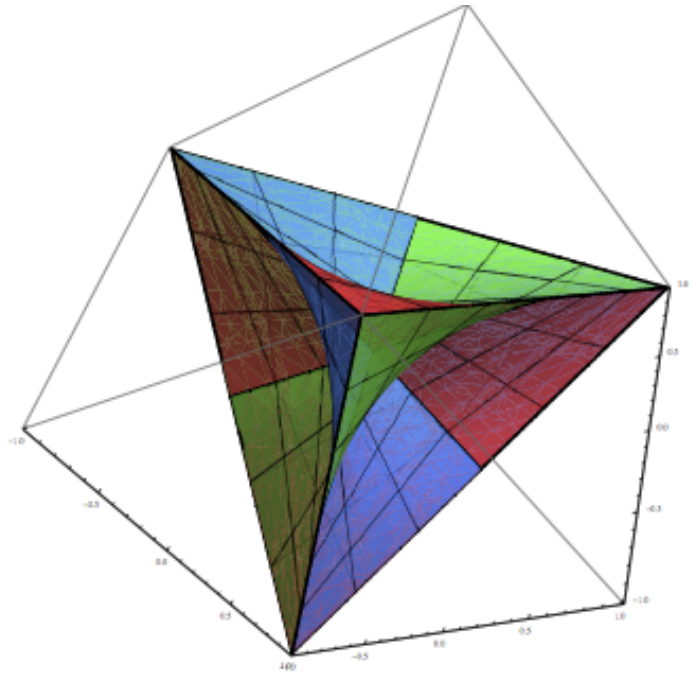


Piotr Zwiernik

Universitat Pompeu Fabra

Nit de la Recerca,
29 September 2018

Simpson's paradox: UC Berkeley admissions example

The admission figures of the grad school at UC Berkeley in 1973: 8442 (44%) men, 4321 (35%) women admitted.

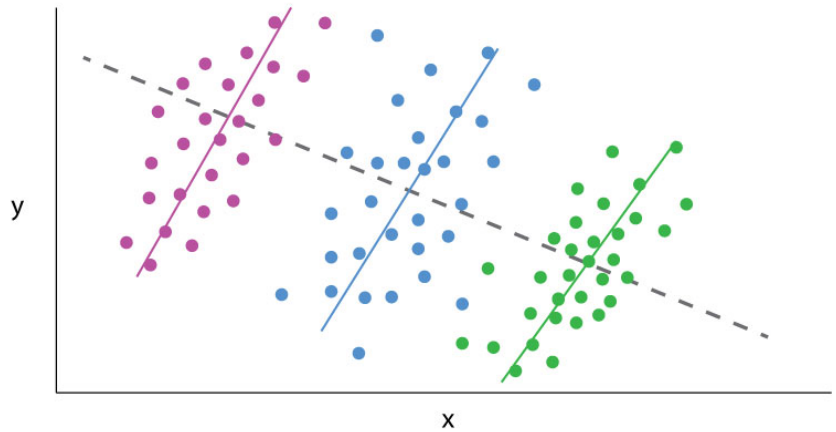
The same data conditioned on the department are:

Department	Men		Women	
	Applicants	Admited	Applicants	Admited
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

"Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation."

(Bickel et al, *Sex Bias in Graduate Admissions: Data From Berkeley*, Science, 1975)

Is there a fix?

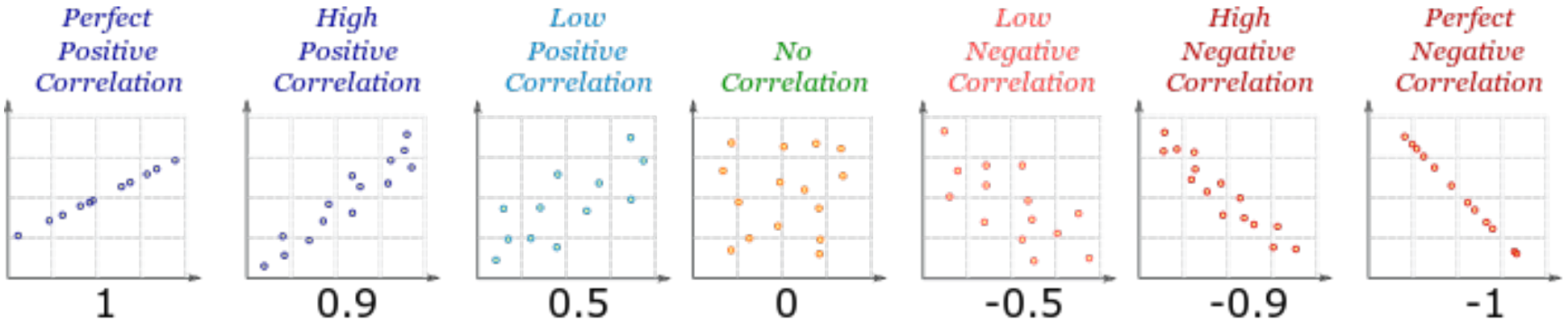


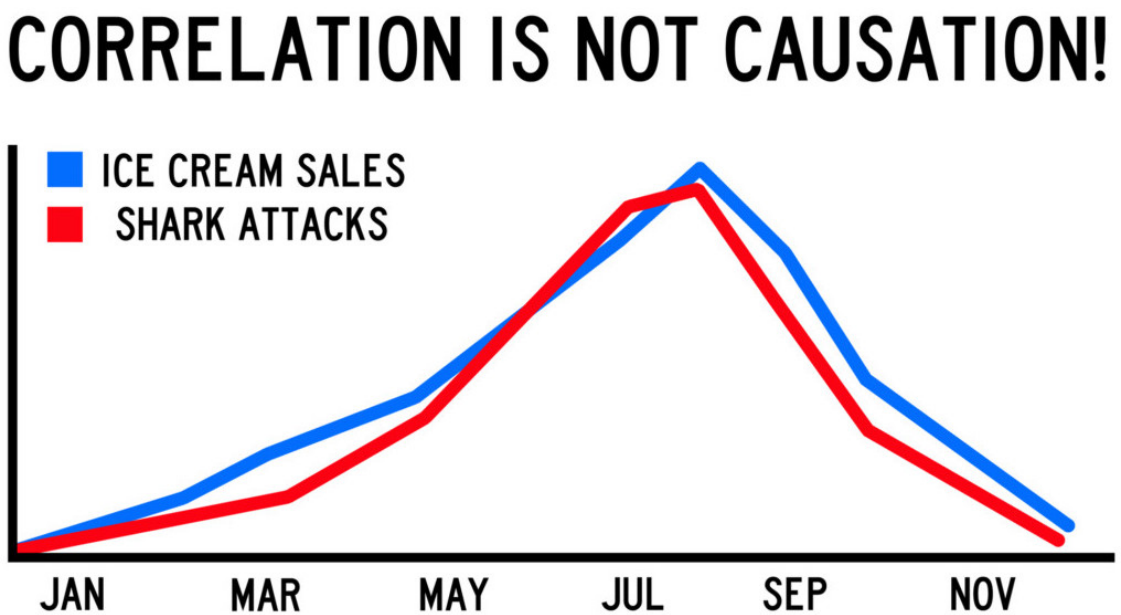
This shows that interpreting data analysis results may be complicated.

Given a data set, can we tell if there is an unobserved variable such that they are independent when controlled for this unobserved variable?

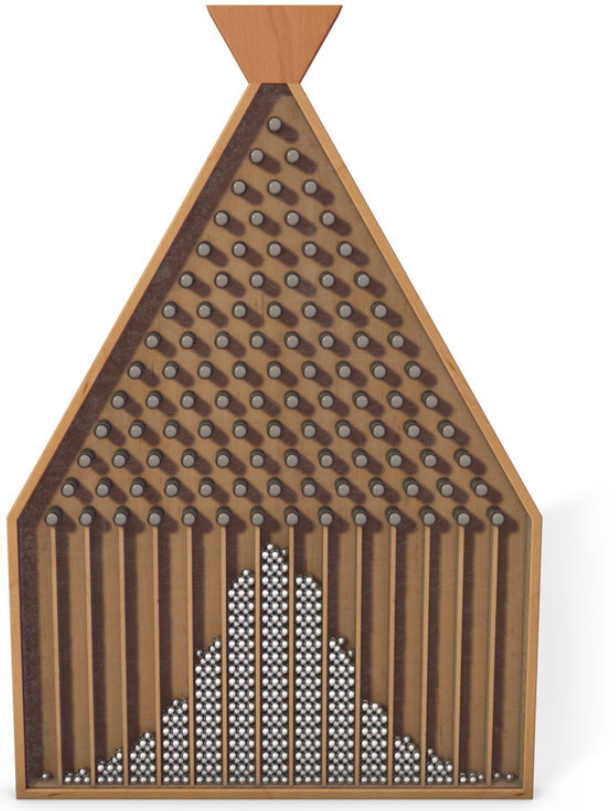
Correlation and independence

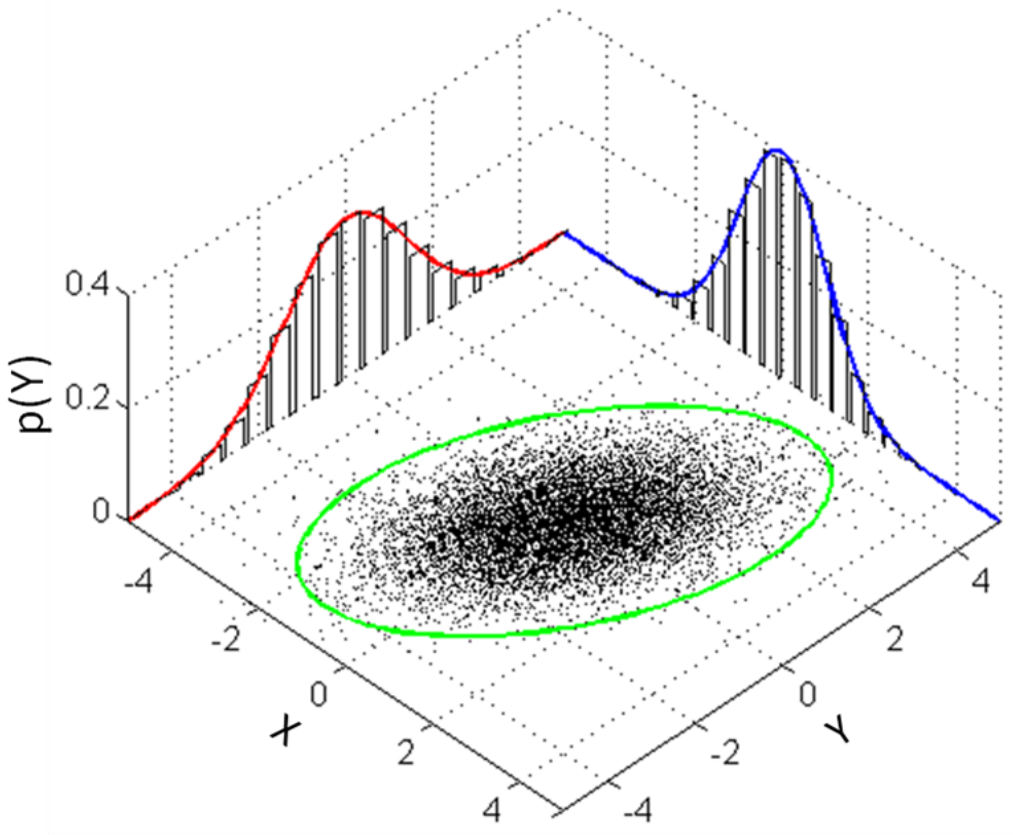
Correlation is the most basic measure of relation between two variables.





Gaussian random variables



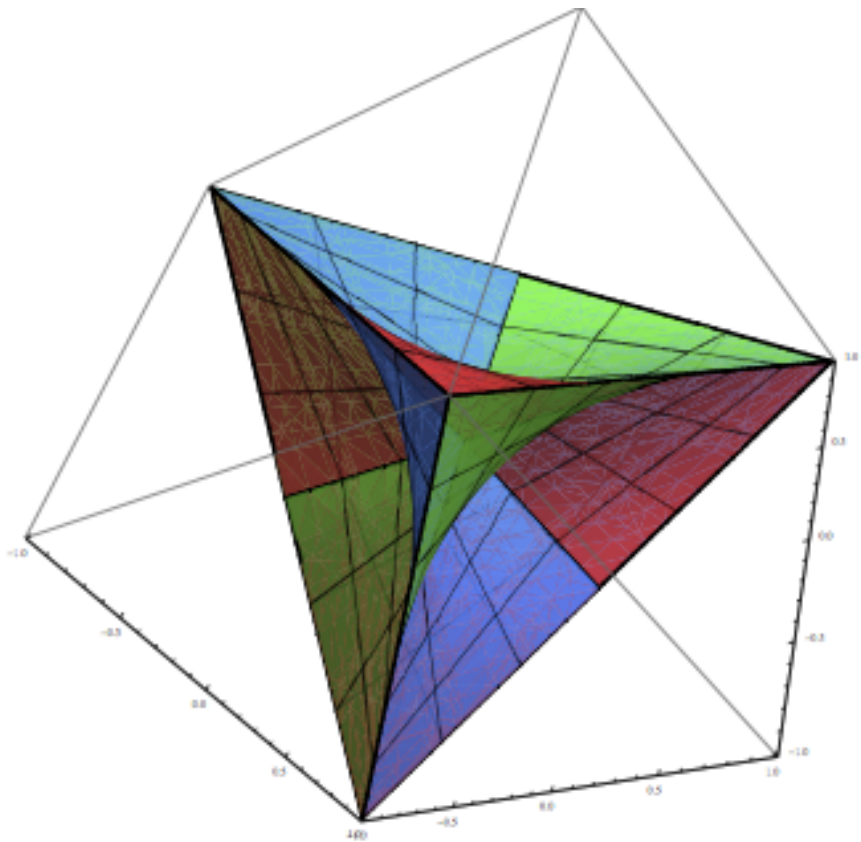


Three Gaussian variables

Gaussian X, Y, Z with correlations $\rho_{XY}, \rho_{XZ}, \rho_{YZ}$.

Theorem: There is an unobserved Gaussian variable that makes them independent if and only if

$$\rho_{XY} \geq \rho_{XZ}\rho_{YZ}, \qquad \rho_{XZ} \geq \rho_{XY}\rho_{YZ}, \qquad \rho_{YZ} \geq \rho_{XY}\rho_{XZ}.$$



Three binary variables

Binary variables: 0/1, true/false, yes/no, on/off, head/tail.

X, Y, Z with probabilities $p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}$.

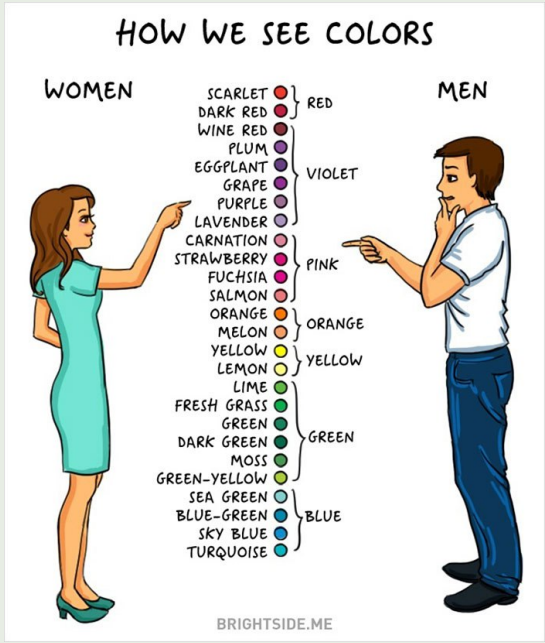
Theorem: There is an unobserved binary variable that makes them independent if and only if

$$p_{000}p_{111} \geq p_{001}p_{110} \quad p_{000}p_{111} \geq p_{010}p_{101} \quad p_{000}p_{111} \geq p_{100}p_{011}$$
$$p_{001}p_{111} \geq p_{011}p_{101} \quad p_{010}p_{111} \geq p_{011}p_{110} \quad p_{100}p_{111} \geq p_{101}p_{110}.$$
$$p_{000}p_{011} \geq p_{001}p_{010} \quad p_{000}p_{101} \geq p_{001}p_{100} \quad p_{000}p_{110} \geq p_{010}p_{100}$$

$X=$ has short hair (yes/no)

$Y=$ likes watching fútbol (yes/no)

$Z=$ recognizes less than 9 colors (yes/no).



Example: EPH-gestosis

- Dataset collected 40 years ago in a study on "Pregnancy and Child Development"
- EPH-gestosis (pre-eclampsia): disease syndrome for pregnant women; three symptoms (high body water retention, high amounts of urinary proteins, elevated blood pressure)
- A **syndrome** is a set of medical symptoms that are **correlated** with each other and, often, with a particular disease or disorder.

The sample distribution

$$\begin{bmatrix} \hat{p}_{000} & \hat{p}_{010} & \hat{p}_{001} & \hat{p}_{011} \\ \hat{p}_{100} & \hat{p}_{110} & \hat{p}_{101} & \hat{p}_{111} \end{bmatrix} = \frac{1}{4649} \begin{bmatrix} 3299 & 107 & 1012 & 58 \\ 78 & 11 & 65 & 19 \end{bmatrix}$$

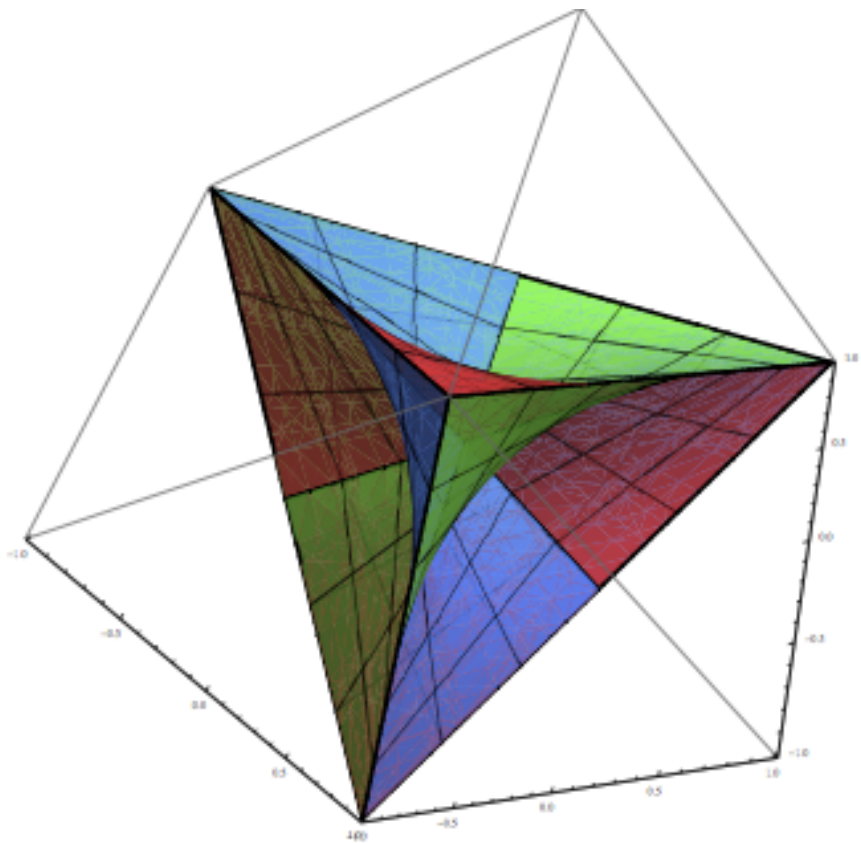
satisfies the given inequalities.

Three binary variables

This set has dimension 7 and so it cannot be drawn. Its slice satisfying

$$p_{000} = p_{111}, \quad p_{100} = p_{011}, \quad p_{010} = p_{101}, \quad p_{001} = p_{110}$$

looks like this...



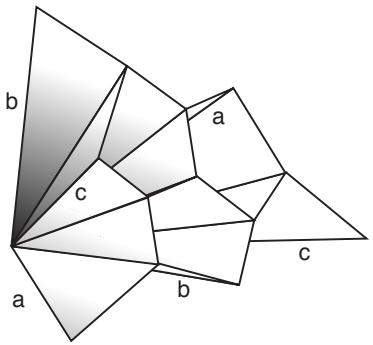
Algebraic statistics and nice pictures

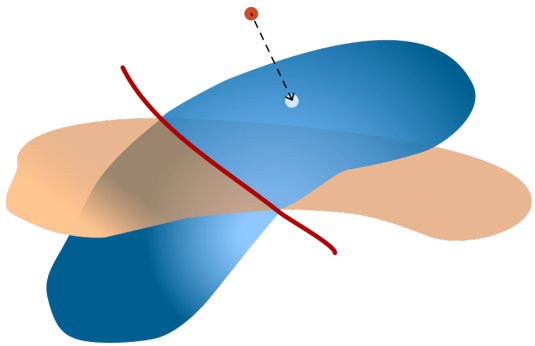
In applied statistics we formulate models to explain what we observe.

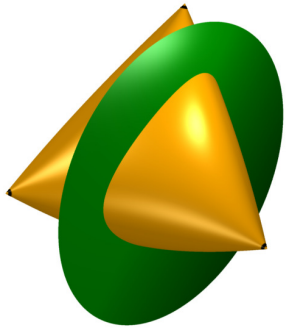
To understand the data we also need to understand the models.

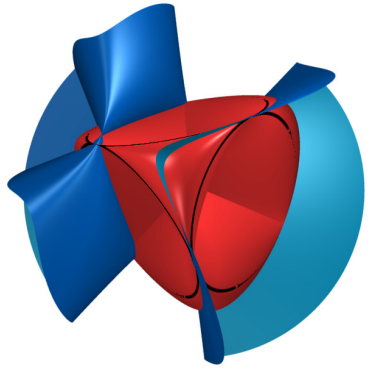
Typical tools to study statistical models are probabilistic and analytic.

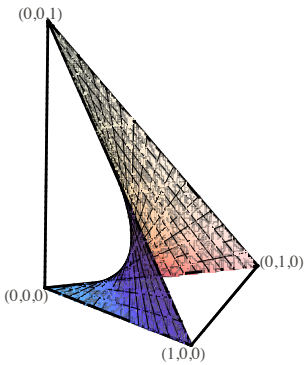
Algebraic statistics complements the standard toolbox with a set of algebraic, combinatorial, and geometric techniques.

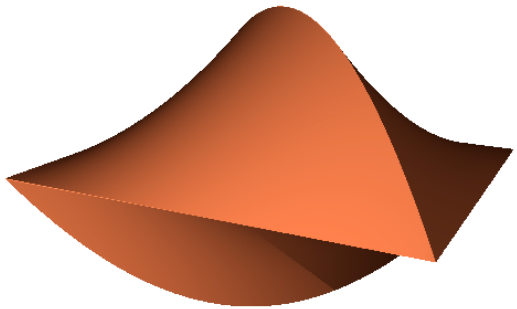


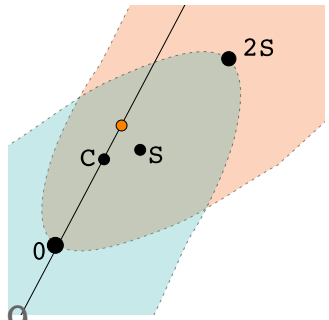


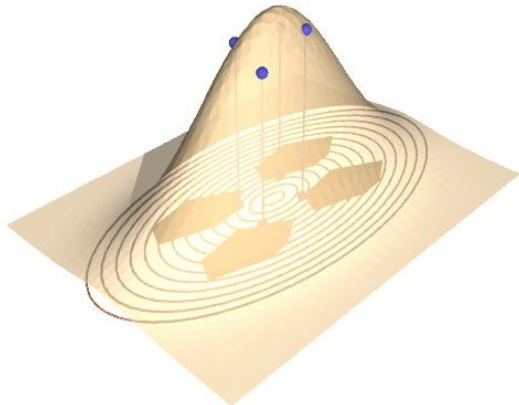


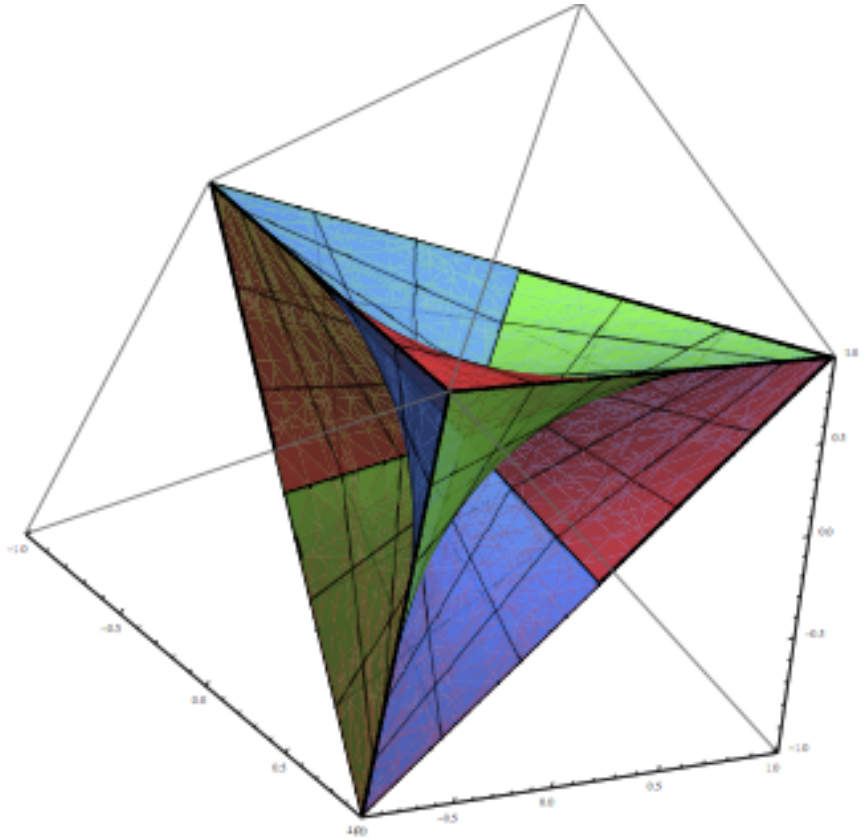












Thank you!