# A shared component hierarchical model to represent how fish assemblages vary as a function of river temperatures and flow regimes

Jeremy PIFFADY[1,2], Éric PARENT[1] & Yves SOUCHON[2]

[1]Équipe Modélisation, Risques, Statistique, Environnement de l'UMR 518 INRA/AgroParisTech, 19, Avenue du Maine, 75732 Paris Cedex 15, France,
[2]Institut de recherche en sciences et technologies pour l'environnement, Pôle d'Hydrobiologie des cours d'eau,3 bis Quai Chauveau, 69336 Lyon

14 Janvier 2011

# Summary

1. How interannual variations of fish assemblages are linked to temperature and flow regimes ?
   - The Bugey case study location
   - The response variables
   - The explanatory variables
2. Challenges to the statistical analyst
   - Challenging features
   - Why not a GLM ?
3. A shared component hierarchical model
   - A *cocktail* model structure
   - Inference
   - Results

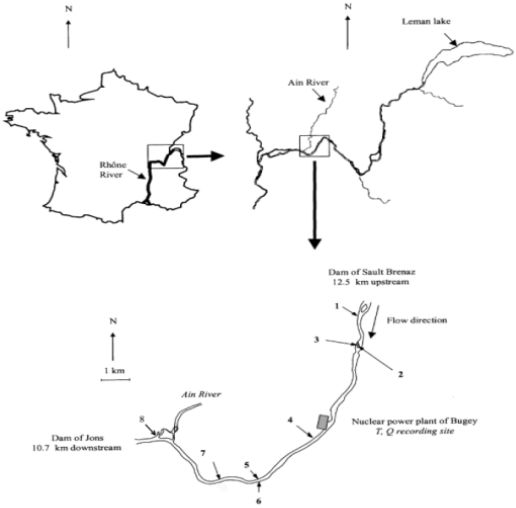
# Jeremy's story of a modelling challenge

$Y = f(X, \varepsilon)$

- $Y$ : ecosystem behavior
- $X$ : environmental variations of interest
- $\varepsilon$ : unknown perturbations, *noise*
- $f$ : ...functional form of the answer ...to be defined as well

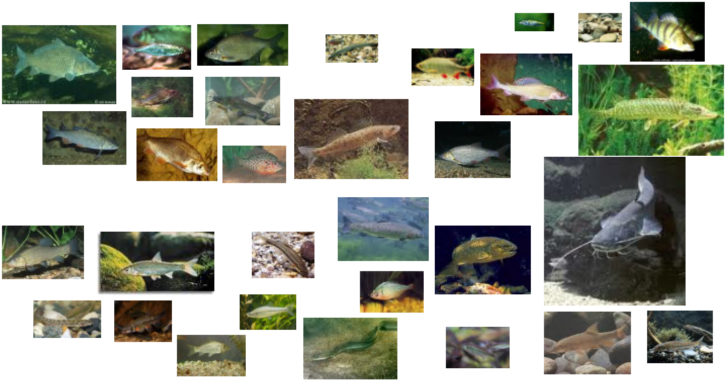
Application to three groupings of juveniles in the upper River Rhone during the 1980-2005 period. Let's find a statistician ! ? But we do have data, let's have a look...

# The Bugey site



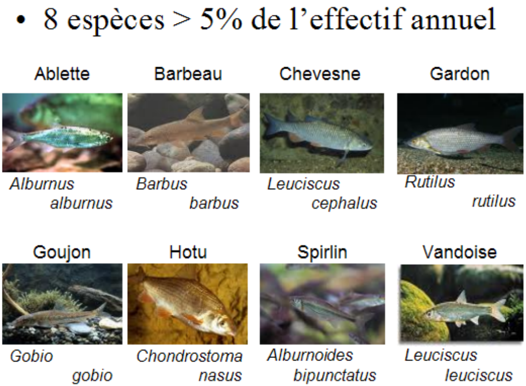
# The Bugey site cont'd

# Many fish species can be found

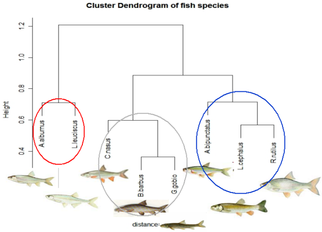


# But only 8 species are permanently caught

# These eight species are common fish

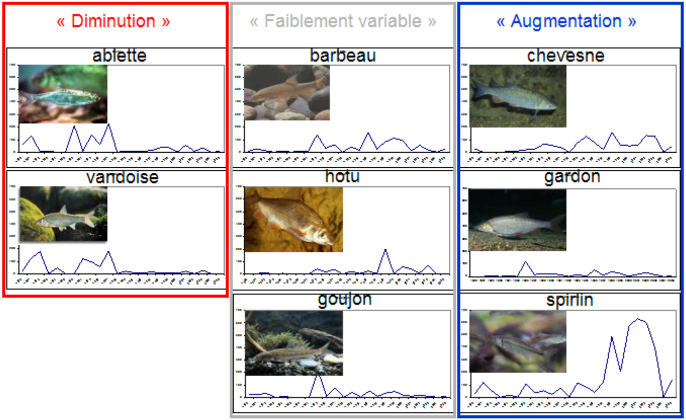


# They can be clustered in three groups

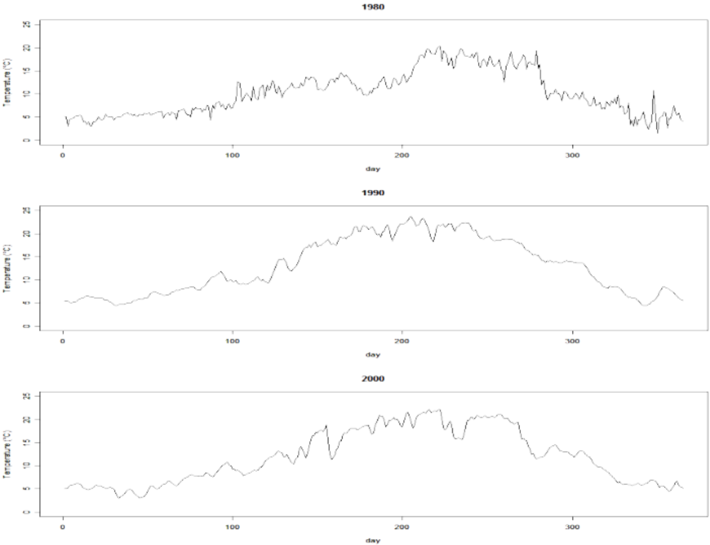


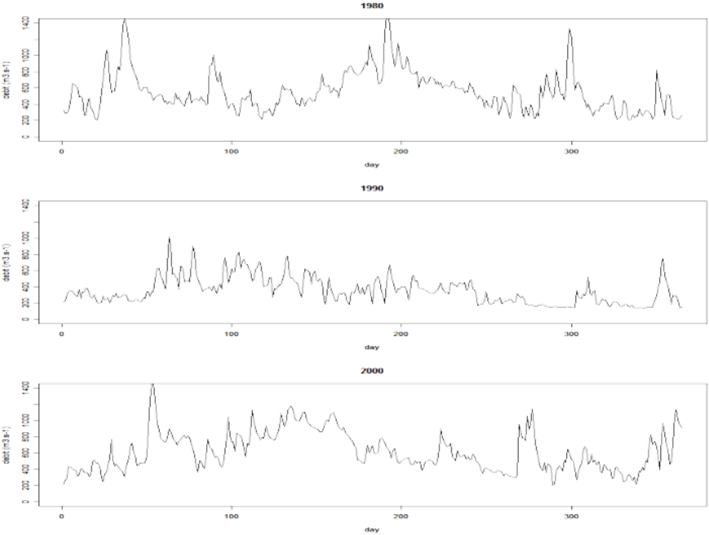

Only the 8 species representing more than 5% each of total abundance were analysed : bleak (Alburnus alburnus), barbel (Barbus barbus), chub (Leuciscus cephalus), roach (Rutilus rutilus), gudgeon (Gobio gobio), nase (Chondrostoma nasus), stream bleak (Alburnoides bipunctatus) and dace (Leuciscus leuciscus). Gp1={bleak and dace} (Cool water group) Gp2={gudgeon, barbel and nase} (Benthic group) Gp3={ stream bleak, roach and chub} (Thermophilic group)

# These three groups exhibits different time patterns

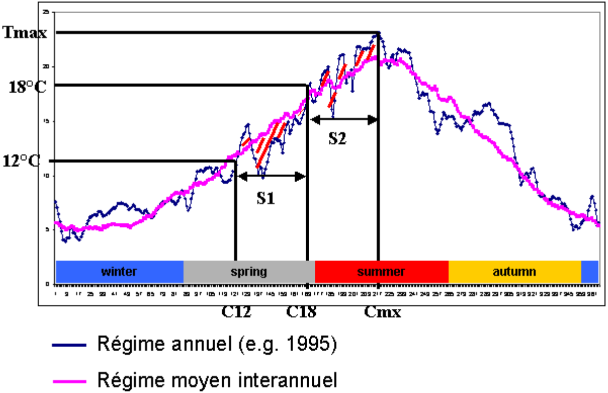


# Temperature rules fish activities(reproduction , etc.)

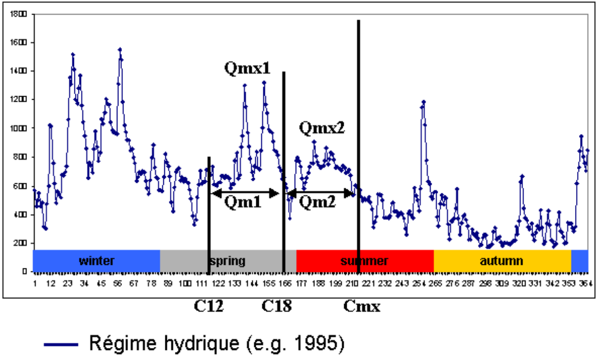



1980
0
100
200
300
day
1990
0
100
200
300
day
2000
0
100
200
300
day

# Biological knowledge is require to extract yearly significant quantities from the daily temperature signal



# Yearly patterns are extracted from flow regimes



# Nine possibly explanatory covariates are extracted and standardized indices are designed

| Covariate | mean | sd |
|---|---|---|
| C12 | $115.4$ | $11.4$ |
| C18 | $162.5$ | $13.9$ |
| Cmx | $214.6$ | $16.1$ |
| S1 | $-12.5$ | $88.0$ |
| S2 | $26.4$ | $71.5$ |
| Qm1 | $565.0$ | $176.2$ |
| Qmx1 | $920.3$ | $275.6$ |
| Qm2 | $580.9$ | $152.7$ |
| Qmx2 | $880.2$ | $221.1$ |

# Statistical challenges

1. The Bugey protocol versus traditionnal ecological hypotheses
   - Only one pass : no difference can be made between capturability and population size
   - Dynamic non linear models such as prey-predator with interactions cannot
   - The system is not closed. Emigration/immigration
   - The system is influenced by the nuclear plant warming the waters
2. The Bugey sampling protocol versus common statistical hypotheses
   - A poorly controlled experiment
   - Variables with different natures and different scales
3. Ambitious observational data study with much lack of contrast
   - Is there anything to see ? Abrupt changes ?
   - Are flows and temperatures the main drivers ? Do they vary enough ?
   - Are not the remaining fish the most adapted (less significant of a change) species ?

# Why not a GLM ?

Write for each group $s$ of species

$$Y_{s,t} \sim dPois(\lambda_{s,t})$$
$$\log(\lambda_{s,t}) = X_t \beta_s + \sigma\varepsilon_t$$

with

- $Y_s$ counts of species $s$ in experiment $t$
- $X_t$ values of the design matrix for experiment $t$
- $\beta_s$ coefficient characterizing answer to species $s$ to environmental variations
- $\varepsilon_t$ $N(0,1)$ overdispersion due to uncontrolled conditions of experiment $t$

Pb :

1. An additional model selection in search of influential variables
2. Poisson assumptions (same capturability,same fishing protocol)
3. Model selection to point out relevant explanatory variables may be tricky

# A cocktail model structure

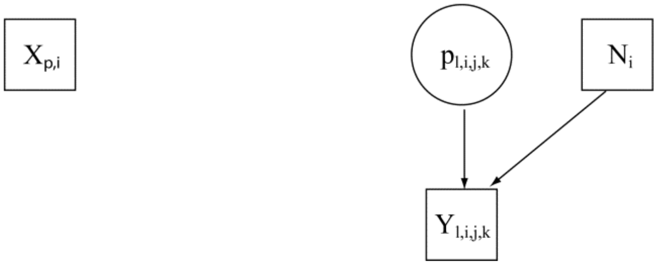
Objectives

- Get rid of the main unstationnarities in the sampling protocol
- i-e work conditionnaly to the total number of captures
- Explain the variations of the specific ratios of species
- Avoid model selection traps

Principles

- Join a multivariate analysis and a logistic regression model within a bayesian hierarchical structure
- A latent variable as a shared component : let's call it the *hypersignal* !
- Similar to a Partial Least Squareregression in the frequentist world

# A cocktail model structure



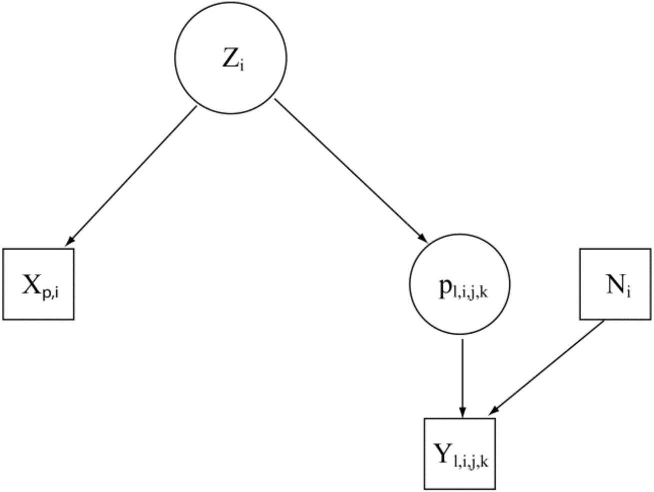
# A cocktail model structure

$$(Y_{1t}, Y_{2t}, Y_{3t}) \sim dmult(p_{1t}, p_{2t}, p_{3t}, N_t)$$

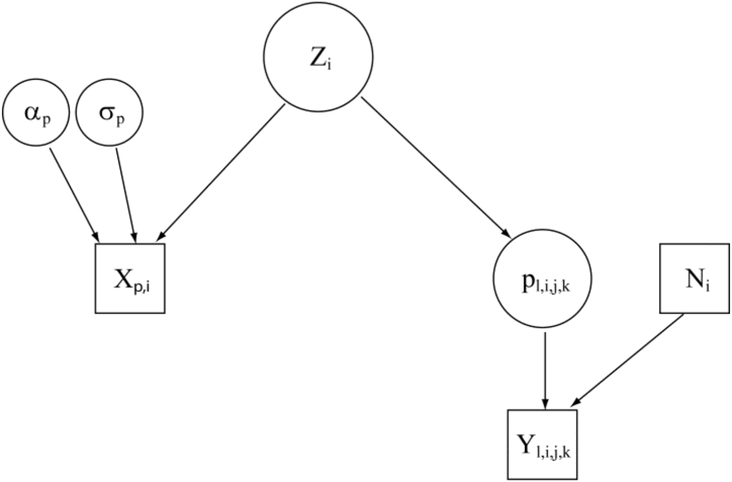
$$p_{1t}, p_{2t}, p_{3t} = f(X_t)$$

which $f$?

# A cocktail model structure

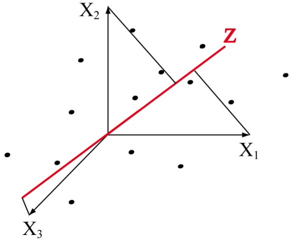


# A cocktail model structure

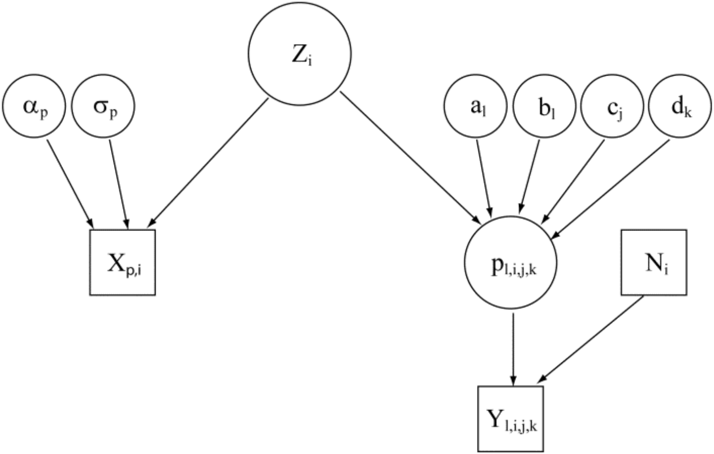


# A principal component-like analysis

$$\begin{cases} X_t^1 = \alpha_1 Z_t + \sigma_1 \varepsilon_{1t} \\ X_t^2 = \alpha_2 Z_t + \sigma_2 \varepsilon_{2t} \\ \quad\quad\quad \dots \\ X_t^9 = \alpha_9 Z_t + \sigma_9 \varepsilon_{9t} \end{cases}$$



# A cocktail model structure



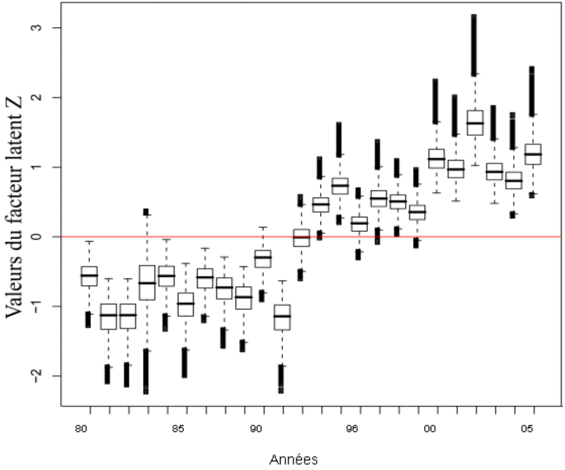
# A cocktail model structure

$$(Y_{1t}, Y_{2t}, Y_{3t}) \sim dmult(p_{1t}, p_{2t}, p_{3t}, N_t)$$

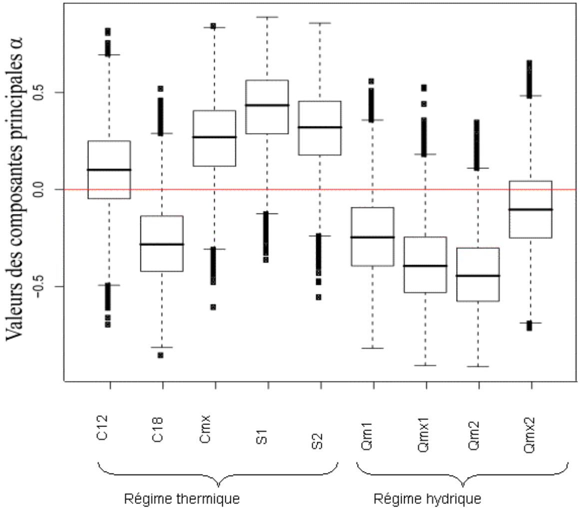
$$\left\{ \begin{array}{c} logit(p_{1,t}) = a_1 Z_t + b_1 + c_{1,site} + d_{1,season} \\ logit(p_{2,t}) = a_3 Z_t + b_3 + c_{3,site} + d_{3,season} \\ p_{1,t} + p_{2,t} + p_{3,t} = 1 \end{array} \right.$$

$$\left\{ \begin{array}{c} X_t^1 = \alpha_1 Z_t + \sigma_1 \varepsilon_{1,t} \\ X_t^2 = \alpha_2 Z_t + \sigma_2 \varepsilon_{2,t} \\ ... \\ X_t^9 = \alpha_9 Z_t + \sigma_9 \varepsilon_{9,t} \end{array} \right.$$

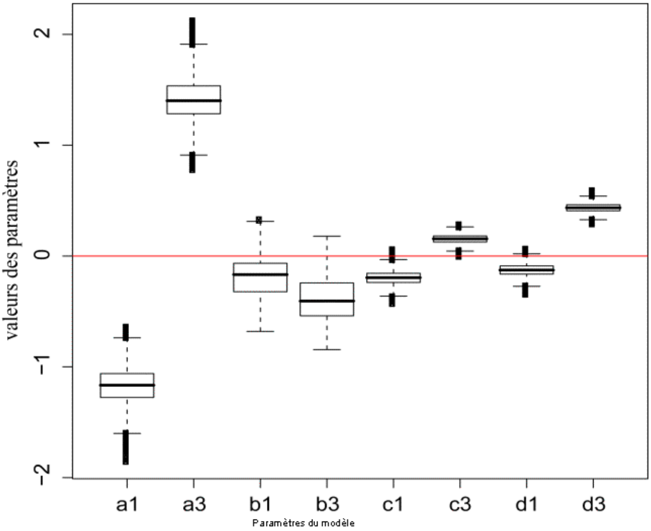
# Inference : the hypersignal versus time



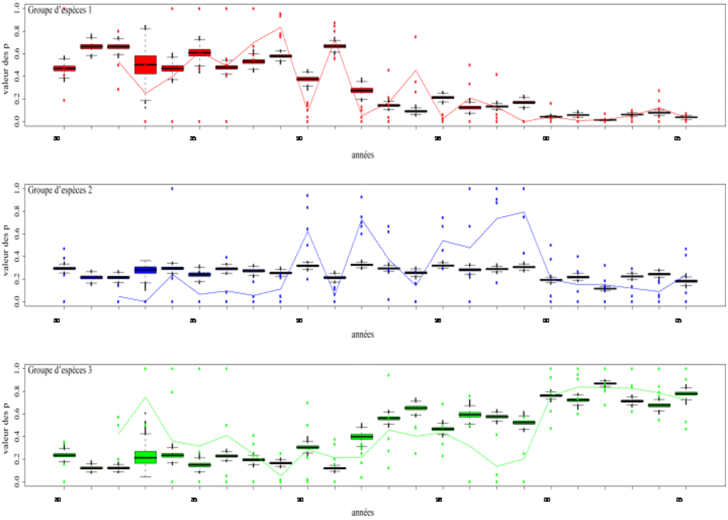
# Inference : trying to explain the hypersignal as a function of environmental covariates



# Inference : trying to understand the hypersignal as an explanation for species relative abundance



# Back to the data

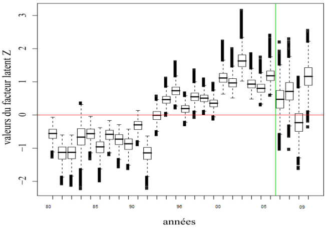


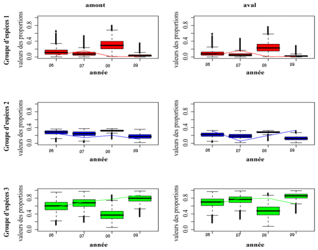
# The four last years were taken as a validation period in predictive mode

$$[Y^{new}|X^{new}, y^{old}] = \int_{\theta, Z} [Y^{new}|\theta, Z, X^{new}][\theta, Z|y^{old}, X^{new}]d\theta dZ$$

with $\theta = (\alpha, \sigma, a, b, c, d)$ and $p_{a,b,c,d}(Z)$ s.t.$logit(p) = aZ + b + c_{site} + d_{season}$

$$[Y^{new}|X^{new}, y^{old}] = \int_{\theta, Z} [Y^{new}|p_{a,b,c,d}(Z), N^{new}][Z, X^{new}, \alpha, \sigma][\theta|y^{old}]d\theta dZ$$

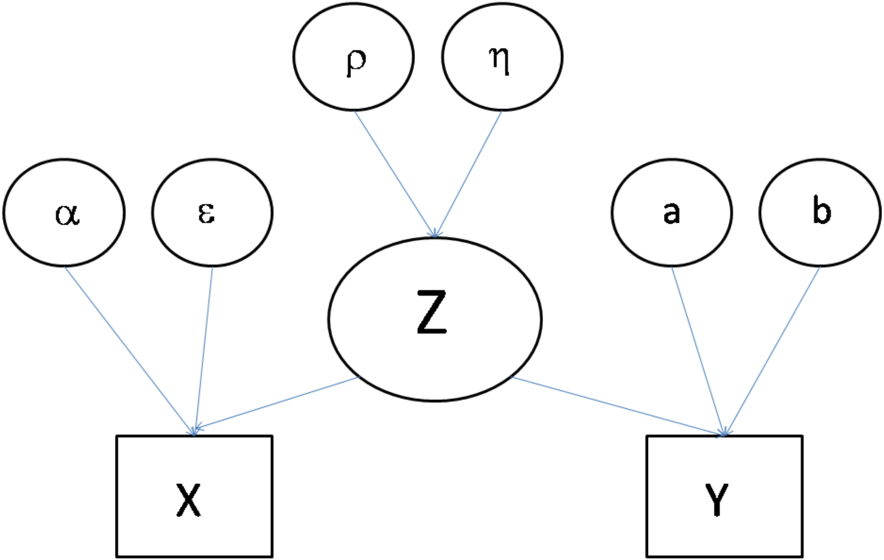


# Discussion

- The protocol variations are somehow stabilized
- An a priori structure might be assumed for the hypersignal : Hidden Markov Model, Shifting level Model, Spline...
- Is the numbering the groups of any relevance ?
- What would be the second principal component ?

# Discussion con'd



# Take home messages

- $Y = f(X, \varepsilon)$ easy mathematical formulation but hard to specify
- Observational data with poor control and few constrast
- Much interplay between data exploratory analysis and model design
- Take into account overdispersion, different natures between inputs & outputs, model choice
- None readymade toolbox solution, design the model of your own !

# Bibliographie


Parent, E. et Bernier, J. (2007).
*Le Raisonnement Bayésien : Modélisation et inférence.*
Springer France, Paris.

Boreux, JJ. , Parent, E. et Bernier, J. (2010).
*Pratique du calcul Bayésien* .
Springer France, Paris.

Hoff, P. (2009).
*A first course in Bayesian Statistical Methods.*
Springer .

Marin, J.M. et Robert, C.P. (2007).(chapitre 3)
*The Bayesian Core.*
Springer .