# penalized Latent Variable Models

**Brice Ozenne**, Esben Budtz-Jørgensen, Klaus Kähler Holst

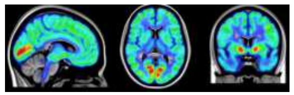
25-08-16 Compstat 2016

# Motivation: depression studies

Investigate factors driving depression and response to treatment
⇒ indirect and correlated measurements

- psychological tests
  → emotional face identification
  → verbal affective memory test

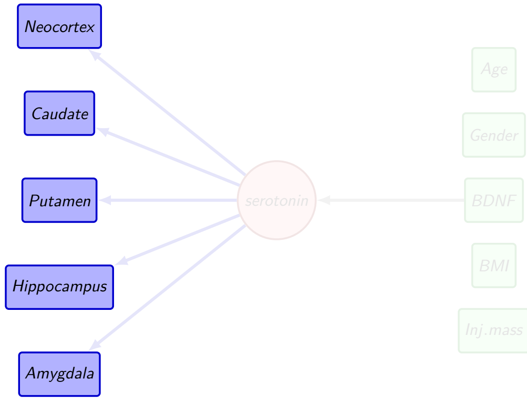
- serotonin level
  → PET[1] imaging
  → average regional value



- covariates, e.g. genetic factors

[1] Positron Emission Tomography

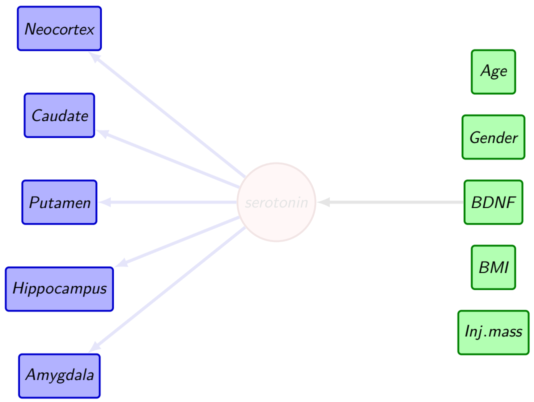
# Example of study - **Fisher2014**[2]




$Y$: regional PET measurements

[2] the model presented here is a simplified version of the published model

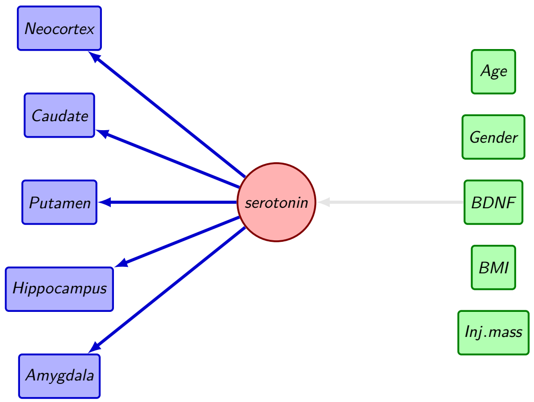
# Example of study - **Fisher2014**[2]




$X$: covariates

[2]the model presented here is a simplified version of the published model

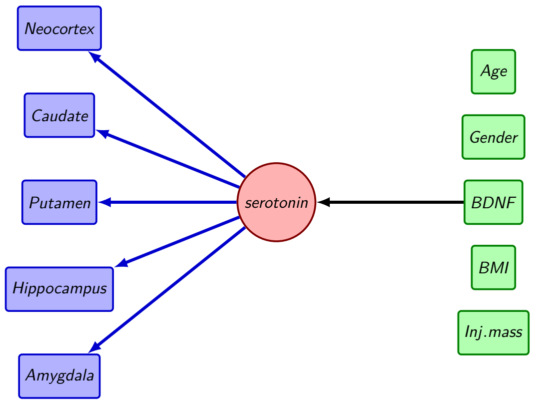
# Example of study - **Fisher2014**[2]




$\eta$: latent variable

[2]the model presented here is a simplified version of the published model

# Example of study - **Fisher2014**[2]



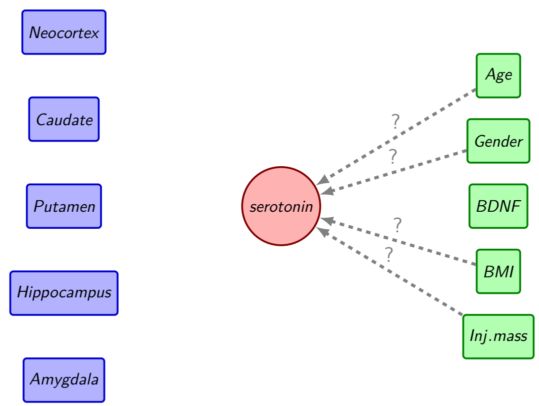

hypothesis to test

[2] the model presented here is a simplified version of the published model

# Challenges

Causal diagram only partially known:

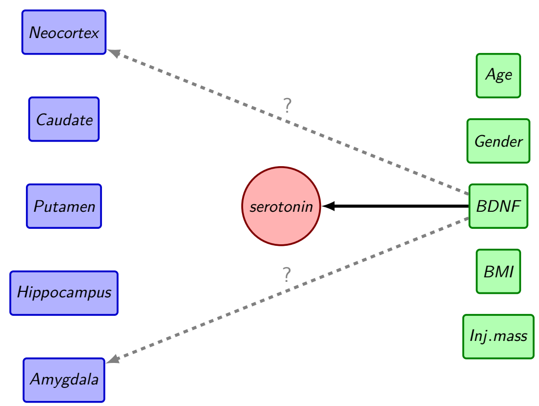
- relevant covariates
- regional specific effects



# Challenges

Causal diagram only partially known:

- relevant covariates
- regional specific effects



# Challenges

Causal diagram only partially known:

- relevant covariates
- regional specific effects

⇒ variable selection procedure

High-dimensional data:

- small samples (e.g. n=73 in **Fisher2014**)
- images
- large number of psychological tests

⇒ regularization

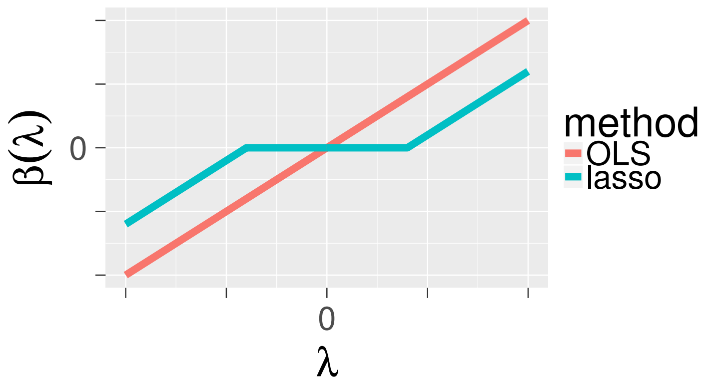
## Different types of regularization

Favours a small number of:

- parameters — `lasso`
  $\mathcal{P}(\Theta) = |\Theta|_1$ — (**Tibshirani1996**)
- group of parameters — `group lasso`
  $\mathcal{P}(\Theta) = \sum_{g=1}^{G} \sqrt{p^g} ||\Theta^{(g)}||_2$ — (**Friedman2010**)
- spatial patterns — `nuclear norm`
  $\mathcal{P}(\Theta) = |eigen(\Theta)|_1$ — (**Zhou2014b**)

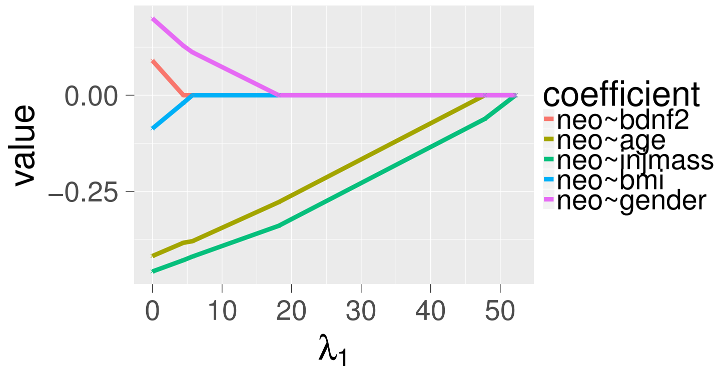
## Some properties of the Lasso regression

- Orthogonal design: $\hat{\beta}_j(\lambda) = sign(Z_j)(|Z_j| - \frac{\lambda}{2})_+$, $Z = X^\top Y$



# Some properties of the Lasso regression

- Orthogonal design: $\hat{\beta}_j(\lambda) = sign(Z_j)(|Z_j| - \frac{\lambda}{2})_+$, $Z = X^\top Y$
- $\beta_j$ piecewise linear with $\lambda$ (**Efron2004**)



## Some properties of the Lasso regression

- Orthogonal design: $\hat{\beta}_j(\lambda) = sign(Z_j)(|Z_j| - \frac{\lambda}{2})_+$, $Z = X^\top Y$
- $\beta_j$ piecewise linear with $\lambda$ (**Efron2004**)
- for suitable $\lambda$, $\mathbb{P}\left[\hat{S}(\lambda) = S_0\right] \xrightarrow[n\to\infty]{} 1$ (**Buhlmann2011**)
  $S_0$ true set of variables
  $\hat{S}$ selected set of variables using lasso

# Contribution

Integrate regularization into the LVM framework:

- estimation algorithm for $\Theta$
- method for choosing the appropriate $\lambda$

$\Theta = (\beta, \sigma, \rho)$: model parameters
$\lambda$: penalisation parameter

# LVM - Estimation

$$\underset{\Theta}{\text{argmin}}\left(\mathcal{L}(\Theta)\right)$$

where:

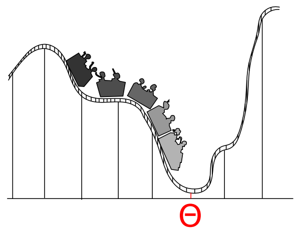
$$\begin{aligned}
\mathcal{L}(\Theta) &\propto \textstyle\sum_{i=1}^{n} \log(|\Sigma(\Theta)|) + (Y_i - \mu_i(\Theta))^{\top}\Sigma(\Theta)^{-1}(Y_i - \mu_i(\Theta)) \\
\mu_i(\Theta) &= \mathbb{E}\left[Y_i|X_i\right] \\
\Sigma(\Theta) &= \mathbb{V}ar\left[Y_i|X_i\right]
\end{aligned}$$

## LVM - Estimation

$$\underset{\Theta}{\text{argmin}} \left( \mathcal{L}(\Theta) \right)$$

Convex and differentiable likelihood:

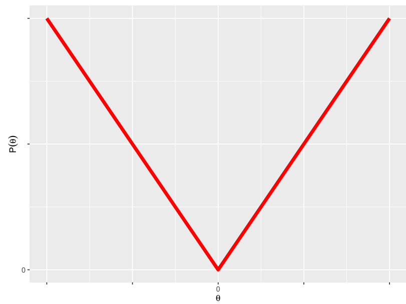
- gradient descent method
- $\Theta^i \leftarrow \Theta^{i-1} - \tau \nabla f(\Theta^{i-1})$
- quadratic convergence rate: $|\epsilon_{i+1}| < M\epsilon_i^2$



## Estimation - penalized LVM

$$f(\Theta) = \mathcal{L}(\Theta) + \lambda \mathcal{P}(\Theta)$$

Non differentiable penalties, e.g. lasso
$\Rightarrow$ cannot use gradient descent methods



# Proximal optimization

Proximal optimization: $f$ convex and differentiable
$g$ convex but not differentiable

- $x$ minimizes $f + g \Leftrightarrow x = prox_{\tau g}(x - \tau \nabla f(x))$

Proximal operator

- $prox_{\tau f} : \mathbb{R}^p \rightarrow \mathbb{R}^p$

$$x \mapsto \underset{v}{\operatorname{argmin}} \left( f(v) + \frac{1}{2\tau} \|v - x\|_2^2 \right)$$

e.g. $prox_{\lambda \|.\|_1}(x) = sign(x)(x - \lambda)^+$

# Proximal optimization

Proximal optimization: $f$ convex and differentiable
$g$ convex but not differentiable

- $x$ minimizes $f + g \Leftrightarrow x = prox_{\tau g}(x - \tau \nabla f(x))$

$\tau$ drives the convergence:

- $\tau \in ]0; \frac{1}{L}]$, $L$ Lipschitz constant of $\nabla f$
- small $\tau \equiv$ slow convergence

$\Rightarrow$ lower bound for $\frac{1}{L}$

# Proximal gradient algorithm

**while** $||f(\Theta^k) - f(\Theta^{k-1})|| > \varepsilon$ **do**

  Find $\tau^k$ by backtracking

  $\Theta^k \leftarrow prox_{\tau^k \lambda \mathcal{P}}(\Theta^{k-1} - \tau^k \nabla \mathcal{L}(\Theta^{k-1}))$

**end**

Backtracking:

- given an intial value $\tau_0$ and $\alpha \in ]0;1[$
- find the first $\tau = \tau_0 \alpha^i, i \in \{0, 1, \dots\}$ satisfying:

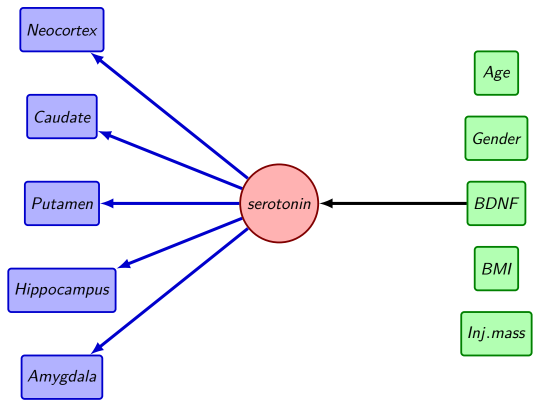
$$f(\Theta^k_\tau) \leq f(\Theta^{k-1}) + \nabla f(\Theta^k_\tau)^\top (\Theta^{k-1} - \Theta^k_\tau) + \frac{1}{2\tau} ||\Theta^{k-1} - \Theta^k_\tau||_2$$

We will have $\hat{\tau} \geq \min\left(\tau_0, \frac{\alpha}{L}\right)$

# Back to our application

Lasso LVM: mispecified model !

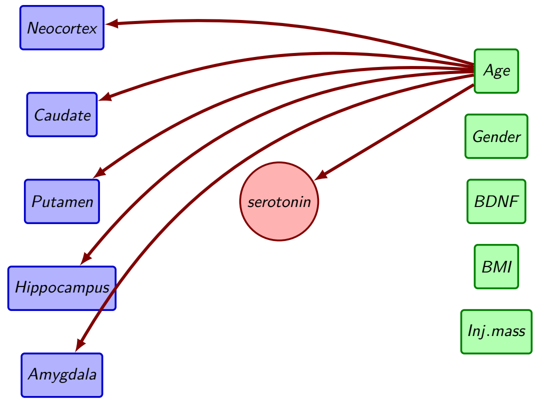
- penalize all links expect those chosen a priori



# Back to our application

Lasso LVM: mispecified model !

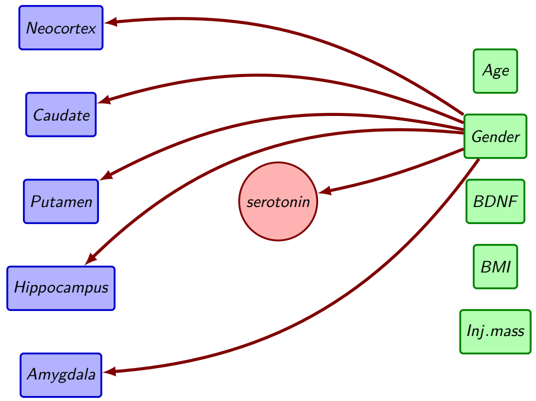
- penalize all links expect those chosen a priori



# Back to our application

Lasso LVM: mispecified model !

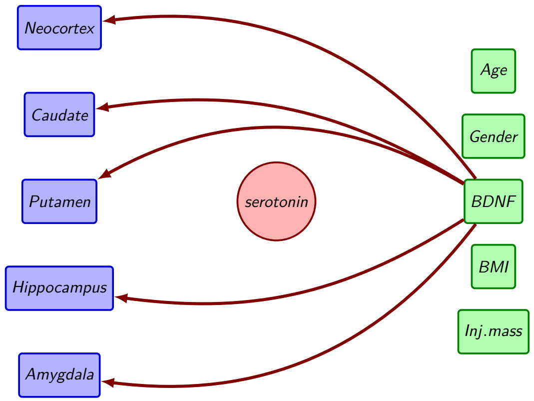
- penalize all links expect those chosen a priori



# Back to our application

Lasso LVM: mispecified model !

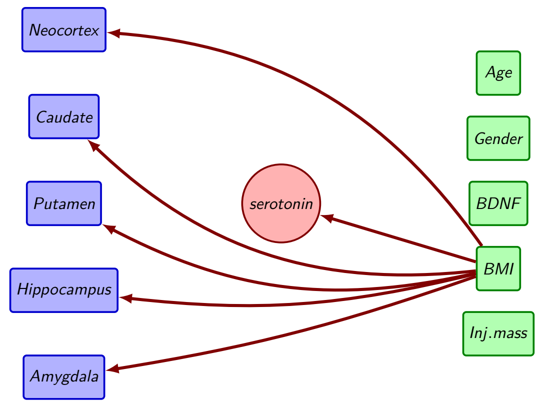
- penalize all links expect those chosen a priori



# Back to our application

Lasso LVM: mispecified model !

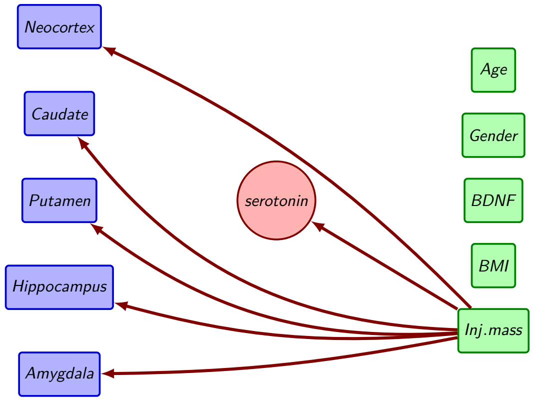
- penalize all links expect those chosen a priori



# Back to our application

Lasso LVM: mispecified model !

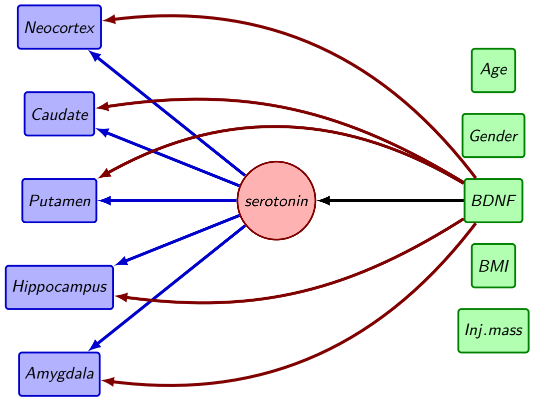
- penalize all links expect those chosen a priori



# Back to our application

Lasso LVM: mispecified model !

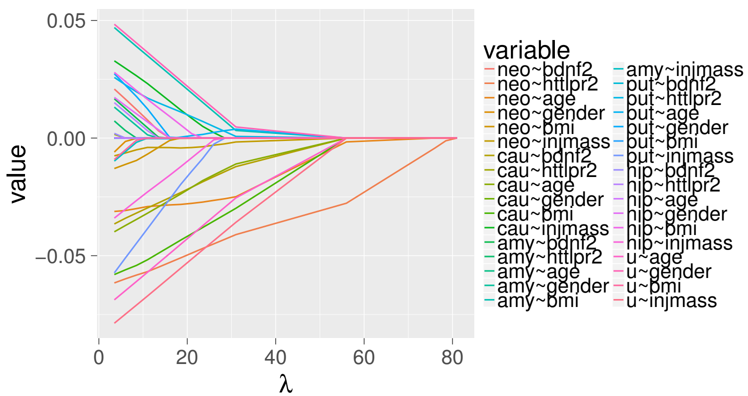
- penalize all links expect those chosen a priori



# Back to our application

Variable selection procedure:

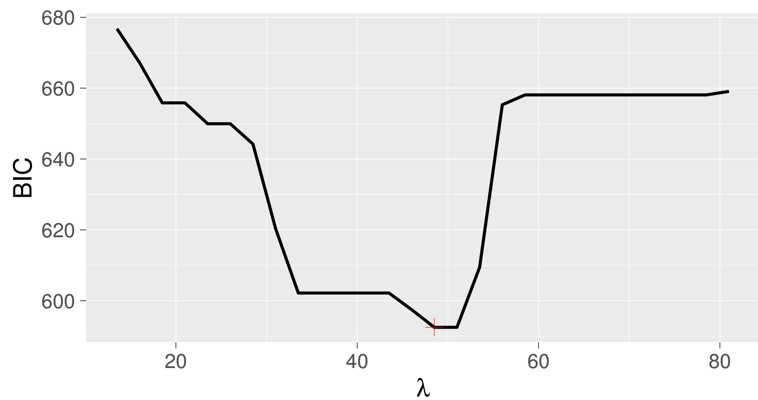
- grid search over $\lambda$
- optimal model according to the BIC



# Back to our application

Variable selection procedure:

- grid search over $\lambda$
- optimal model according to the BIC



# Simulation study

Match lasso regression estimations

- low dimensional case
- high dimensional case

Convergence of lasso LVM

- low dimensional case: ok
- high dimensional case: ok if $\lambda$ is high enough

Variable selection with lasso LVM

- conservative method

# Choosing $\lambda$

Limitation of grid search

- may miss interesting $\lambda$
- time consuming

Regularization path

- set of $\lambda$ where the set of non 0 coefficients changes
  $\Rightarrow$ called "breakpoints"
- likely to be the set of relevant $\lambda$

EPSODE algorithm

- (**Zhou2014a**) proposed a generalization of LARS to convex functions
  $\Rightarrow$ applicable to LVM ?

# Penalization path for LVM

$$\frac{\partial \Theta}{\partial \lambda} = ?$$

# Penalization path for LVM

$$f_\lambda(\Theta) = \mathcal{L}(\Theta) + \lambda||\Theta||_1 = \mathcal{L}(\Theta) + \lambda(\Theta^+ + \Theta^0 - \Theta^-)$$

For a small $d\lambda$:

$$\begin{aligned}
&\underset{d\Theta}{\text{argmin}}\,(f_{\lambda+d\lambda}(\Theta + d\Theta) - f_\lambda(\Theta)) \\
=&\underset{d\Theta,\eta}{\text{argmin}}\Big(\nabla\mathcal{L}(\Theta)d\Theta + \frac{1}{2}\nabla^2\mathcal{L}(\Theta)(d\Theta)^2 + o((d\Theta)^2) \\
&\quad +(\lambda + d\lambda)(d\Theta^+ + d\Theta^-) + \eta d\Theta^0\Big),\ \eta \text{ lagrange multiplier} \\
=& \ldots
\end{aligned}$$

So

$$\frac{d\Theta}{d\Lambda} = -P(\nabla^2\mathcal{L}(\Theta), sign(\Theta))u_z(sign(\Theta))$$

## Estimation - Penalization path

$$\frac{d\Theta}{d\Lambda} = -P(\nabla^2\mathcal{L}(\Theta), sign(\Theta))u_z(sign(\Theta))$$

$P$ matrix

$u_z$ vector

Linear regression:

- $\nabla^2\mathcal{L}(\Theta)$ piecewise constant

$\Rightarrow$ $P$ piecewise constant (**Efron2004** - LARS)

# Estimation - Penalization path

$$\frac{d\Theta}{d\Lambda} = -P(\nabla^2\mathcal{L}(\Theta), sign(\Theta))u_z(sign(\Theta))$$

$P$ matrix

$u_z$ vector

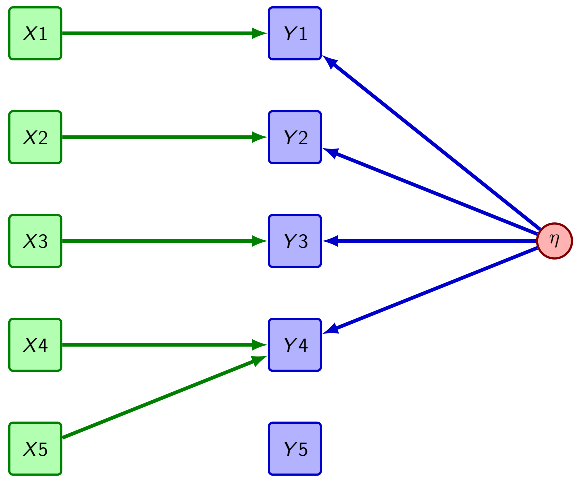
LVM:

- $\nabla^2\mathcal{L}(\Theta)$ not constant

$\Rightarrow$ Solve differential equation
**Assumption:** $\nabla^2\mathcal{L}(\Theta)$ constant between two discretization points

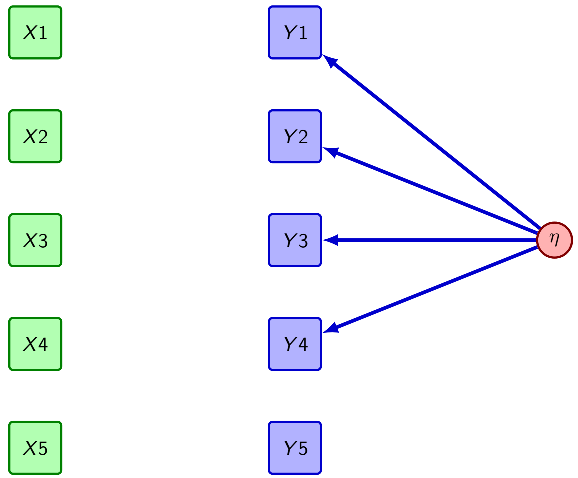
# TRUE Model

$Yi \sim \mathcal{N}(0,1)$ $n = 500$



# pLVM containing the TRUE model

All links are penalized except those shown below:



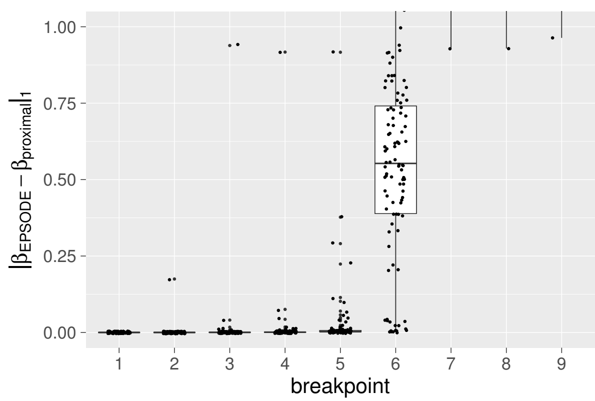
# Simulation study

For each sample:

- Simulate data according to the TRUE model
- Estimate the breakpoints for the pLVM using EPSODE:
  $\rightarrow \{\lambda_1, \dots \lambda_p\}$
- Keep the coefficients of the pLVM estimated by EPSODE
  $\rightarrow \beta_{EPSODE}$
- Proximal gradient for the pLVM applied at $\{\lambda_1, \dots \lambda_p\}$
  $\rightarrow \beta_{proxGrad}$

- agreement: $\sum_{j=1}^{p} |\beta_{proxGrad,j} - \beta_{EPSODE,j}|$

# Accuracy of the regularization path




⇒ incorrect after a number of breakpoints (here 5)

## Summary

Integration of regularization into LVM:

- proximal gradient
- lasso, ridge, elastic net, group lasso penalty
- nuclear norm

$\Rightarrow$ user-specific penalty terms can be used specifying the proximal operator

Regularization path:

- lasso, ridge, elastic net
- increasing bias along the path
  - need explicit formulation for the hessian ?
  - need thinner mesh ?

# Perspectives

- nuclear norm penalty (n=500,p=4096+5)