
1	11.25	7.43	15.48	2.27	48.90	15.13	49.28	4.7	61.32	268.95
2	10.87	7.45	14.97	1.97	47.71	14.46	44.36	5.1	61.76	273.02
3	11.18	7.44	14.20	1.97	48.29	14.81	43.66	5.2	64.16	263.20
4	10.62	7.38	15.02	2.03	49.06	14.72	44.80	4.9	64.04	285.11
5	11.02	7.43	12.92	1.97	47.44	14.40	41.20	5.2	57.46	256.64
6	10.83	7.72	13.58	2.12	48.34	14.18	43.06	4.9	52.18	274.07

Question: What are the predominant types of competitors. Which ones are more successful overall?

Example 2: Star Wars

The Star Wars dataset comes from the dplyr package. It describes 13 characteristics of 87 characters from the Star Wars universe.

```
> data("starwars", package = "dplyr")  
> head(starwars[,c(1,2,3,5,6,7,8,11)])
```

	name	height	mass	skin_color	eye_color	birth_year	sex	species
1	Luke Skywalker	172	77	fair	blue	19.0	male	Human
2	C-3P0	167	75	gold	yellow	112.0	none	Droid
3	R2-D2	96	32	white, blue	red	33.0	none	Droid
4	Darth Vader	202	136	white	yellow	41.9	male	Human
5	Leia Organa	150	49	light	brown	19.0	female	Human
6	Owen Lars	178	120	light	blue	52.0	male	Human

Example 3: Pottery

Chemical analysis data on Romano-British pottery made in three different regions (kiln 1, kilns 2-3, and kilns 4-5):

```
> data("pottery", package = "HSAUR2")  
> head(pottery)
```

	Al2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO	kiln
1	18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015	1
2	16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018	1
3	18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014	1
4	16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019	1
5	17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019	1
6	18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017	1

Question: Do the chemical profiles of each pot suggest different types of pots and if any such types are related to kiln or region.

Example 4: US pollution data

Air pollution in cities in the USA.

```
> data("USairpollution", package = "HSAUR2")  
> head(USairpollution)
```

	S02	temp	manu	popul	wind	precip	predays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55.0	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148

The following variables were obtained for 41 US cities:

S02: SO2 content of air in micrograms per cubic metre;

temp: average annual temperature in degrees Fahrenheit;

manu: number of manufacturing enterprises employing 20 or more workers;

popul: population size (1970 census) in thousands;

wind: average annual wind speed in miles per hour;

precip: average annual precipitation in inches;

predays: average number of days with precipitation per year.

Example 5: Academic salaries

The Salaries dataset comes from the carData package. It describes the 9- month academic salaries of 397 US college professors at a single institution in 2008-2009.

```
> data(Salaries, package="carData")
> head(Salaries)
```

	rank	discipline	yrs.since.phd	yrs.service	sex	salary
1	Prof	B	19	18	Male	139750
2	Prof	B	20	16	Male	173200
3	AsstProf	B	4	3	Male	79750
4	Prof	B	45	39	Male	115000
5	Prof	B	40	41	Male	141500
6	AssocProf	B	6	6	Male	97000

Example 6: A bigger example

Gene expression microarray dataset that reports the transcriptomes of 101 individual cells from mouse embryos at different time points in early development. 45101 genes are probed.

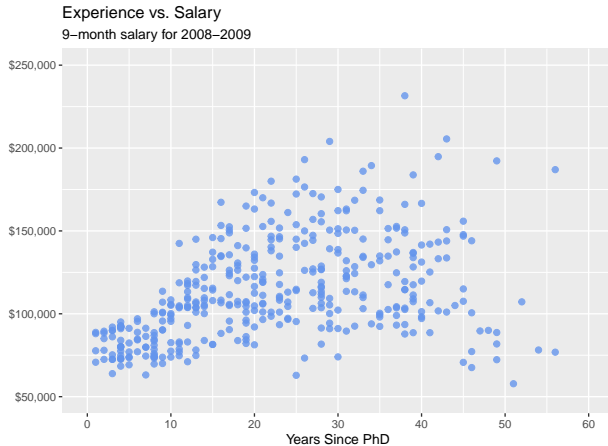
```
> if (!require("BiocManager", quietly = TRUE))
+   install.packages("BiocManager")
> BiocManager::install("Hiiragi2013")
> library("Hiiragi2013")
> data("x")
> dim(Biobase::exprs(x))
[1] 45101  101
> dfx = as.data.frame(Biobase::exprs(x))
```

Main challenges

- ▶ Having 25,000 dimensions of variation to consider at once is daunting; we will show how to reduce our data to a smaller number of most important dimensions without losing too much information.
- ▶ How to visualize multivariate data?
- ▶ Sometimes we do not want dimensionality reduction techniques and need to model a high-dimensional data sets directly. We will learn about high-dimensional statistics.

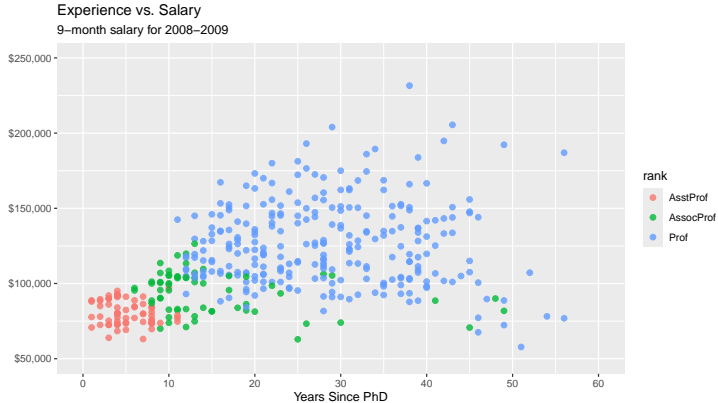
Scatterplot for 2D datasets

The scatterplot is the standard for representing continuous bivariate data.



Scatterplot for 2D datasets + categorical variable

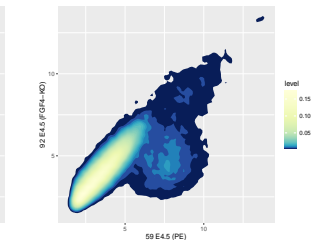
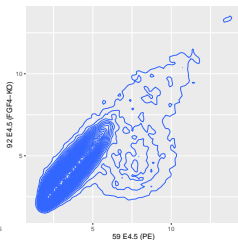
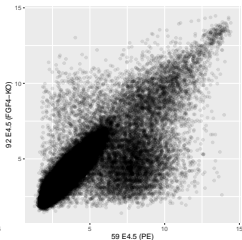
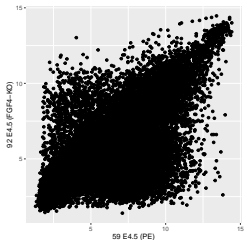
The scatterplot is the standard for representing continuous bivariate data.



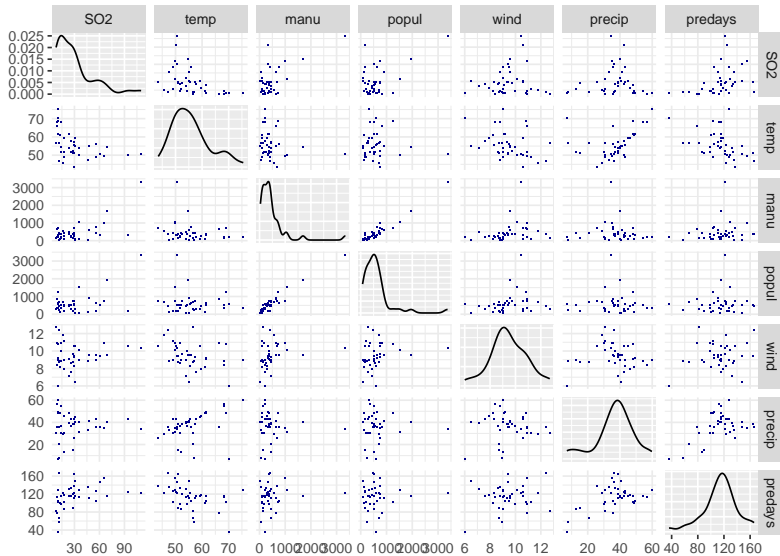
2D kernel density estimation

In our bigger gene expression dataset, take a look at differential expression between a wildtype and an FGF4-KO sample.

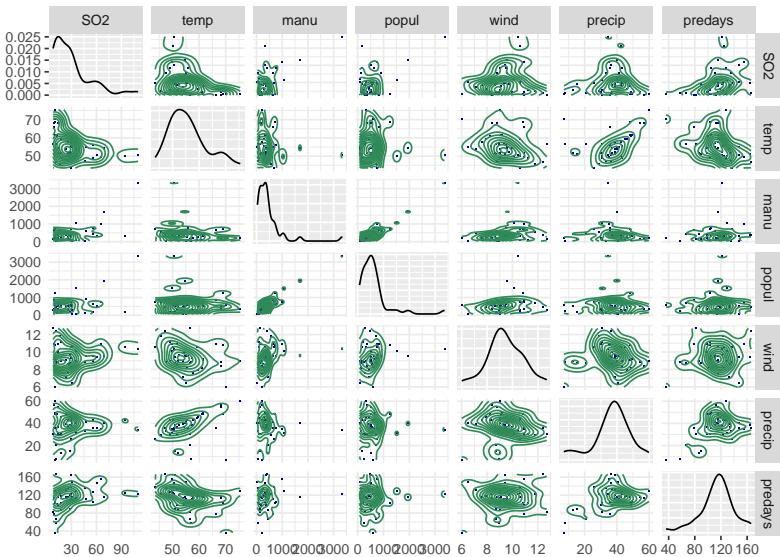
```
> scp <- ggplot(dfx, aes(x = `59 E4.5 (PE)`, y = `92 E4.5 (FGF4-KO)`))  
> plot1 <- scp + geom_point()  
> plot2 <- scp + geom_point(alpha = 0.1)  
> plot3 <- scp + geom_density2d(h = 0.5, bins = 60)  
> library("RColorBrewer")  
> colorscale = scale_fill_gradientn(colors = rev(brewer.pal(9, "YlGnBu")),  
+   values = c(0, exp(seq(-5, 0, length.out = 100))))  
> plot4 <- scp + stat_density2d(h = 0.5, bins = 60, aes( fill = after_stat(level)), geom = "polygon") + colorscale + coord_fixed()
```



The scatterplot matrix and lattice plots



The scatterplot matrix and lattice plots + 2D kernel density estimators



Interactive graphics: Plotly

Plotly is a powerful library for creating interactive, web-based visualizations. It's available in several programming languages including R and Python. With Plotly, users can create interactive plots like scatter plots, bar charts, heatmaps, and 3D plots.

```
> library(ggplot2)
> library(plotly)
> p <- ggplot(mpg, aes(x=displ, y=hwy, color=class)) + geom_point(size=3) +
+ y = "Highway Mileage", color = "Car Class") + theme_bw()
> ggplotly(p)
```

Mousing over a point displays information about that point. Clicking on a legend point, removes that class from the plot. Clicking on it again returns it. Popup tools on the upper right of the plot allow you to zoom in and out of the image, pan, select, reset axes, and download the image as a png file.

Summary of this part

- ▶ Vectors, matrices, random vectors
- ▶ Interpretations for Ax and AB
- ▶ Linearity of expectation, bilinearity of covariance