Review of the article *Robust mixture of experts modeling using the t distribution* by F. Chamroukhi

Victoria BRAMI - Margot COSSON

Master Mathématiques Vision Apprentissage

victoria.brami@eleves.enpc.fr - margot.cosson@eleves.enpc.fr

November 29, 2022

Introduction

Mixture of experts is a statistical tool for modeling heterogeneity in data.

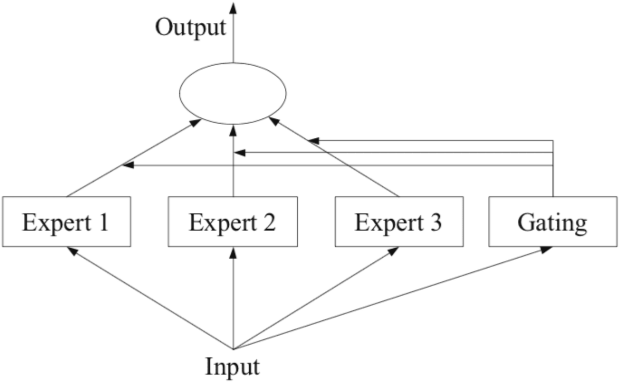


Figure: Mixture of Experts

Introduction

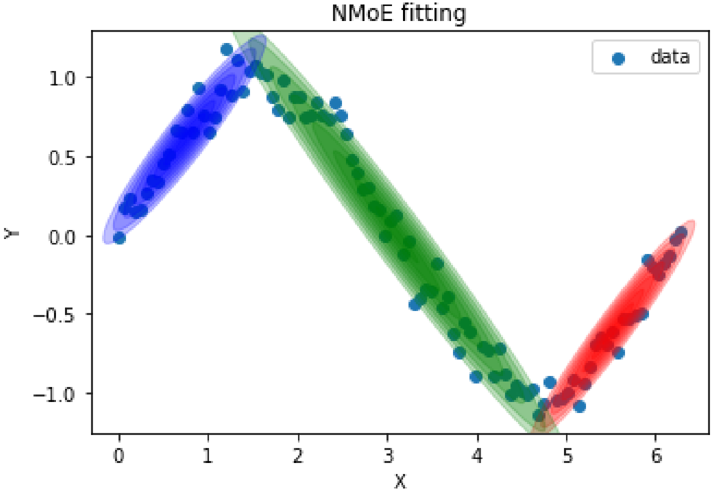
Mixture of experts is a statistical tool for modeling heterogeneity in data.

$$f(y|x;\psi) = \sum_{k=1}^{K} \pi_k(r;\alpha) f_k(y|x;\psi_k) \tag{1}$$

with

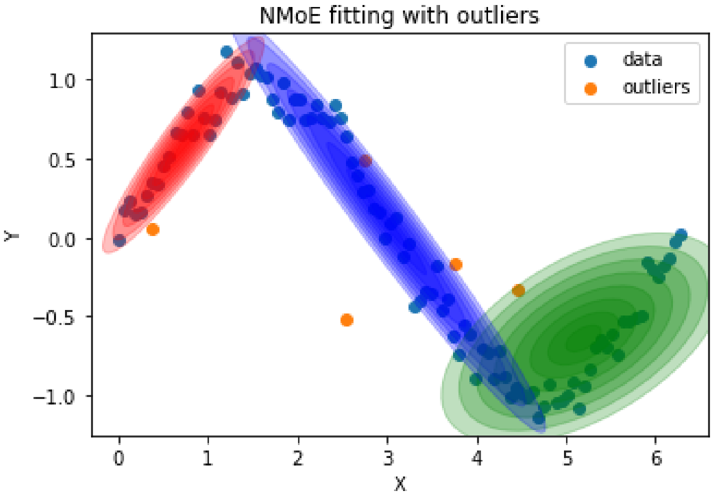
- π_k, the gating functions
- f_k, the expert functions
- x and r, the inputs
- α and ψ_k, the parameters to learn

The most used MoE is the Normal MoE which is a mixture of gaussian experts.



Introduction

However, while efficient in most cases, NMoE is not robust to outliers and heavy tailed data distribution.



Introduction

Question

Does it exist a MoE based on another distribution which fits data as well as NMoE and is more robust to outliers and heavy tailed data ?

Author's answer

Yes, the TMoE which is a mixture of experts based on the t-distribution.

Overview

1. Theoretical foundations
   - TMoE: Mixture of experts based on t distribution
   - Training method of TMoE: E(C)M algorithm

2. Critical analysis
   - Reinforce tests on simulated data
     - Reproduce results
     - More complex outliers
   - On climatic data

Table of Contents

The t distribution

The p.d.f of a t distribution $t_\nu(\mu, \sigma^2, \nu)$ writes:

$$f(y; \mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{d_y^2}{\nu})^{-\frac{\nu+1}{2}} \tag{2}$$

with

- $\mu \in \mathbb{R}$ the location parameter.
- $\sigma^2 \in \mathbb{R}_+$ the scale parameter.
- $\nu \in \mathbb{R}_+$ the degrees of freedom.
- $d_y^2 = (\frac{y-\mu}{\sigma})^2$ the squared mahalanobis distance.

The TMoE model

We define a K-component TMoE model by:

$$f(y|r, x, \Psi) = \sum_{k=1}^{K} \pi_k(r; \alpha) t_{\nu_k}(y, \mu(x, \beta_k), \sigma_k^2, \nu_k) \tag{3}$$

Notice

A t-expert component approaches a normal expert when $\nu_k \to \infty$.
Thus a TMoE model approaches an NMoE model.

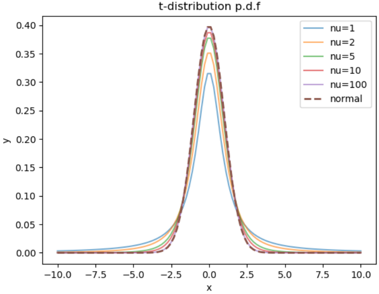


Figure: Probability density function of a t-distribution

Representation of the TMoE model

For each sample i, the hidden class label Z_i follows the multinomial distribution:

$$\mathbf{Z}_i|\mathbf{r}_i \sim \mathrm{Mult}(1, \pi_1(\mathbf{r}_i, \alpha), \ldots, \pi_K(\mathbf{r}_i, \alpha))$$

If a sample Y_i "belongs" to class k then we let $Z_{ik} = 1$ and, from the characterization of the $t-$distribution:

$$Y_i = \mu(\mathbf{x}_i, \beta_k) + \sigma_k \frac{E_i}{\sqrt{W_i}}$$

with:

- $E \sim \mathcal{N}(0, 1)$.
- $W \sim \Gamma(\frac{\nu_k}{2}, \frac{\nu_k}{2})$

Representation of the TMoE model: Advantages

Identifiability of the TMoE model

A TMoE model is **identifiable**: if the following hypothesis hold:

- ordered: $(\beta_1, \sigma_1^2, \nu_1) \prec \ldots (\beta_K, \sigma_K^2, \nu_K)$.
- initialized: α_K is set to 0 at initialization.
- irreductibility: $i \neq j \implies (\beta_i, \sigma_i^2, \nu_i) \neq (\beta_j, \sigma_j^2, \nu_j)$.

Then

$$f(y|\mathbf{x}, \mathbf{r}, \beta^j, (\sigma^2)^j, \nu^j) = f(y|\mathbf{x}, \mathbf{r}, \beta^k, (\sigma^2)^k, \nu^k)$$
$$\implies \beta^j, (\sigma^2)^j, \nu^j) = (\beta^k, (\sigma^2)^k, \nu^k)$$

This ensures the uniqueness of the solution when fitting a TMoE on data.

Maximum Likelihood Estimation

We are looking for the maximum likelihood estimators of the TMoE model. The observed-data log-likelihood writes :

$$log\ L(\psi) = \sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k(r_i; \alpha) t(y_i; \mu(x_i; \beta_k), \sigma_k^2, \nu_k) \tag{4}$$

Maximization of this function cannot be done directly so we use Expectation-Maximization algorithm which alternes between:

- **Step E**: computes the conditional expectation of complete-data log-likelihood (Q-function);
- **Step M**: maximizes it.

E step

The Q function writes:

$$Q(\psi;\psi^{(m)}) = Q_1(\alpha;\psi^{(m)}) + \sum_{k=1}^{K}[Q_2(\theta_k;\psi^{(m)}) + Q_3(\nu_k;\psi^{(m)})] \quad (5)$$

with

- $Q_1(\alpha;\psi^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m)} log\pi_k(r_i;\alpha)$
- $Q_2(\theta_k;\psi^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}[-\frac{1}{2}log(2\pi) - \frac{1}{2}log(\sigma_k^2) - \frac{1}{2}w_{ik}^{(m)}d_{ik}^2]$
- $Q_3(\nu_k;\psi^{(m)}) = \sum_{i=1}^{n}\tau_{ik}^{(m)}[-log\Gamma(\frac{\nu_k}{2})+(\frac{\nu_k}{2})log(\frac{\nu_k}{2})-(\frac{\nu_k}{2})w_{ik}^{(m)}+(\frac{\nu_k}{2}-1)e_{1,ik}^{(m)}]$
- $\tau_{ik}^{(m)} = \mathbb{E}_{\psi^{(m)}}[Z_{ik}|y_i,x_i,r_i]$
- $w_{ik}^{(m)} = \mathbb{E}_{\psi^{(m)}}[W_i|y_i,Z_{ik}=1,x_i,r_i]$
- $e_{1,ik}^{(m)} = \mathbb{E}_{\psi^{(m)}}[log(W_i)|y_i,Z_{ik}=1,x_i,r_i]$

M step

Due to its form, the Q function can be maximized independently with respect to α (1) and to each θ_k (2_k) and ν_k (3_k).

- (1) $\max Q_1(\alpha; \psi^{(m)})$: no closed form $\Rightarrow$ Iteratively Reweighted Least Squares algorithm
- (2_k) $\max Q_2(\theta_k; \psi^{(m)})$: closed form solution for β_k and σ_k^2
- (3_k) $\max Q_3(\nu_k; \psi^{(m)})$: equation form solution $\Rightarrow$ root finding algorithm

Expectation-Conditional-Maximization (ECM) algorithm

Adding an additional E-step between M-steps 2_k and 3_k such that the algorithm iteratively computes:

- E1-step: $Q_1(\alpha; \psi^{(m)})$ and $Q_2(\theta_k; \psi^{(m)})$
- M1-step: $\alpha^{(m+1)} = \max Q_1(\alpha; \psi^{(m)})$ and $\theta^{(m+1)} = \max Q_2(\theta_k; \psi^{(m)})$
- E2-step: $Q_3(\nu_k; \alpha^{(m)}, \theta_{\mathbf{k}}^{(\mathbf{m+1})}, \nu_k^{(m)})$ with updated parameter $\theta_k^{(m+1)}$
- M2-step: $\max Q_3(\nu_k; \alpha^{(m)}, \theta_{\mathbf{k}}^{(\mathbf{m+1})}, \nu_k^{(m)})$

Convergence properties of EM and ECM

- Stable convergence: at each iteration, the likelihood is increased.
- Converges towards a stationary point of the observed-data log-likelihood if the sequence $\{L_{\text{obs}}(\psi^{(k)} | Y_{\text{obs}}), k \geq 0\}$ is bounded above.

Pros and cons

Pros of TMoE

- accurate model for regression, clustering, density estimation
- relatively low computation times
- robust to outliers unlike NMoE
- accurate to fit heavy tailed distribution unlike NMoE

Cons of TMoE

- EM training algorithm: highly depends from initial point
- empirical number of experts (however we could rely on AIC, BIC, ICL)
- robustness to outliers is also ensured by Laplace MoE in most cases
- computation times are higher than NMoE ones

Table of Contents

Computation cost comparison

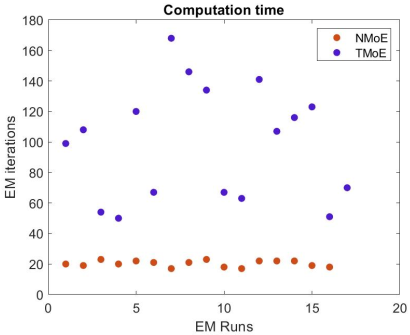


Figure: EM computation time (standard simulated data $n = 500$)

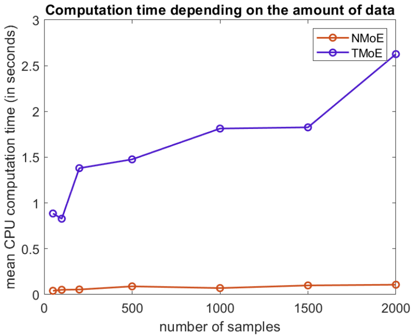


Figure: Mean EM computation time depending on the number of samples

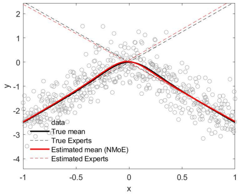
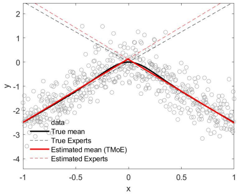
Conclusion: EM algorithm converges much faster when it intends to fit an NMoE distribution on the data

Reproduction of results

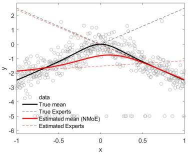
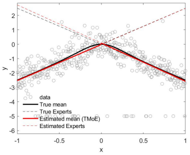
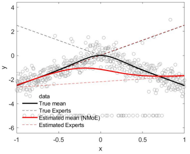
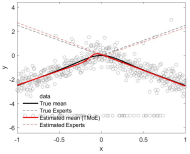
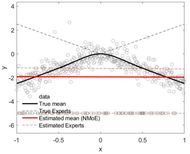
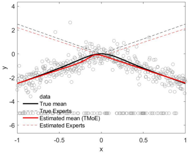
First dataset: **Synthetic simulated dataset**.
The article evaluates ECM:

1. On perfect data.
2. On data containing constant ordinates outliers (up to 5%).

When fitting perfect data, NMoE fitting is slightly better:

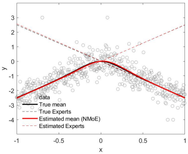
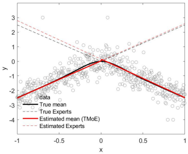
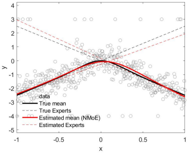
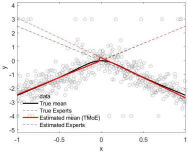
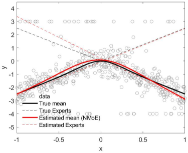
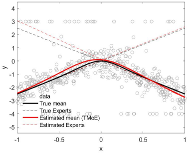
Outliers	NMoE fit	TMoE fit
0% (perfect data)		

Reproduction of results - Constant Outliers

Outliers	NMoE fit	TMoE fit
5%		
10%		
20%		

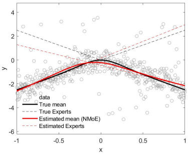
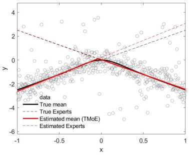
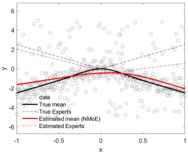
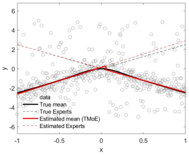
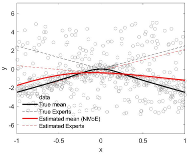
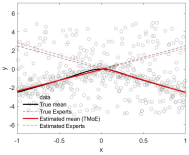
Observation: TMoE is **much more robust** than NMoE when fitting data containing constant-ordinates outliers.

Reproduction of results - Constant Outliers

Outliers	NMoE fit	TMoE fit
1%		
5%		
10%		

Critic: Robustness analysis limited to the study of constant outliers. For other similarly arbitrary chosen outliers, both models **are equivalent.**

Tests on random anomalies

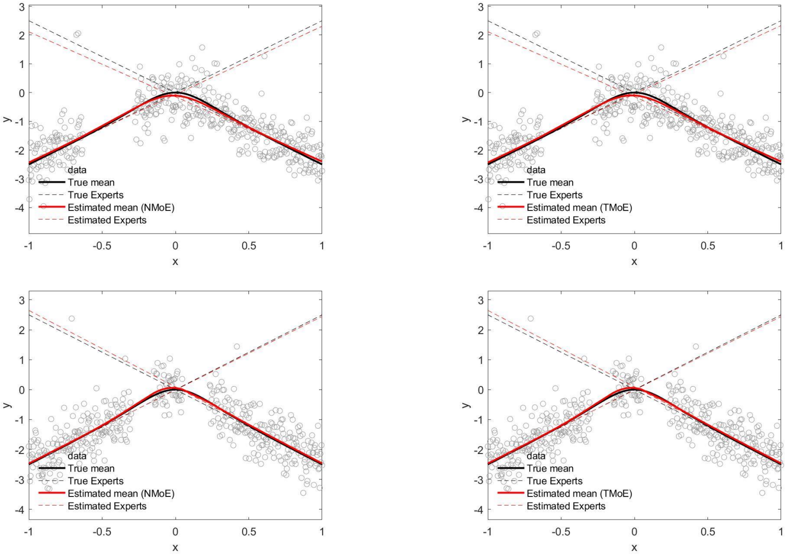
Outliers	NMoE fit	TMoE fit
10%		
30%		
49%		

Outliers drawn uniformly in the interval [-5, 5] (more similar to noise effects).

Conclusion

tMoE very robust to random outliers: almost perfect fitting with 49% of outliers.

Tests on incomplete data



Temperature anomalies dataset

Bayesian inference for environmental studies: the example of the global warming

- Nasa GISS Surface Temperature Dataset[1].
- Temperature anomalies: how much warmer or colder it is than normal. for a particular place and time.
- For this dataset, the baseline is the mean over the period 1951-1980
- Example: in 2000, the averaged global warming was + 0.55 °Ccompared to the temperatures of 1951-1980.

[1] `https://data.giss.nasa.gov/gistemp/`

Global warming model

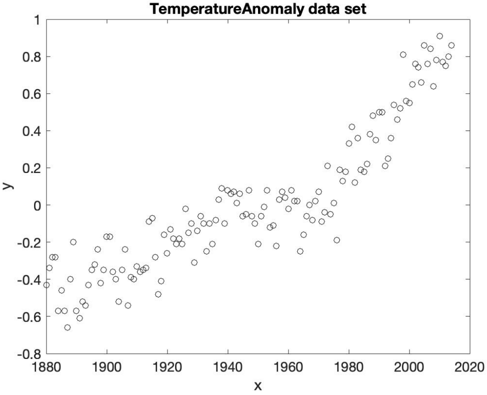


Figure: Yearly temperature anomalies from 1880-2014

Goal: Derive clusters to highlight the different period og global warming.

Specificity of data: natural noise due to fluctuations of global temperatures by natural change of ocean current, solar power and volcanic activity.

IPCC statement

IPCC Fifth Assessment Report, The Physical Science Basis, p. 193

Since 1901 almost the whole globe has experienced surface warming. Warming has not been linear; **most warming occurred in two periods: around 1900 to around 1940 and around 1970 onwards.**

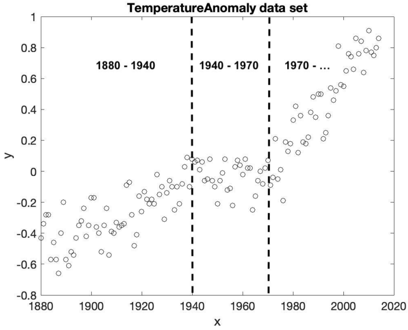


Figure: Global warming periods

Reproduction of the results

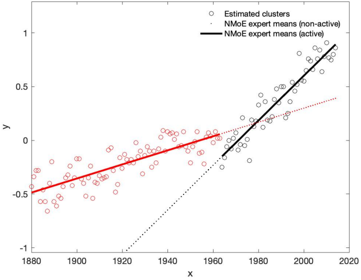


Figure: NMoE, K=2, periods: 1880-1964; 1964 - 2014

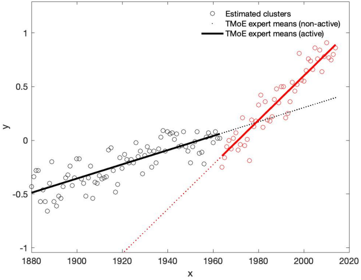


Figure: TMoE, K=2, periods: 1880-1963; 1963 - 2014

$\Rightarrow$ Same result for NMoE and TMoE ($K = 2$ is also the value corresponding to highest value of BIC) but it does not fit with IPCC analysis.

More realistic yearly model

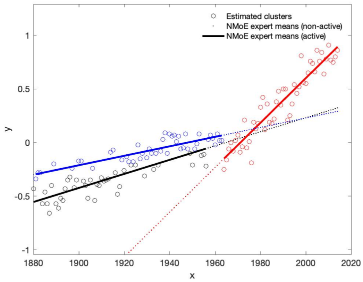


Figure: NMoE, K=3, periods: ? ; 1963 - 2014

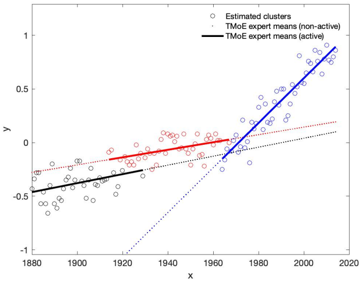


Figure: TMoE, K=3, periods: 1880-1920 ; 1920 - 1966 ; 1966 - 2014

$\Rightarrow$ NMoE gives non-sens results in term of physics while TMoE gives 3 periods but they overlap and they do not match IPCC analysis.

Monthly temperature anomalies model

⇒ We try with monthly temperature anomalies.

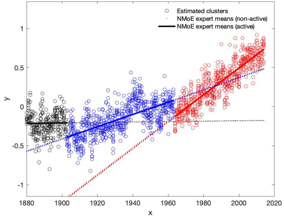


Figure: NMoE, K=3, periods: 1880-1902 ; 1902 - 1963 ; 1963 - 2014

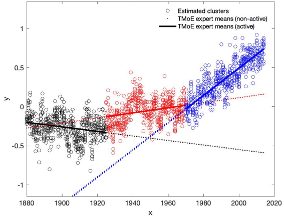


Figure: TMoE, K=3, periods: 1880-1925 ; 1925 - 1970 ; 1970 - 2014

⇒ Clusters of the TMoE fit the expected periods of the IPCC while NMoE does not give relevant ones.

Choice of period's number

Without any prior on the number of period, how many clusters would be chosen ?

K =	1	2	3	4	5	6
NMoE	374.7	701.8	802.9	842.0	**872.0**	862.0
TMoE	370.3	627.5	752.5	**809.5**	805.0	807.0

Table: Bayesian Information Criterion for monthly temperature anomalies dataset

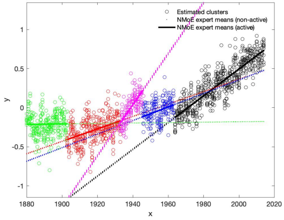


Figure: NMoE, K=5, periods: 1880-1903 ; 1903 - 1934 ; 1934 - 1946 ; 1946 - 1964 ; 1964 - 2014

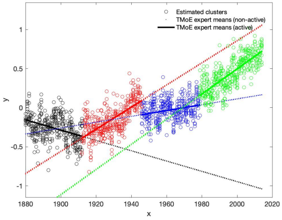


Figure: TMoE, K=4, periods: 1880-1912 ; 1912 - 1946 ; 1946 - 1978; 1978 - 2014

Questions

Questions

References

F. Chamroukhi (2016)
Robust mixture of experts modeling using the t distribution
Neural networks 79, 20 – 36. Code repository:
https://github.com/fchamroukhi/tMoE_m.

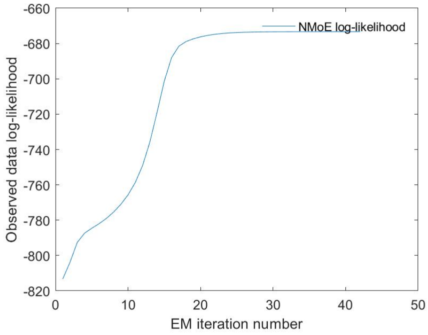
Hartmann, D. et al. (2013)
Climate Change 2013: The Physical Science Basis
IPCC report 2, 187 – 197, url: https:
//www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_all_final.pdf.

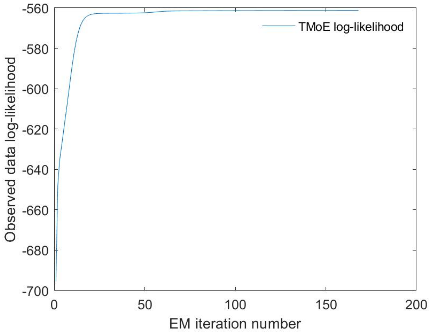
Meng, X.L., Rubin, D.B. (2013)
Maximum likelihood estimation via the ECM algorithm: A general framework
Biometrika 80.2, 267 – 278.

Annex - (Multi-cycle)-ECM

- EM : $L(\psi_{k+1}) \geq L(\psi_k)$ since $Q(\psi_{k+1};\psi_k) \geq Q(\psi_k;\psi_k)$.
- ECM : replace M steps by conditional M steps when the maximization over all paramaters at once is too complicated $\Rightarrow$ here it is naturally the case since the Q function is separable. So, $L(\psi_{k+1}) \geq L(\psi_k)$ because
  $Q(\psi_{k+1};\psi_k) \geq Q(\psi_{k+\frac{S-1}{S}};\psi_k) \geq Q(\psi_{k+\frac{S-2}{S}};\psi_k) \geq ... \geq Q(\psi_k;\psi_k)$.
- Multi-cycle ECM : add E steps between CM steps.
  $Q(\psi_{k+1};\psi_k) \geq Q(\psi_{k+\frac{S-1}{S}};\psi_k) \geq Q(\psi_{k+\frac{S-2}{S}};\psi_k) \geq ... \geq Q(\psi_k;\psi_k)$ may not hold since Q function is updated but $L(\psi_{k+1}) \geq L(\psi_k)$ is still ensured.

Annex - LogLikelihood on Simulated data





Annex - Climate change data

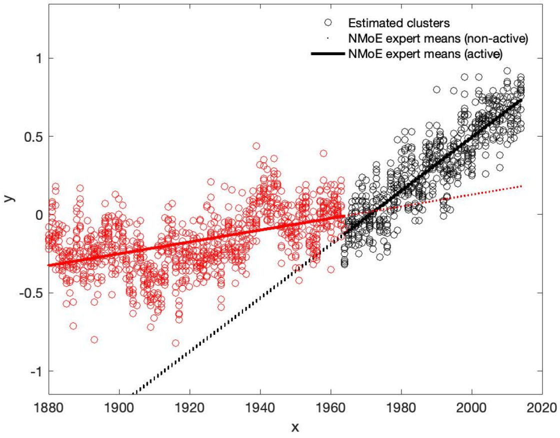


Figure: NMoE, K=2, periods: 1880-1964 ; 1964 - 2014

Annex - Climate change data

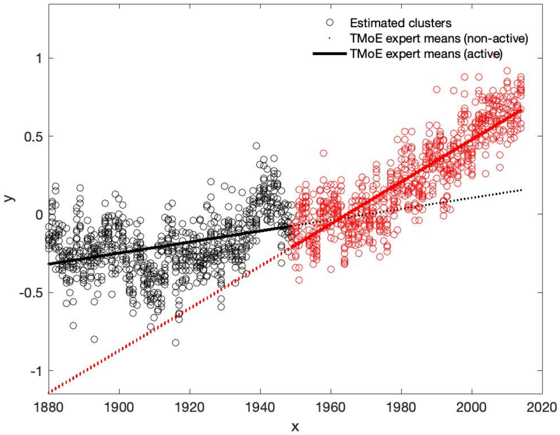


Figure: TMoE, K=2, periods: 1880-1949 ; 1949 - 2014

Annex - Climate change data

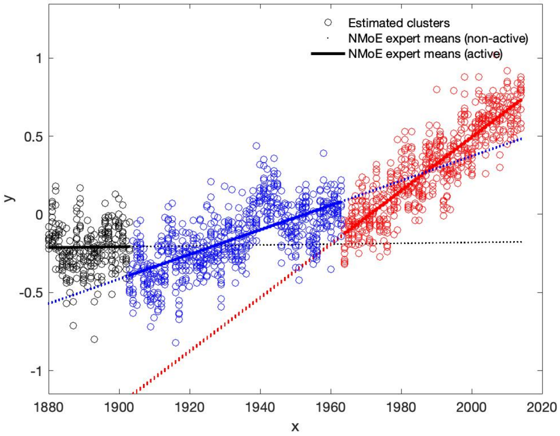


Figure: NMoE, K=3, periods: 1880-1902 ; 1902 - 1963 ; 1963 - 2014

Annex - Climate change data

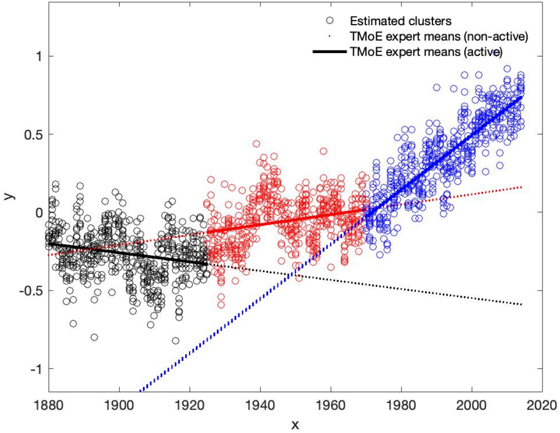


Figure: TMoE, K=3, periods: 1880-1925 ; 1925 - 1970 ; 1970 - 2014

Annex - Climate change data

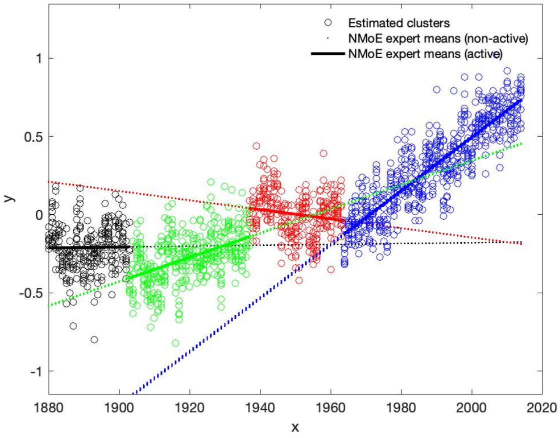


Figure: NMoE, K=4, periods: 1880-1902 ; 1902 - 1936 ; 1936 - 1963; 1963 - 2014

Annex - Climate change data

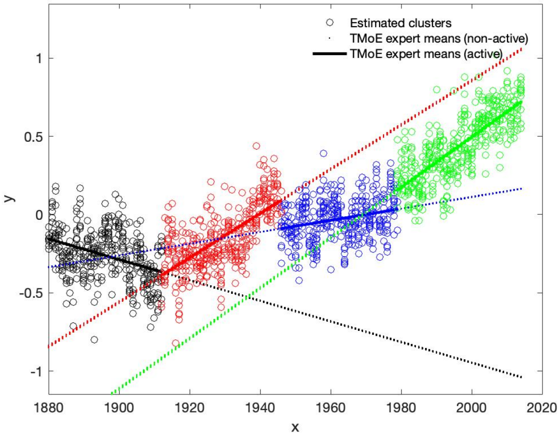


Figure: TMoE, K=4, periods: 1880-1912 ; 1912 - 1946 ; 1946 - 1978; 1978 - 2014

Annex - Climate change data

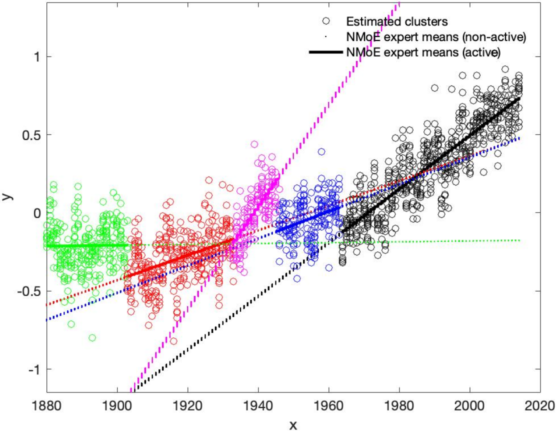


Figure: NMoE, K=5, periods: 1880-1903 ; 1903 - 1934 ; 1934 - 1946 ; 1946 - 1964 ; 1964 - 2014

Annex - Climate change data

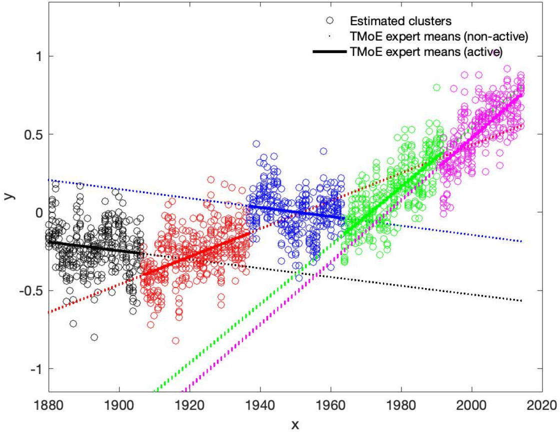


Figure: TMoE, K=5, periods: 1880-1906 ; 1906 - 1936 ; 1936 - 1964 ; 1964 - 1991 ; 1991 - 2014

Annex - Climate change data

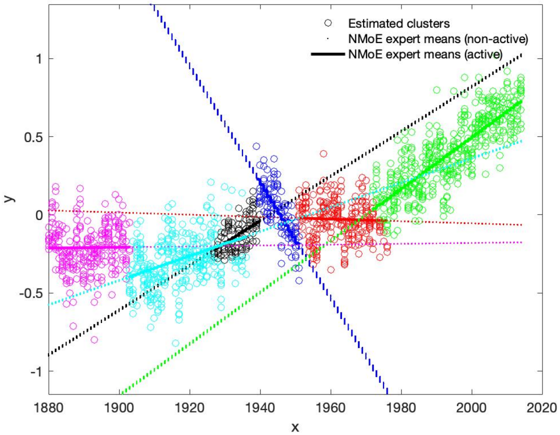


Figure: NMoE, K=6, periods: 1880-1903 ; 1903 - 1930 ; 1930 - 1940 ; 1940 - 1951 ; 1951 - 1973 ; 1973 - 2014

Annex - Climate change data

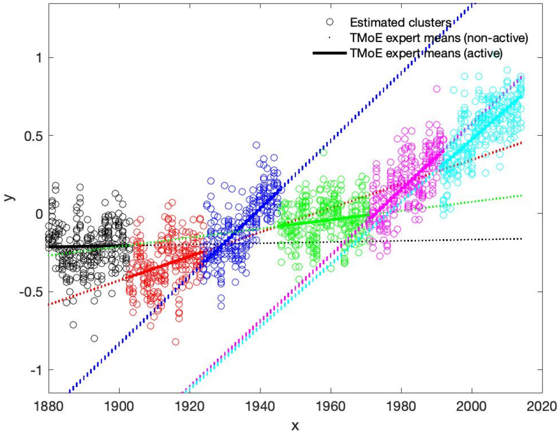


Figure: TMoE, K=6, periods: 1880-1903 ; 1903 - 1924 ; 1924 - 1946 ; 1946 - 1971 ; 1971 - 1992 ; 1992 - 2014