Robust and Nonparametric Statistical Tools for Big Data in Neuroscience

UNIVERSIDAD EAFIT

Henry Laniado.
Seminar
PhD in Mathematical Engineering

August 18, 2017

Motivation: fMRI problem

Figure: **Functional Magnetic Resonance**

Motivation: fMRI problem

Figure: **Stimulus**

Motivation: fMRI problem

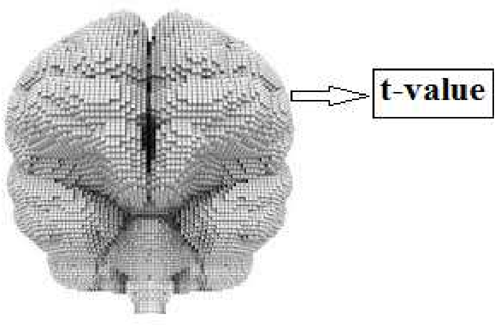


Figure: **900000 Voxels**

Motivation: fMRI problem

The theoretical activity is compared with that observed in each voxel.

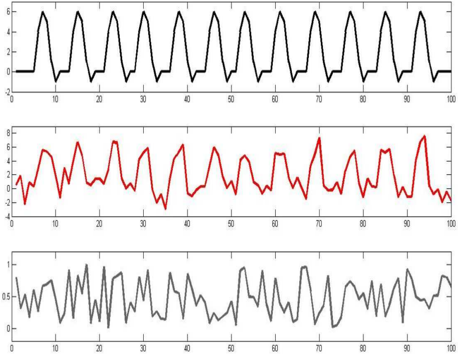


Figure: **Activation and non-activation**

Motivation: fMRI problem

The theoretical activity is compared with that observed in each voxel.

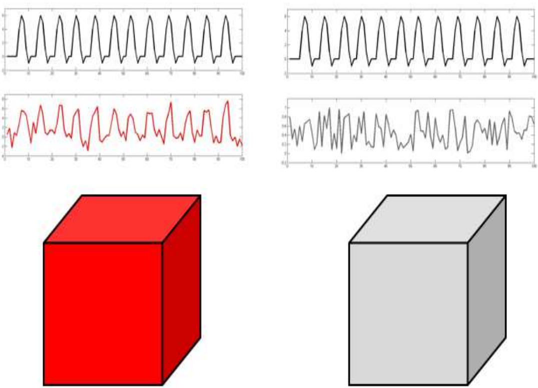


Figure: **Activation and non-activation**

Motivation: fMRI problem

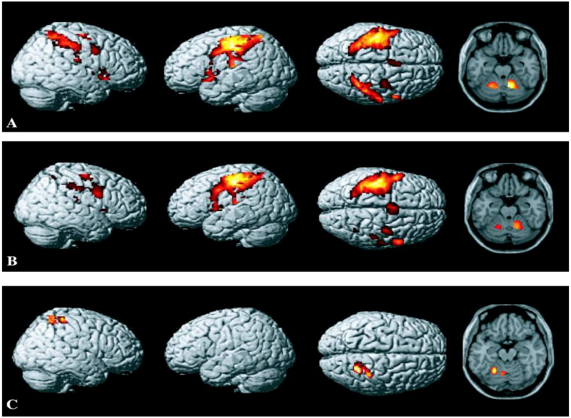


Figure: **Activation and non-activation**

Motivation: fMRI problem

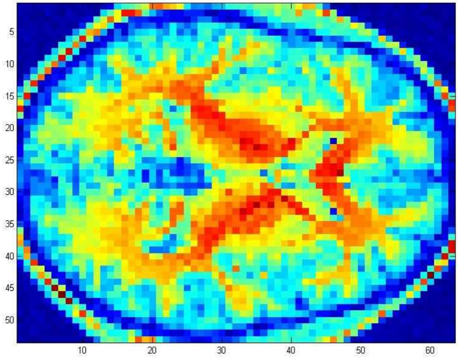


Figure: **Activation and non-activation**

Motivation: fMRI problem

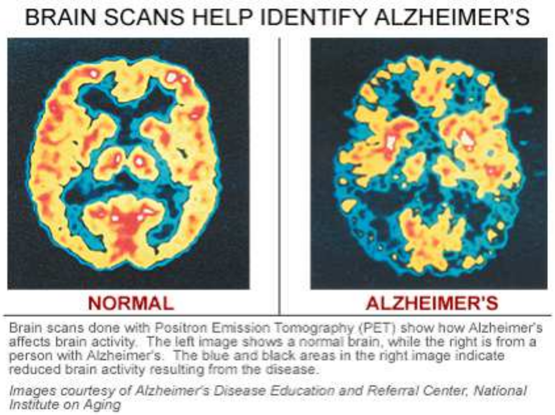


Figure: **Identification**

Main Goal

Find robust and non parametric estimators to explain linear dependence and apply them in the construction of covariance matrices to improve the significance of linear multivariate statistical models in the presence of outliers

Main Goal

Find robust and non parametric estimators to explain linear dependence and apply them in the construction of covariance matrices to improve the significance of linear multivariate statistical models in the presence of outliers

Hypothesis 1

Significance in linear models is too sensitive in the presence of outliers.

Main Goal

Find robust and non parametric estimators to explain linear dependence and apply them in the construction of covariance matrices to improve the significance of linear multivariate statistical models in the presence of outliers

Hypothesis 1

Significance in linear models is too sensitive in the presence of outliers.

Hypothesis 2

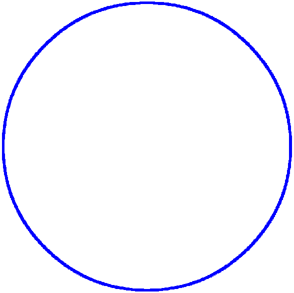
Significance in linear models can be improved with robust estimators or non parametric estimators

Outline

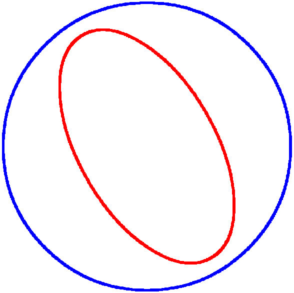
1. **Robustness: a fast review**
2. **General Lineal Model (GLM)**
3. **Robust GLM, with an explanatory variables**
4. **Robust GLM, multivariate case**
5. **Non-parametric estimators**
6. **Conclusions**
7. **Research lines**

Population: unknown parameters. $\theta = \int$

Population: unknown parameters. $\theta = \int$

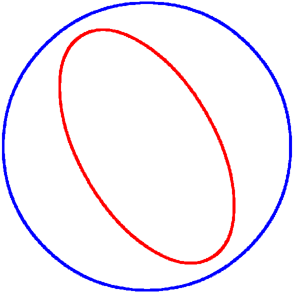


Population: unknown parameters. $\theta = \int$



Sample: estimators. $\hat{\theta} = \sum$

Population: unknown parameters. $\theta = \int$



Sample: estimators. $\hat{\theta} = \sum$

$\mu = \int f(x)dx$ **the estimator** $\bar{X} = \frac{\sum x_i}{n}$

Example

Table: Simulated data from $N(\color{red}{0}, 1)$

0.354	-0.838	0.1204	-0.353	-0.3853	2.1261	-0.268
-0.16	-1.038	-0.273	0.7939	0.45153	0.858	0.6697
-1.838	2.5976	-0.282	0.1757	-0.5986	0.9407	-0.036
-0.531	0.6208	1.4303	-1.044	-0.1349	0.2476	-0.397
-1.432	1.0849	-1.12	0.6788	-1.1078	-0.245	0.5031
-1.136	-0.011	1.5668	0.6654	-0.6194	-1.481	-0.446
2.0077	-0.898	-0.886	-2.145	-2.435	-0.396	-0.116

Example

Table: Simulated data from $N(0, 1)$

0.354	-0.838	0.1204	-0.353	-0.3853	2.1261	-0.268
-0.16	-1.038	-0.273	0.7939	0.45153	0.858	0.6697
-1.838	2.5976	-0.282	0.1757	-0.5986	0.9407	-0.036
-0.531	0.6208	1.4303	-1.044	-0.1349	0.2476	-0.397
-1.432	1.0849	-1.12	0.6788	-1.1078	-0.245	0.5031
-1.136	-0.011	1.5668	0.6654	-0.6194	-1.481	-0.446
2.0077	-0.898	-0.886	-2.145	-2.435	-0.396	-0.116

$$\bar{X} = \frac{\sum_{i=1}^{49} x_i}{49} = -0.09$$

Example

Table: Data from $N(0, 1)$. Contaminated data

0.354	-0.838	0.1204	-0.353	-0.3853	2.1261	-0.268
-0.16	-1.038	-0.273	8.2356	0.45153	0.858	0.6697
-1.838	2.5976	-0.282	0.1757	-0.5986	0.9407	12.266
-3.82	0.6208	1.4303	-1.044	-0.1349	0.2476	-0.397
-1.432	1.0849	-1.12	0.6788	-1.1078	-0.245	0.5031
-1.136	-0.011	1.5668	0.6654	-0.6194	-1.481	-0.446
2.0077	-0.898	-0.886	-2.145	-5.1321	-0.396	9.266

Example

Table: Data from $N(0, 1)$. Contaminated data

0.354	-0.838	0.1204	-0.353	-0.3853	2.1261	-0.268
-0.16	-1.038	-0.273	8.2356	0.45153	0.858	0.6697
-1.838	2.5976	-0.282	0.1757	-0.5986	0.9407	12.266
-3.82	0.6208	1.4303	-1.044	-0.1349	0.2476	-0.397
-1.432	1.0849	-1.12	0.6788	-1.1078	-0.245	0.5031
-1.136	-0.011	1.5668	0.6654	-0.6194	-1.481	-0.446
2.0077	-0.898	-0.886	-2.145	-5.1321	-0.396	9.266

$$\bar{X} = \frac{\sum_{i=1}^{49} x_i}{49} = 0.4252$$

Example

Table: Data from $N(0, 1)$. Contaminated data

-5.132	-3.82	-2.1448	-1.4809	-1.4318	-1.14	-1.1199
-1.108	-1.044	-1.0384	-0.8979	-0.886	-0.84	-0.6194
-0.599	-0.531	-0.4462	-0.3974	-0.3956	-0.39	-0.2819
-0.273	-0.268	-0.2455	-0.1604	-0.1349	-0.01	0.1204
0.176	0.2476	0.354	0.45153	0.5031	0.621	0.6654
0.67	0.6788	0.7939	0.85803	0.9407	1.085	1.4303
1.567	2.0077	2.1261	2.59761	8.2356	9.266	12.266

$$\bar{X} = \frac{\sum_{i=1}^{49} x_i}{49} = 0.4252$$

Example

Table: Data from $N(0, 1)$. Contaminated data

X	X	X	-1.4809	-1.4318	-1.14	-1.1199
-1.108	-1.044	-1.0384	-0.8979	-0.886	-0.84	-0.6194
-0.599	-0.531	-0.4462	-0.3974	-0.3956	-0.39	-0.2819
-0.273	-0.268	-0.2455	-0.1604	-0.1349	-0.01	0.1204
0.176	0.2476	0.354	0.45153	0.5031	0.621	0.6654
0.67	0.6788	0.7939	0.85803	0.9407	1.085	1.4303
1.567	2.0077	2.1261	2.59761	X	X	X

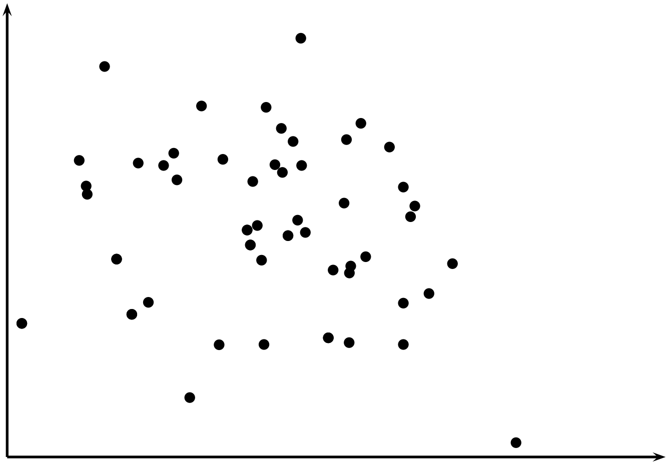
Trimmed mean

$$\bar{X}_\alpha = \frac{\sum_{i=k_1}^{k_2} x_{[i]}}{k_2 - k_1 + 1}$$

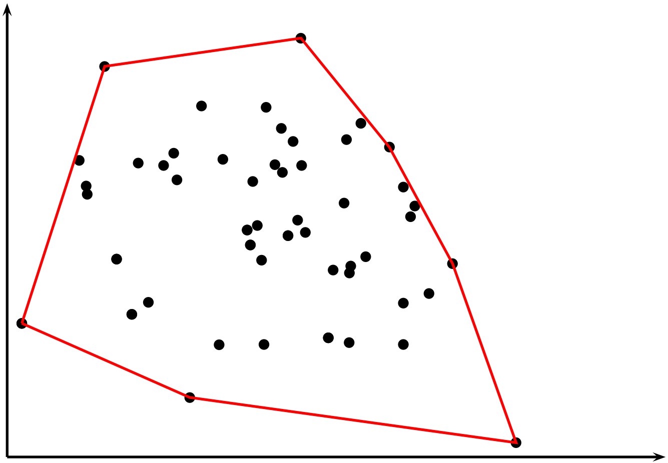
$k_1 = [\frac{\alpha}{2} n] \quad k_2 = [n(1 - \frac{\alpha}{2})]$

$$\bar{X}_{0.1} = 0.0503$$

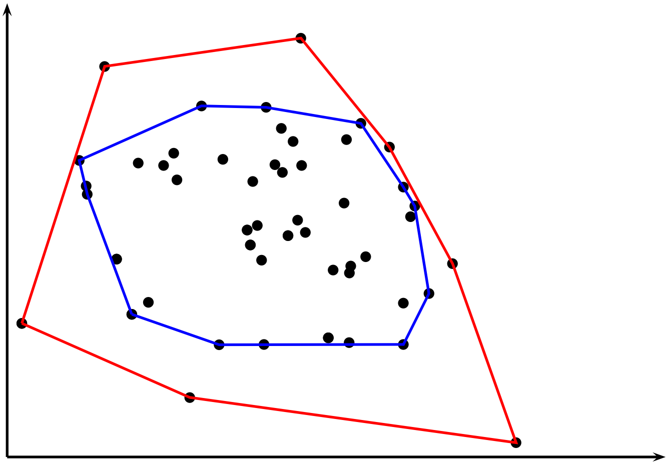
Depth



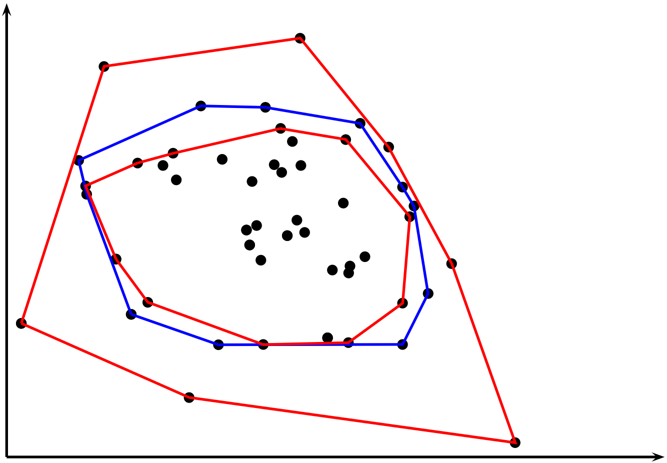
Depth



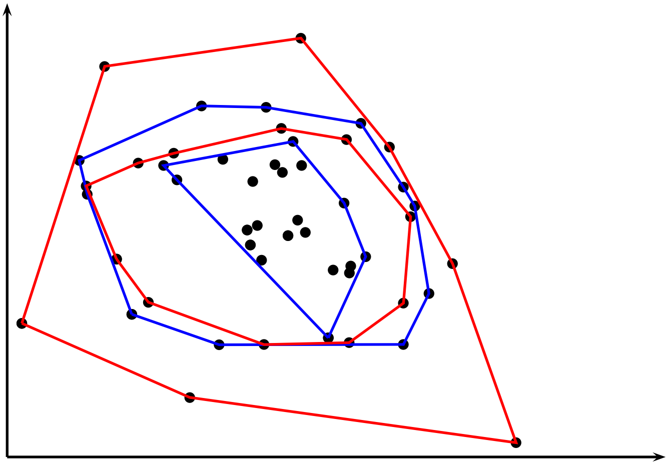
Depth



Depth



Depth



Dependent uniform distribution in dimension 2

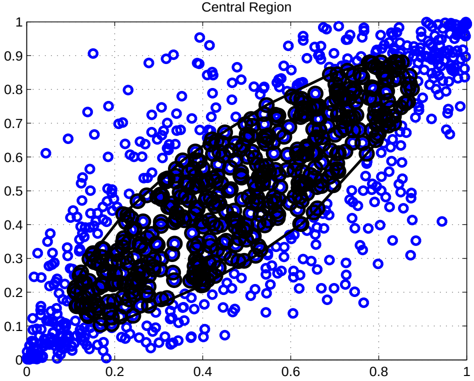


Figure: Random Tukey Depth

Normal distribution in dimension 3

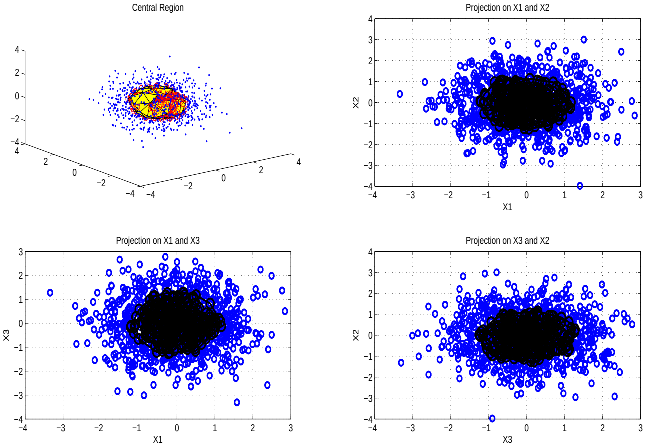


Figure: Random Tukey Depth

Central Region: univariate case

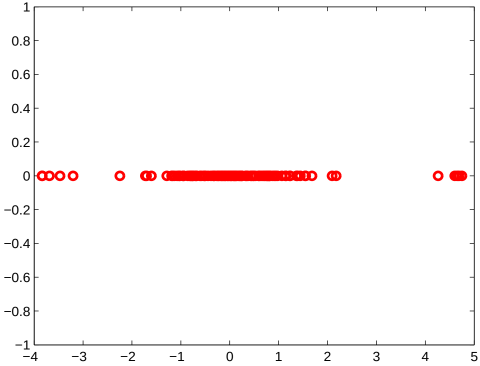


Figure: Data

Central Region: univariate case

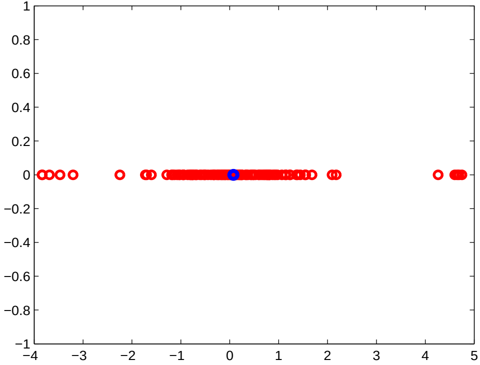


Figure: **Median**

Central Region: univariate case

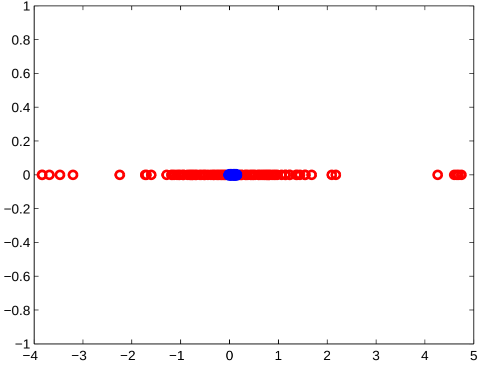


Figure: **Central Region** 10%

Central Region: univariate case

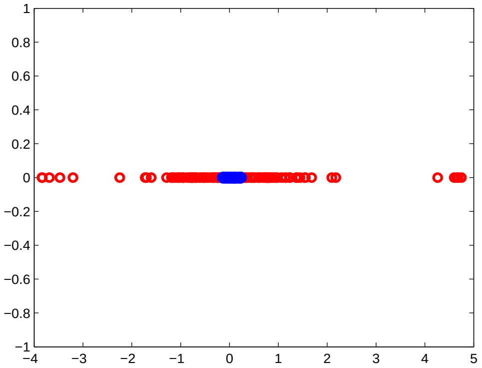


Figure: **Central Region** 20%

Central Region: univariate case

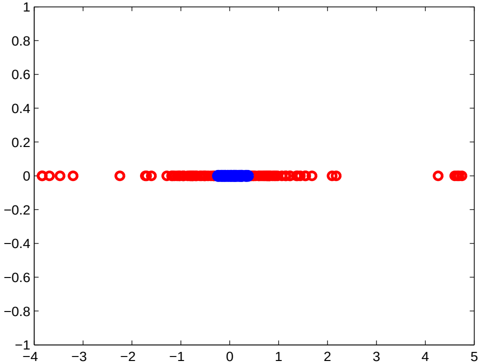


Figure: **Central Region** 30%

Central Region: univariate case

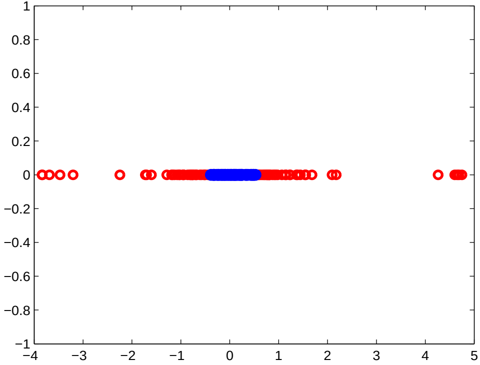


Figure: **Central Region** 40%

Central Region: univariate case

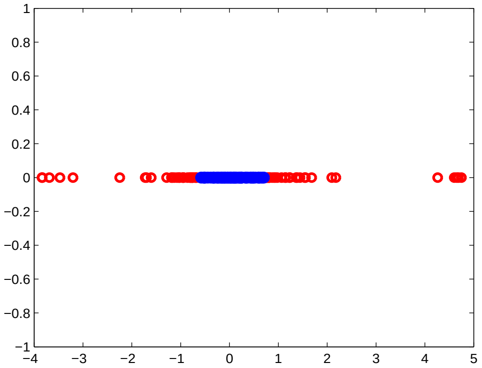


Figure: **Central Region** 50%

Central Region: univariate case

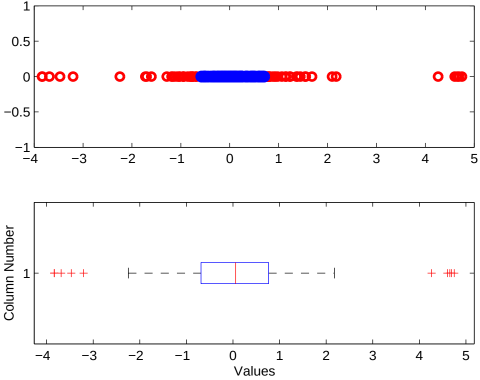


Figure: **Box plot**

Central Region: bivariate case

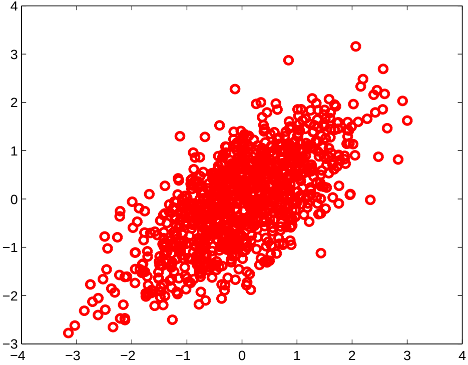


Figure: Data

Central Region: bivariate case

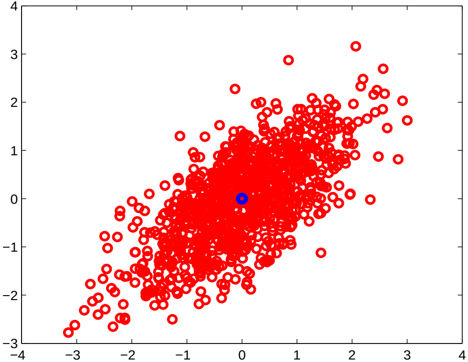


Figure: **deepest point**

Central Region: bivariate case

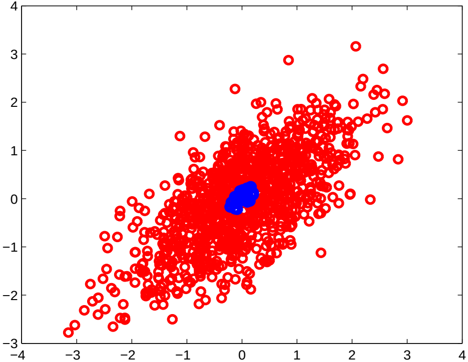


Figure: **Central Region** 10%

Central Region: bivariate case

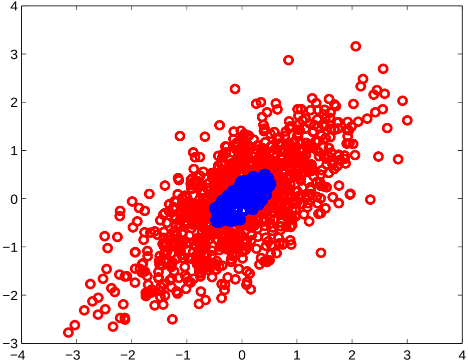


Figure: **Central Region** 30%

Central Region: bivariate case

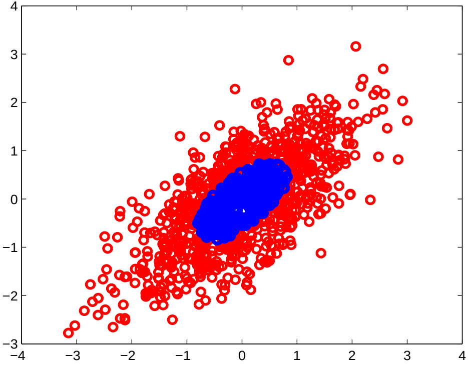


Figure: **Central Region** 40%

Central Region: bivariate case

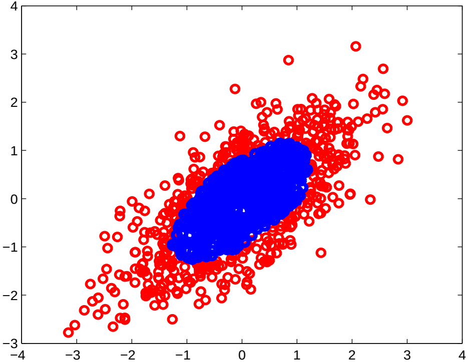


Figure: **Central Region** 50%

Central Region: functional case

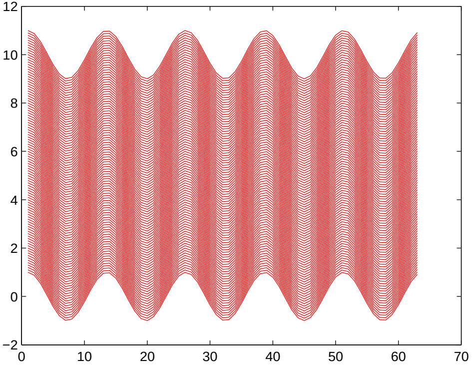


Figure: Data

Central Region: functional case

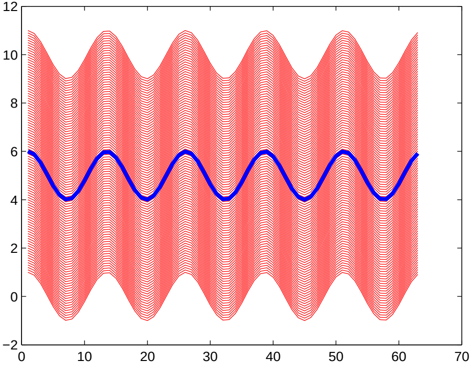


Figure: **deepest curve**

Central Region: functional case

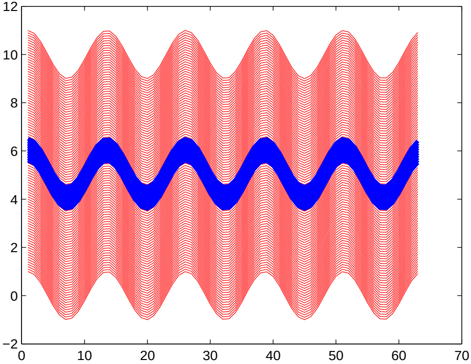


Figure: **Central Region** 10%

Central Region: functional case

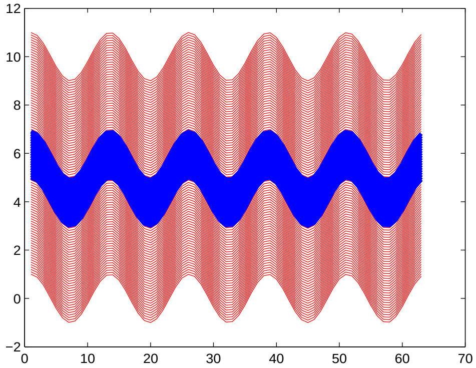


Figure: **Central Region** 20%

Central Region: functional case

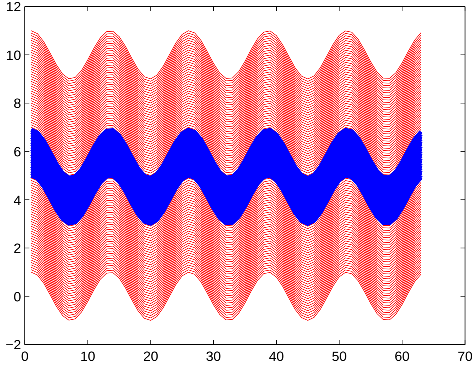


Figure: **Central Region** 20%

Central Region: functional case

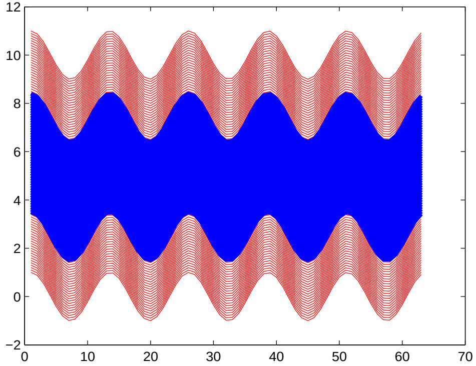


Figure: **Central Region** 50%

Central Region: functional case

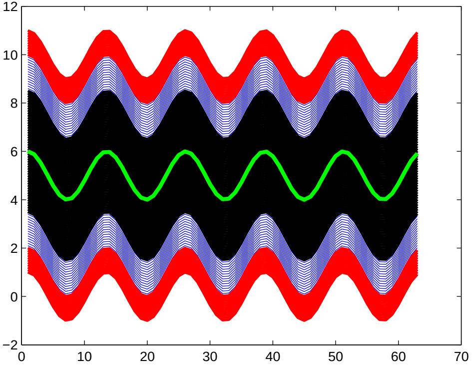


Figure: **Central Region, Extremes**

Central Region: functional case

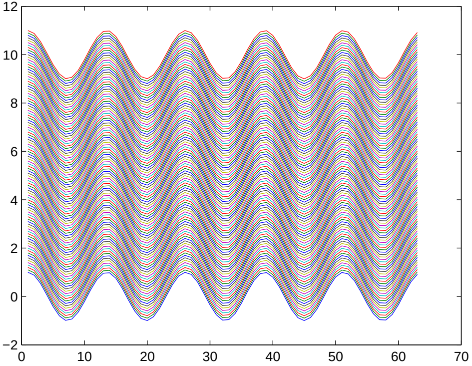


Figure: Data

Central Region: functional case

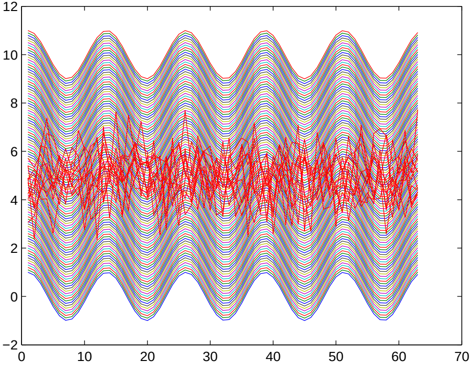


Figure: Contaminated Data

Central Region: functional case

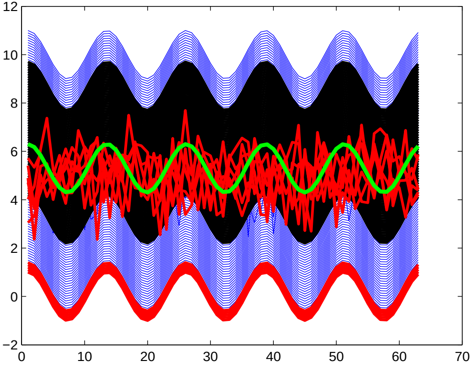
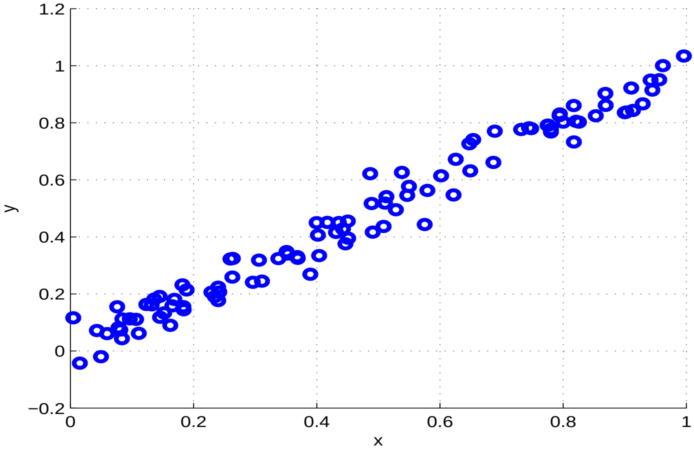


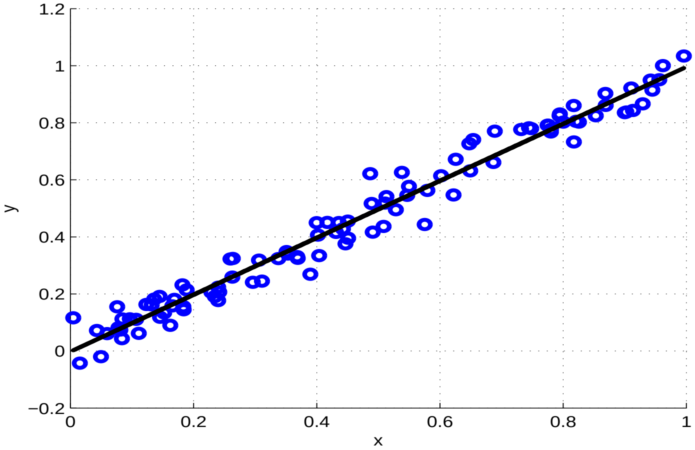
Figure: **Central Region, Extremes**

- Robustness:a fast review
- Central Region: bivariate case
- Central Region: functional case
- **General Linear Model (GLM)**
- Robust GLM, with an explanatory variable
  - Contaminating a 5%
  - Contaminating a 10%
- Robust GLM, Multiavarite case
  - Contaminating a 5%
  - Contaminating a 10%
- Non parametric estimators
- Conclusions
- Research Lines

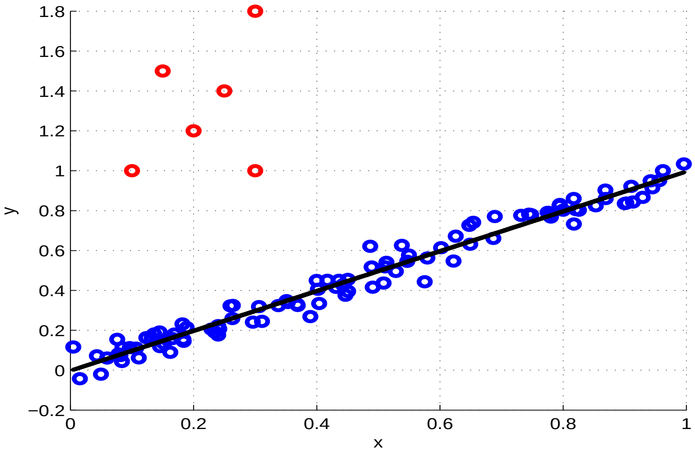
General Linear Model



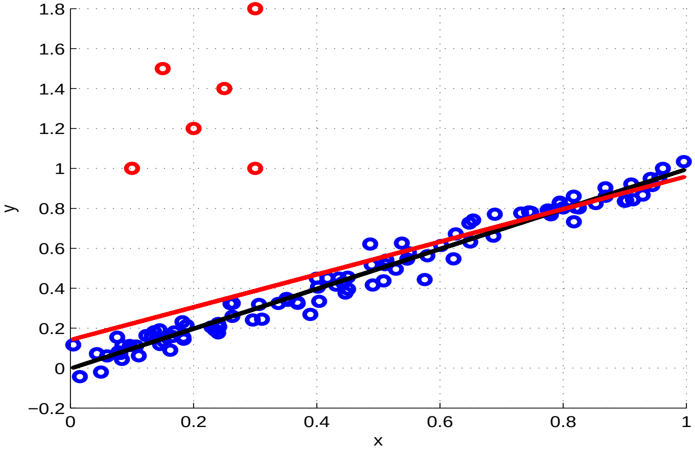
General Linear Model



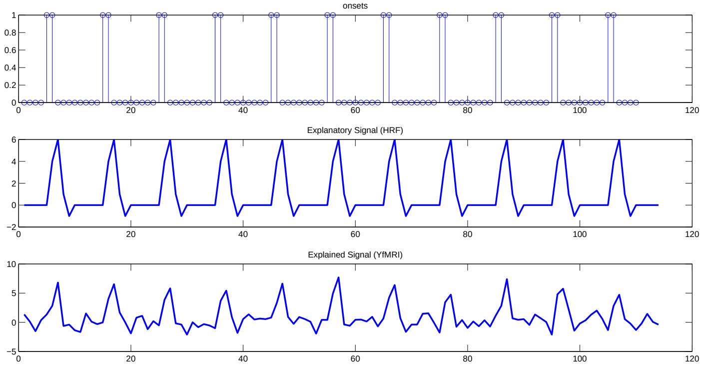
General Linear Model



General Linear Model

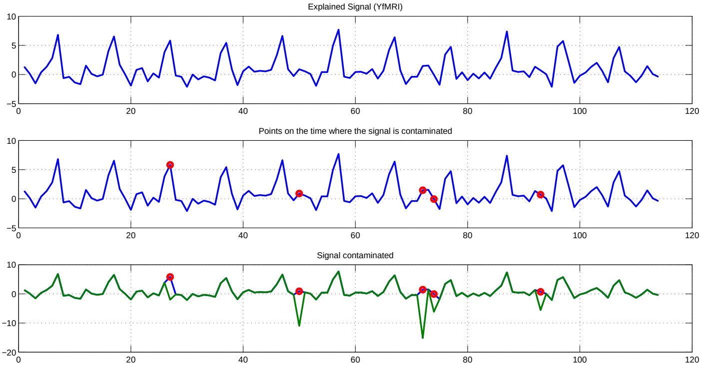


Significant variable



$$\mathbf{Y} = \mathbf{X} + N(0,1)$$

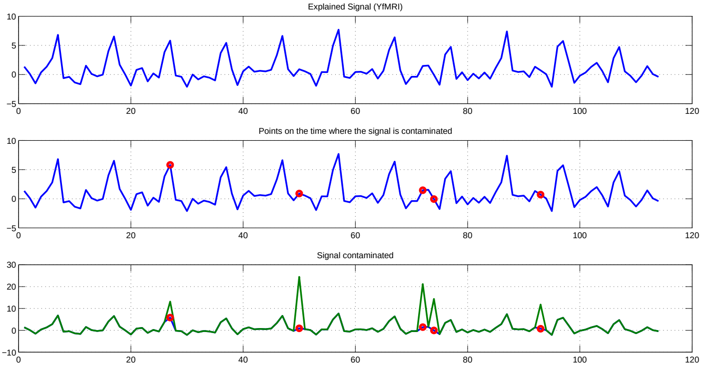
Significant variable and Contaminating a 5%



The red points will be generated by:

- **Normal**$(0, \sigma^2)$, σ^2 **varying from 1 to 100.**

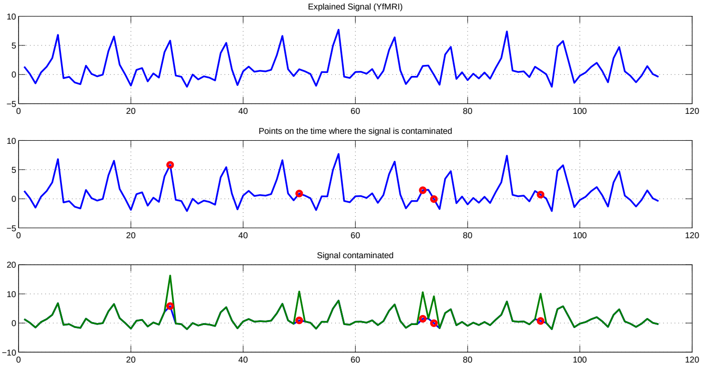
Significant variable and Contaminating a 5%



The red points will be generated by:

- **Normal**(μ, σ^2), **both varying from 1 to 100.**

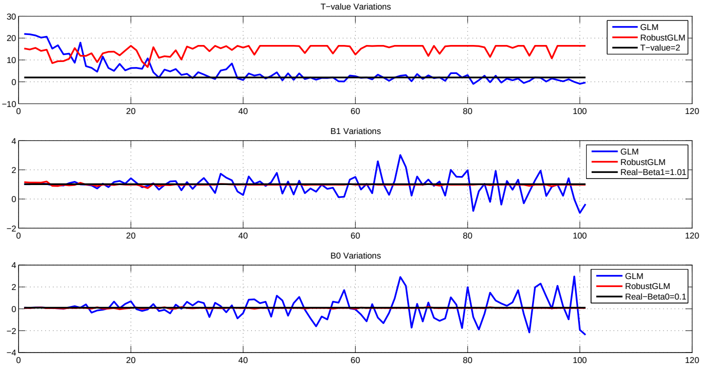
Significant variable and Contaminating a 5%



The red points will be generated by:

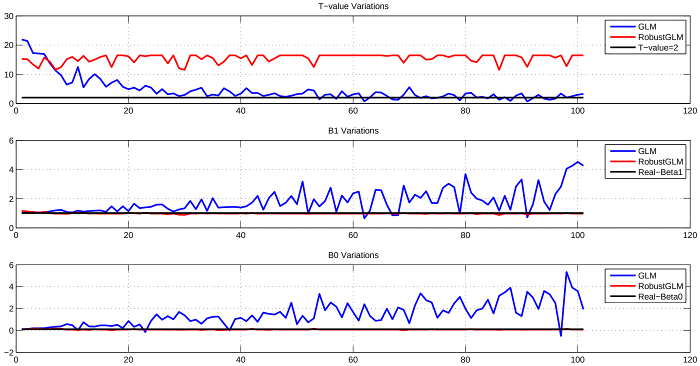
- **Normal**$(\mu, 1)$, μ **varying from 1 to 100.**

Significant variable and Contaminating a 5%



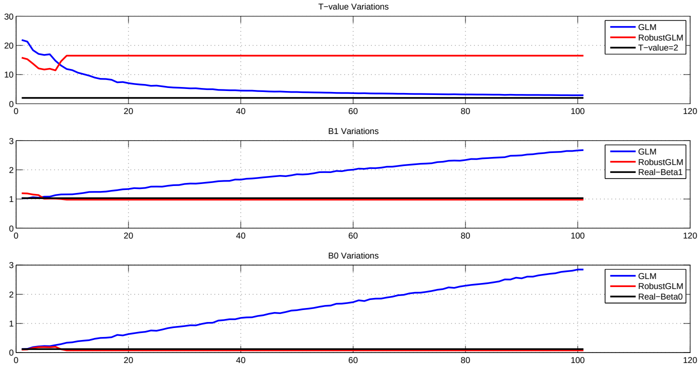
- **Normal$(0, \sigma^2)$, σ^2 varying from 1 to 100.**
- **X-axis represents σ^2 values varying from 1 to 100.**

Significant variable and Contaminating a 5%



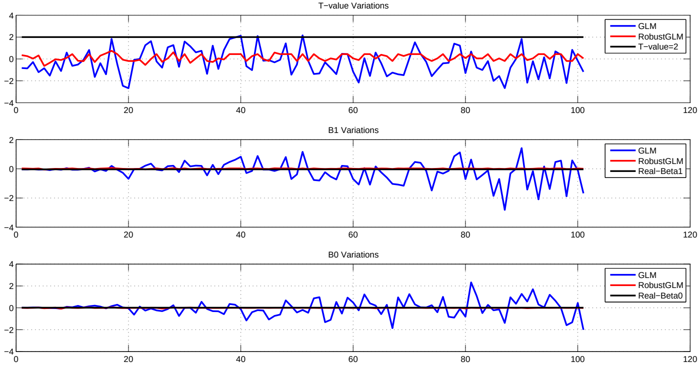
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \ \sigma^2]$ **values both varying from 1 to 100.**

Significant variable and Contaminating a 5%



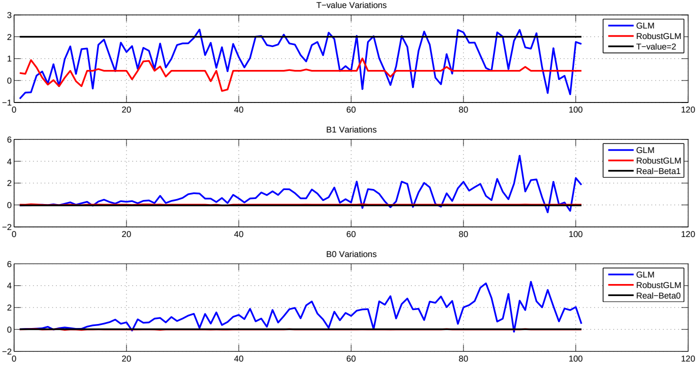
- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

No Significant variable and Contaminating a 5%



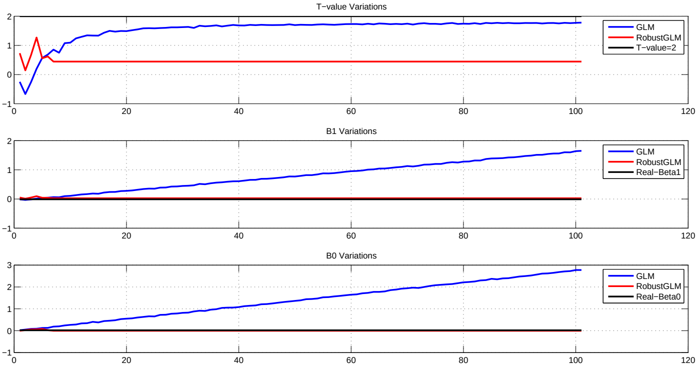
- **Normal$(0, \sigma^2)$, σ^2 varying from 1 to 100.**
- **X-axis represents σ^2 values varying from 1 to 100.**

No Significant variable and Contaminating a 5%



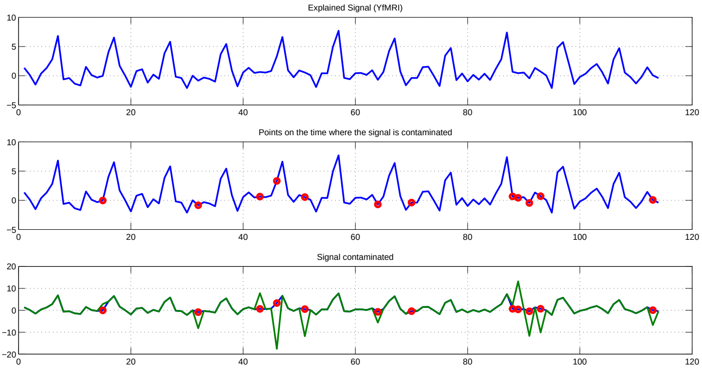
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \ \sigma^2]$ **values both varying from 1 to 100.**

No Significant variable and Contaminating a 5%



- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

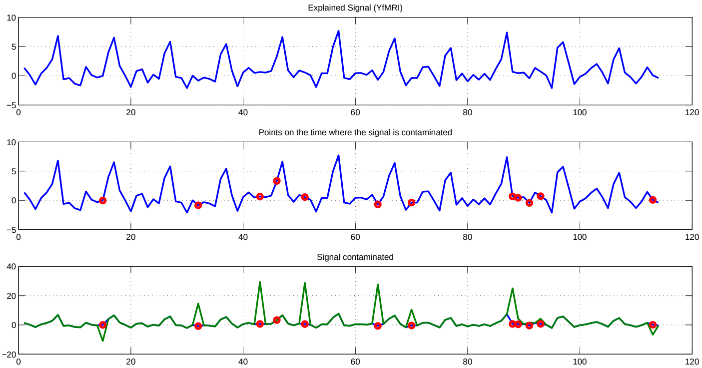
Significant variable and Contaminating a 10%



The red points will be generated by:

- **Normal**$(0, \sigma^2)$, σ^2 **varying from 1 to 100.**

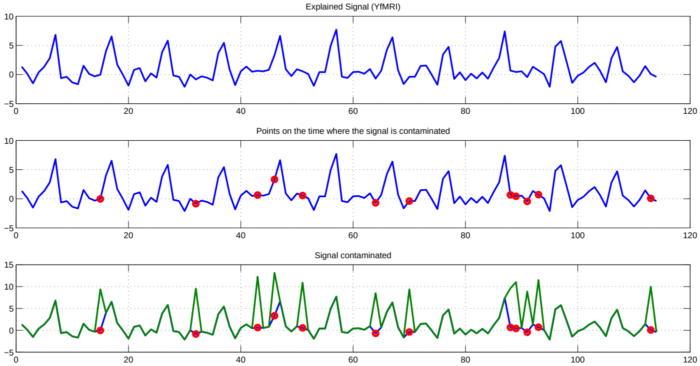
Significant variable and Contaminating a 10%



The red points will be generated by:

- **Normal**(μ, σ^2), **both varying from 1 to 100.**

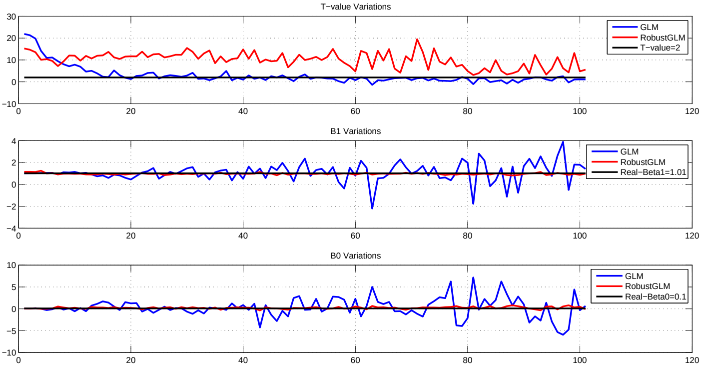
Significant variable and Contaminating a 10%



The red points will be generated by:

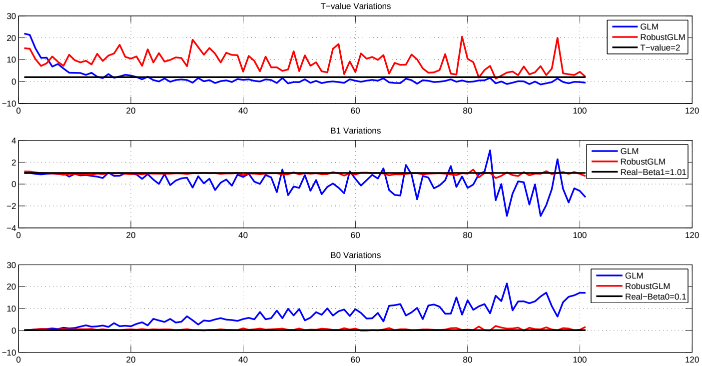
- **Normal**$(\mu, 1)$, μ **varying from 1 to 100.**

Significant variable and Contaminating a 10%



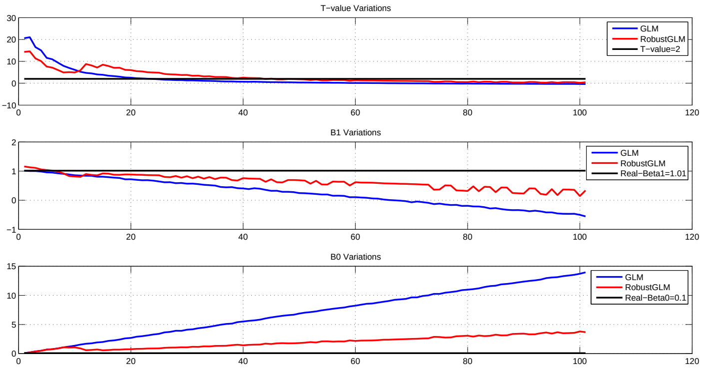
- **Normal$(0, \sigma^2)$, σ^2 varying from 1 to 100.**
- **X-axis represents σ^2 values varying from 1 to 100.**

Significant variable and Contaminating a 10%



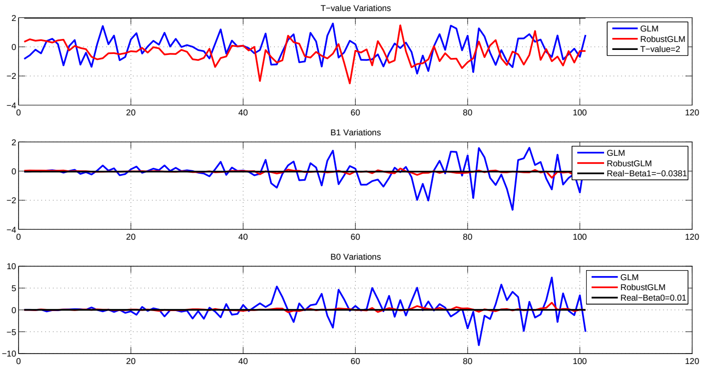
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \sigma^2]$ **values both varying from 1 to 100.**

Significant variable and Contaminating a 10%



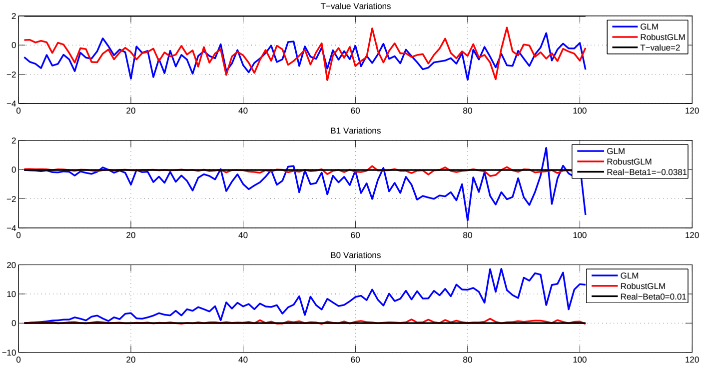
- **Normal**$(\mu, 1)$**,** μ **varying from 1 to 100.**
- X**-axis represents** μ **values varying from 1 to 100.**

No Significant variable and Contaminating a 10%



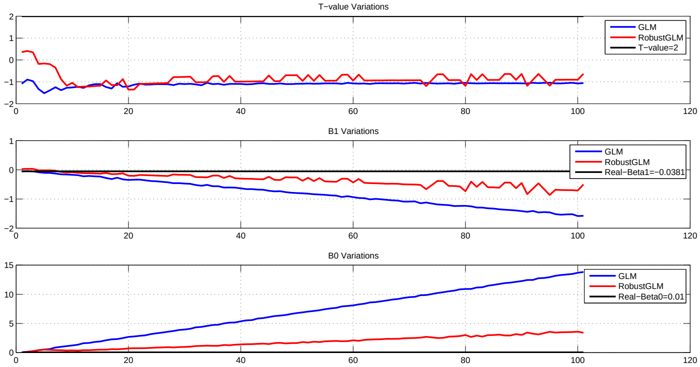
- **Normal**$(0, \sigma^2)$**,** σ^2 **varying from 1 to 100.**
- X**-axis represents** σ^2 **values varying from 1 to 100.**

No Significant variable and Contaminating a 10%



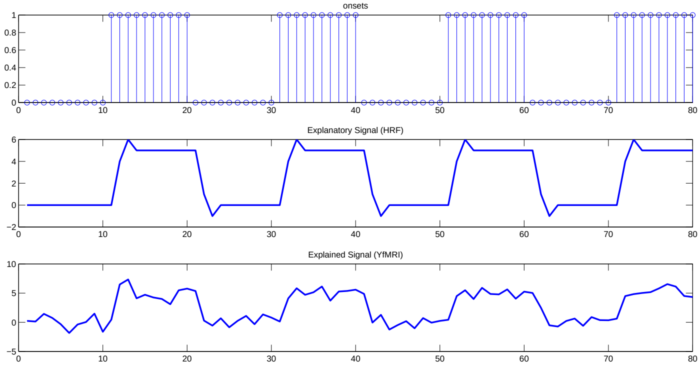
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \ \sigma^2]$ **values both varying from 1 to 100.**

No Significant variable and Contaminating a 10%



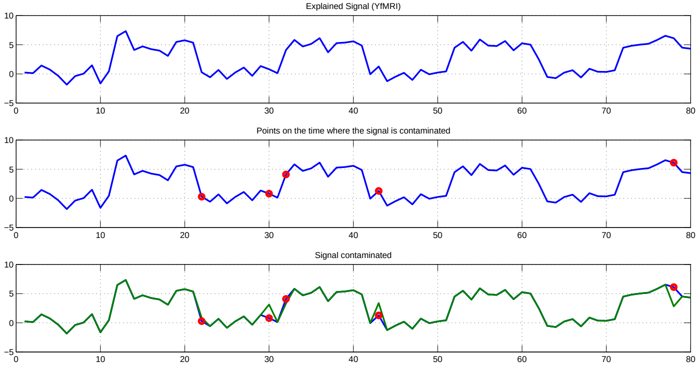
- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

Significant variable



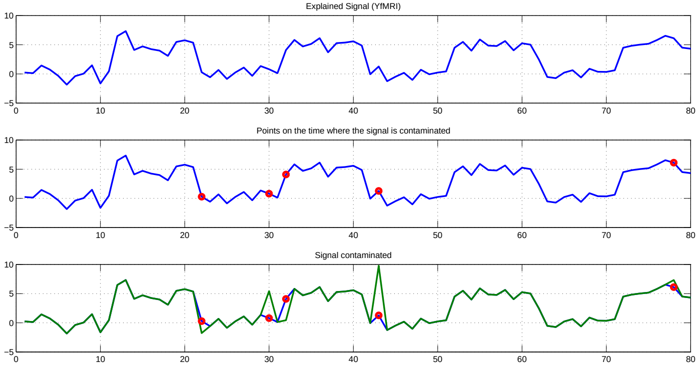
$$\mathbf{Y} = X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + N(0,1)$$

Significant variable and Contaminating a 5%



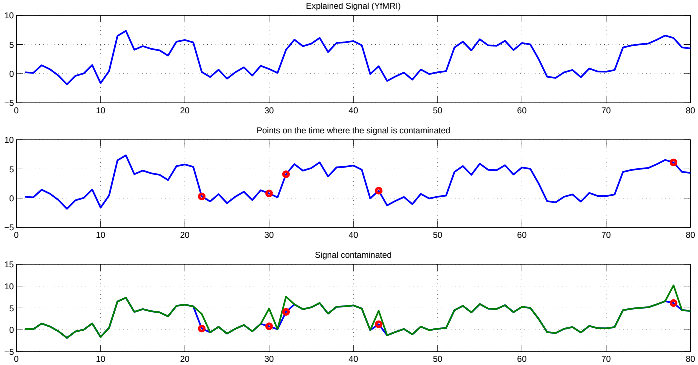
- **Normal**$(0, \sigma^2)$, σ^2 **varying from 1 to 100.**

Significant variable and Contaminating a 5%



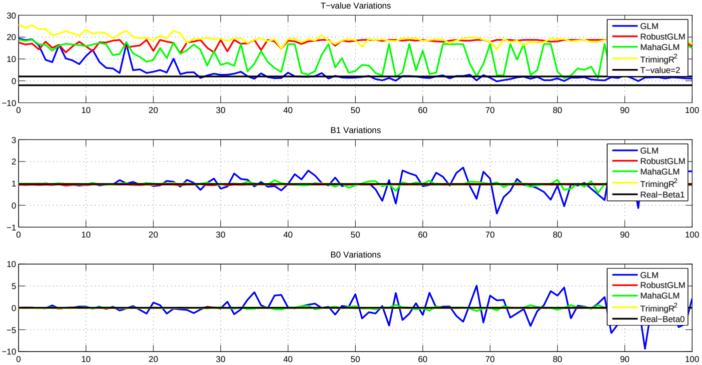
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**

Significant variable and Contaminating a 5%



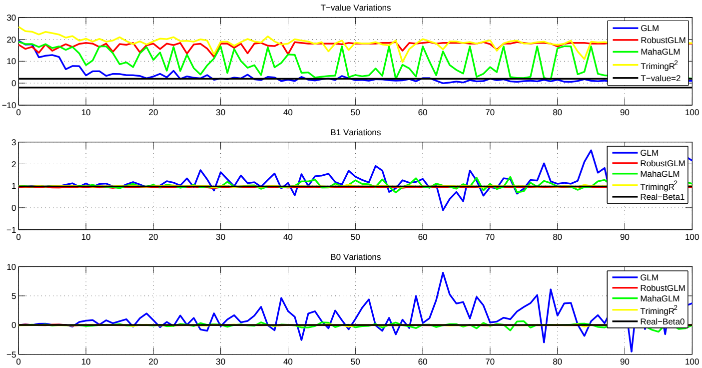
- **Normal** $(\mu, 1)$**,** μ **varying from 1 to 100.**

Significant variable and Contaminating a 5%



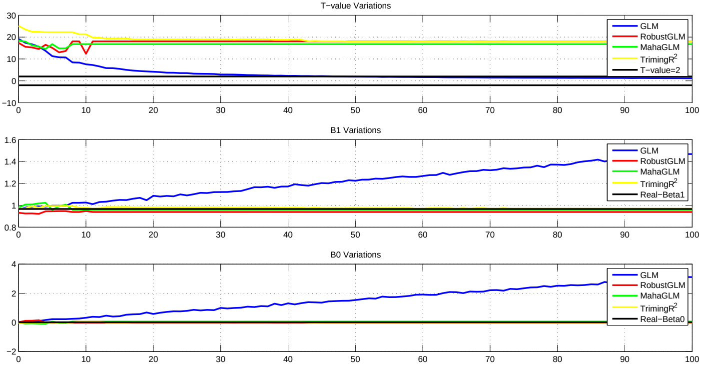
- **Normal**$(0, \sigma^2)$**,** σ^2 **varying from 1 to 100.**
- X**-axis represents** σ^2 **values varying from 1 to 100.**

Significant variable and Contaminating a 5%



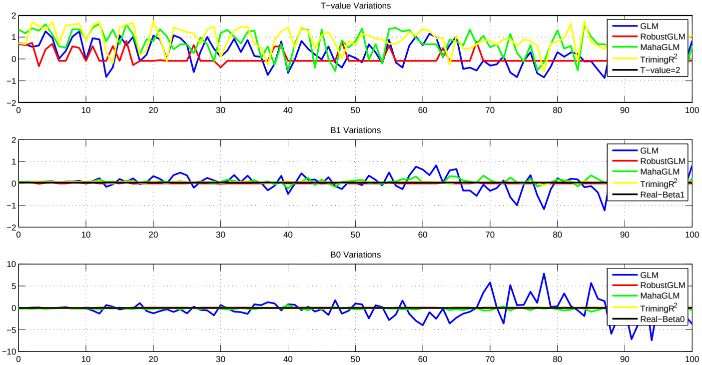
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \ \sigma^2]$ **values both varying from 1 to 100.**

Significant variable and Contaminating a 5%



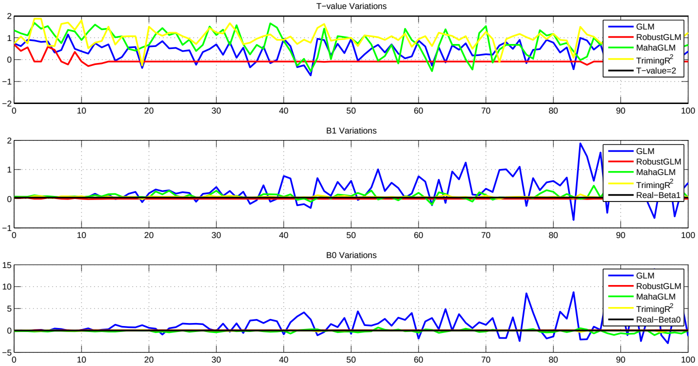
- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

No Significant variable and Contaminating a 5%



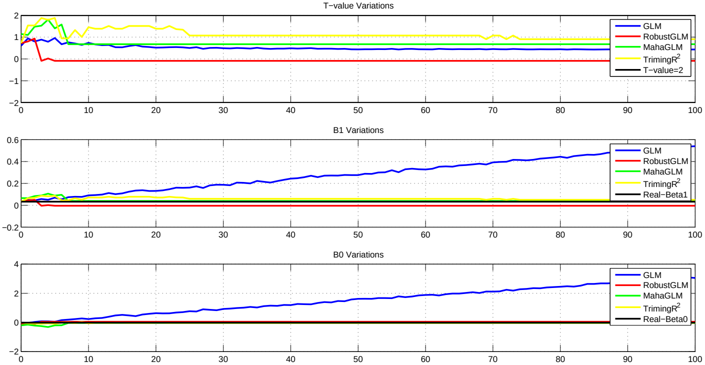
- **Normal**$(0, \sigma^2)$**,** σ^2 **varying from 1 to 100.**
- X**-axis represents** σ^2 **values varying from 1 to 100.**

No Significant variable and Contaminating a 5%



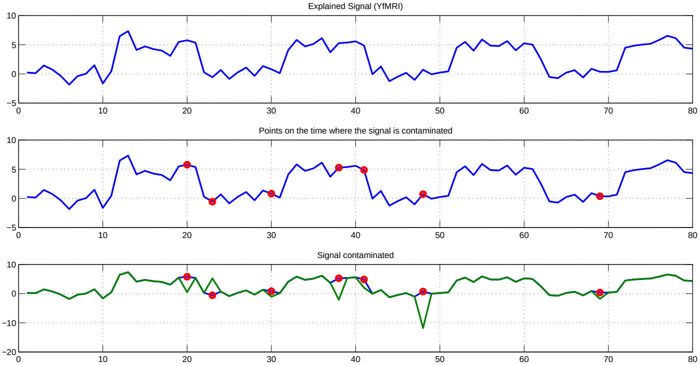
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \sigma^2]$ **values both varying from 1 to 100.**

No Significant variable and Contaminating a 5%



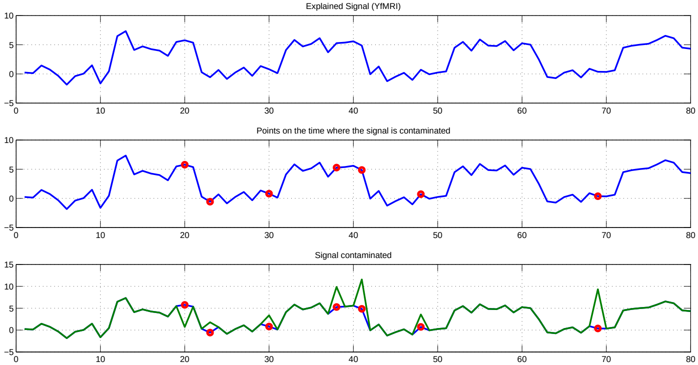
- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

Significant variable and Contaminating a 10%



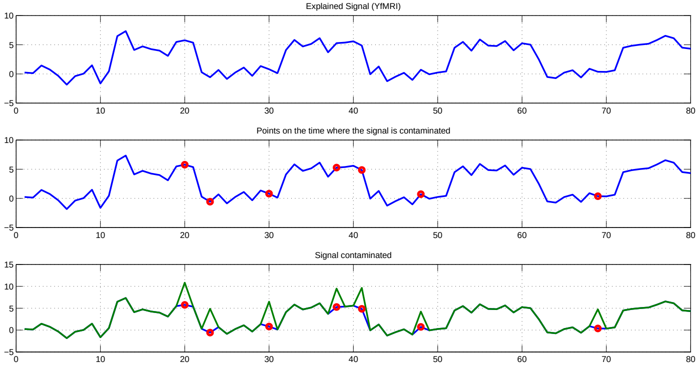
- **Normal**$(0, \sigma^2)$**,** σ^2 **varying from 1 to 100.**

Significant variable and Contaminating a 10%



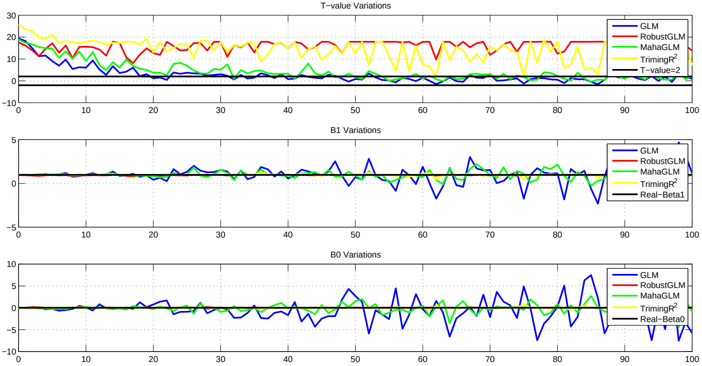
- **Normal**(μ, σ^2)**, both varying from 1 to 100.**

Significant variable and Contaminating a 10%



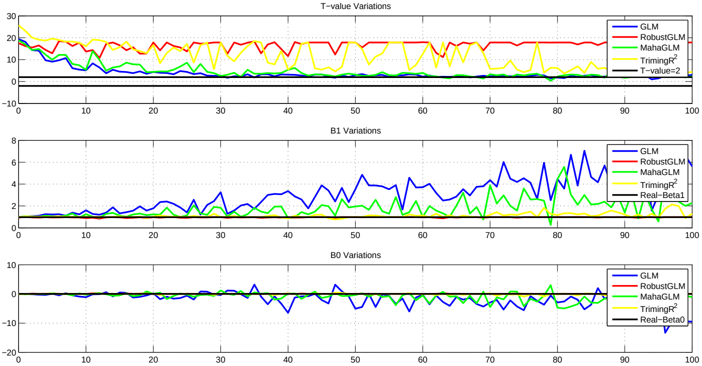
- **Normal$(\mu, 1)$, μ varying from 1 to 100.**

Significant variable and Contaminating a 10%



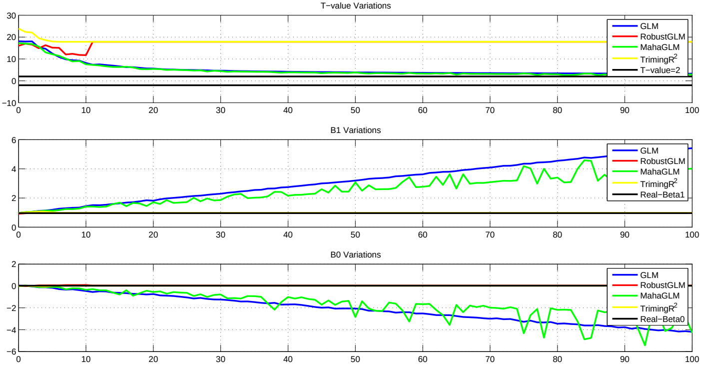
- **Normal$(0, \sigma^2)$, σ^2 varying from 1 to 100.**
- **X-axis represents σ^2 values varying from 1 to 100.**

Significant variable and Contaminating a 10%



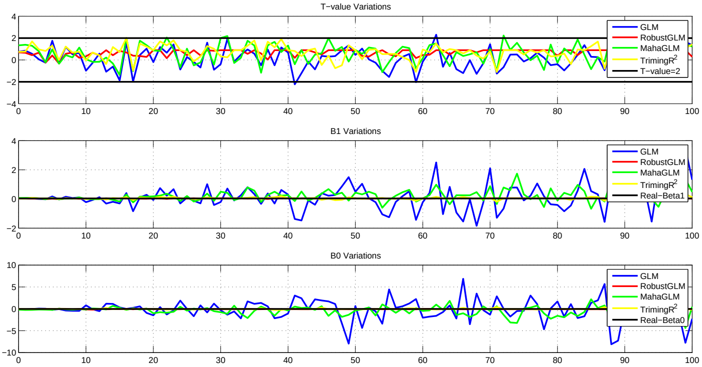
- **Normal**(μ, σ^2), **both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \sigma^2]$ **values both varying from 1 to 100.**

Significant variable and Contaminating a 10%



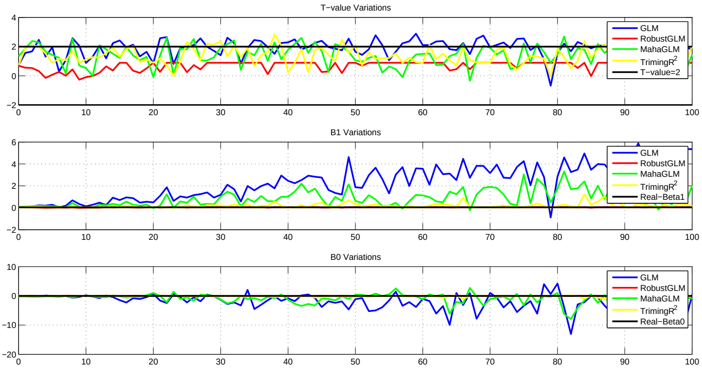
- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

No Significant variable and Contaminating a 10%



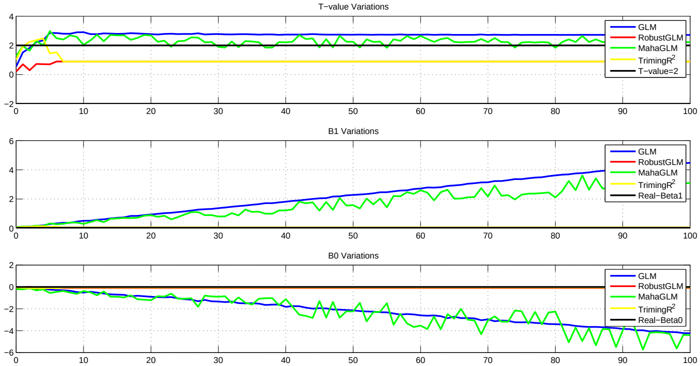
- **Normal$(0, \sigma^2)$, σ^2 varying from 1 to 100.**
- **X-axis represents σ^2 values varying from 1 to 100.**

No Significant variable and Contaminating a 10%



- **Normal**(μ, σ^2)**, both varying from 1 to 100.**
- X**-axis represents** $[\mu; \ \sigma^2]$ **values both varying from 1 to 100.**

No Significant variable and Contaminating a 10%



- **Normal$(\mu, 1)$, μ varying from 1 to 100.**
- **X-axis represents μ values varying from 1 to 100.**

Non-parametric estimation for Σ

Pearson correlation coefficient

The Pearson correlation matrix is related to the covariance matrix by

$$\rho_p(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \tag{1}$$

Our methodology basically changes in (1) the Pearson correlation coefficient $\rho_p(i,j)$ by a non parametric correlation coefficient as:

Non-parametric estimation for Σ

Pearson correlation coefficient

The Pearson correlation matrix is related to the covariance matrix by

$$\rho_p(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \quad (1)$$

Our methodology basically changes in (1) the Pearson correlation coefficient $\rho_p(i,j)$ by a non parametric correlation coefficient as:

- **Kendall** $\rho_k(i,j)$
- **Spearman** $\rho_s(i,j)$

Non-parametric estimation for Σ

Pearson correlation coefficient

The Pearson correlation matrix is related to the covariance matrix by

$$\rho_p(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \quad (1)$$

Our methodology basically changes in (1) the Pearson correlation coefficient $\rho_p(i,j)$ by a non parametric correlation coefficient as:

- **Kendall** $\rho_k(i,j)$
- **Spearman** $\rho_s(i,j)$
- **Robust version** $\rho_r(i,j)$

Kendall

$$\rho_k(i,j) = \frac{2 * (N_c - N_d)}{N(N-1)}$$

Spearman

$$\rho_s(i,j) = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Robust version

$$\rho_r(i,j) = \frac{C_r(i,j)}{\sqrt{C_r(i,i) C_r(j,j)}}$$

where,

$$C_r(i,j) = \text{median}\{(x_i - \text{median}(x_i))(x_i - \text{median}(x_i))\}$$

Non-parametric estimation for Σ

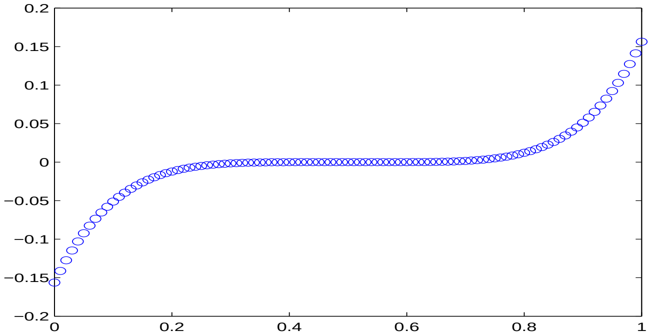


Figure: $\rho_p = 0.81, \quad \rho_k = \rho_s = \rho_r = 1$

Non-parametric estimation for Σ

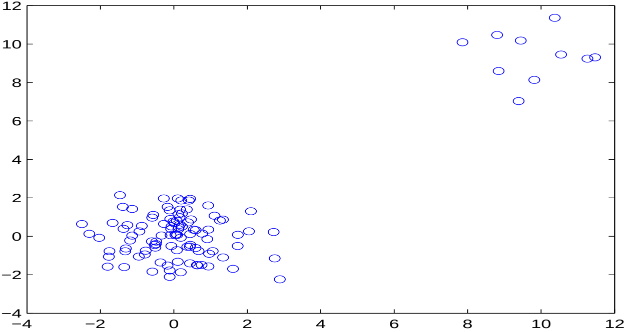


Figure: $\rho_p = 0.86, \quad \rho_k = 0.15, \quad \rho_s = 0.23, \quad \rho_r = 0.005$

Non-parametric estimation for Σ

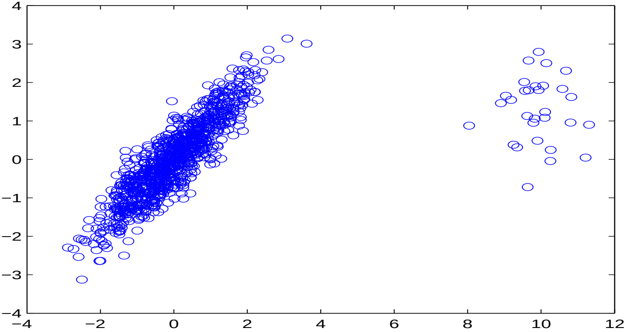


Figure: $\rho_p = 0.61, \quad \rho_k = 0.77, \quad \rho_s = 0.89, \quad \rho_r = 0.83$

Non-parametric estimation for Σ

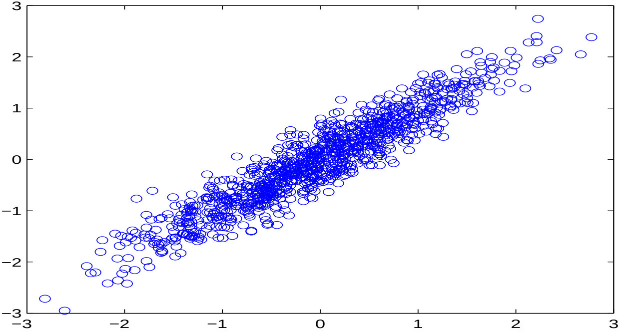


Figure: $\rho_p = 0.94, \quad \rho_k = 0.85, \quad \rho_s = 0.93, \quad \rho_r = 0.92$

Conclusions

- **We have introduced the fMRI problem and the importance of using robust statistical techniques to study the brain**

Conclusions

- **We have introduced the fMRI problem and the importance of using robust statistical techniques to study the brain**
- **A fast review of different approaches to estimate the covariance matrix has been made**

Conclusions

- **We have introduced the fMRI problem and the importance of using robust statistical techniques to study the brain**
- **A fast review of different approaches to estimate the covariance matrix has been made**
- **The GLM is quite sensitive in the presence of outliers**

Conclusions

- **We have introduced the fMRI problem and the importance of using robust statistical techniques to study the brain**
- **A fast review of different approaches to estimate the covariance matrix has been made**
- **The GLM is quite sensitive in the presence of outliers**
- **The fMRI based on GLM has very hight false discovery rate.**

Conclusions

- **We have introduced the fMRI problem and the importance of using robust statistical techniques to study the brain**
- **A fast review of different approaches to estimate the covariance matrix has been made**
- **The GLM is quite sensitive in the presence of outliers**
- **The fMRI based on GLM has very hight false discovery rate.**
- **The estimations introduced here improve the performance of GLM**

Research lines

- **Find good estimations for the variance and covariance matrix and its properties**

Research lines

- **Find good estimations for the variance and covariance matrix and its properties**
- **Implement the estimations to the fMRI problem**

Research lines

- **Find good estimations for the variance and covariance matrix and its properties**
- **Implement the estimations to the fMRI problem**
- **Find some criterium to know the optimal cut-off in the robust estimations**

Research lines

- **Find good estimations for the variance and covariance matrix and its properties**
- **Implement the estimations to the fMRI problem**
- **Find some criterium to know the optimal cut-off in the robust estimations**
- **To design estimators with a good computational performance in high dimensional data and big data**