tigaR: temporal integrative genomics analysis in R

Integration Meeting
19-06-2015, Delft
The Netherlands

Contributors

Statistics

- Viktorian Miok
- Wessel van Wieringen
- Mark van de Wiel

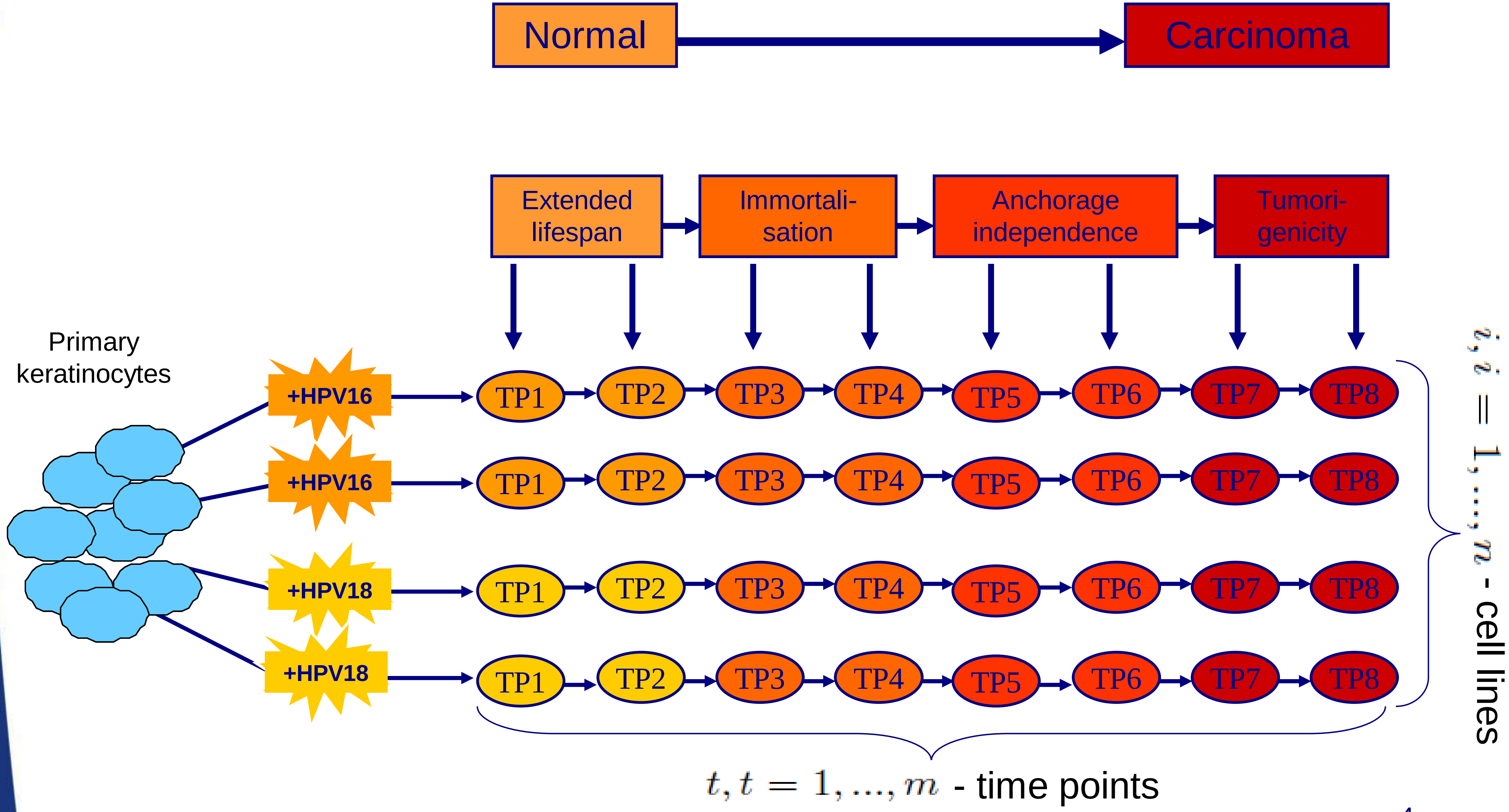
Biology

- Saskia Wilting
- Annelieke Jaspers
- Renske Steenbergen
- Peter Snijders
- Paula van Noort
- Ruud Brakenhoff

Cervical cancer study

- Second most common cancer in women worldwide.
- Caused by HPV virus, in 80% cases HPV16 and HPV18.
- Cell line model – in vitro model system of HPV-induced transformation.
- Integration – high-throughput multi level molecular data sets.
- Aim: identification of key genes.

Experiment


Normal
Carcinoma
Extended lifespan
Immortalisation
Anchorage independence
Tumorigenicity
Primary keratinocytes
+HPV16
+HPV16
+HPV18
+HPV18
TP1
TP2
TP3
TP4
TP5
TP6
TP7
TP8
$i, i = 1, ..., n$ - cell lines
$t, t = 1, ..., m$ - time points

Model

$j, j = 1, ..., p$ - genes

$\mathbf{Y}_{*,*,t} = (\mathbf{Y}_{1,*,t}, ..., \mathbf{Y}_{n,*,t})$ - mRNA gene expression

Bayesian GLMM: $Y_{i,j,t} \sim \mathcal{N}(\mu_{i,j,t}, \sigma^2_{\varepsilon,j})$

$$\mu_{i,j,t} = \overbrace{f(i;\boldsymbol{\alpha}_j)}^{\text{Cell line effect}} + \overbrace{h(t;\boldsymbol{\gamma}_j)}^{\text{Time effect}}$$

α, γ - Gaussian distribution assumption

Fixed and random effects

Fixed effect:

$$f(i;\boldsymbol{\alpha_j}) = \boldsymbol{\alpha}_{i,j}$$

Random effect:

$$h(t;\boldsymbol{\gamma}_j) = \sum_{k=1}^{K} \gamma_{j,k} \left| t - \kappa_k \right|^3$$

Matrix notation: $$Y_{i,j,t} = \alpha_{i,j} + \tilde{\mathbf{Z}}_t \tilde{\gamma}_j + \varepsilon_{i,j,t}$$

$$\tilde{\mathbf{Z}}_t = \mathbf{Z}_t \mathbf{V}_\omega \mathbf{D}_\omega^{1/2}$$

$$\tilde{\gamma}_j = \mathbf{D}_\omega^{-1/2} \mathbf{V}_\omega^{\mathrm{T}} \gamma_j$$

Spline basis:

$$\mathbf{Z}_t = (|t-\kappa_1|^3, \ldots, |t-\kappa_K|^3)$$

Spline coefficients:

$$\boldsymbol{\gamma}_j = (\gamma_j, \ldots, \gamma_{j,K})^{\mathrm{T}}$$

INLA estimation

$\boldsymbol{\theta}$ - model parameters

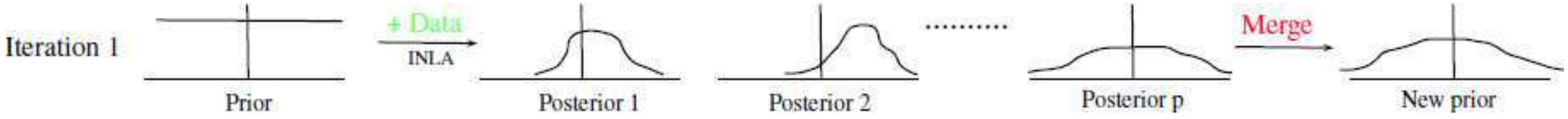
ϕ - hyper-parameters

INLA (Rue et al., 2009) procedure consist in:

- Approximate full posterior of $\pi(\phi|\boldsymbol{y})$ and $\pi(\theta_l|\phi, \boldsymbol{y})$ using Laplace approximation.

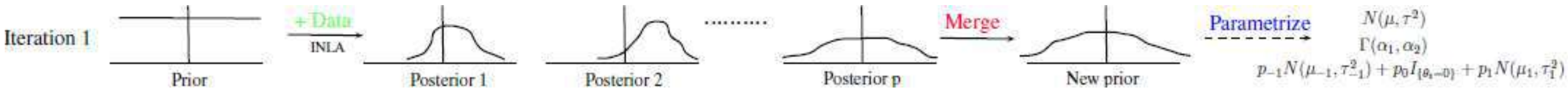
- Approximate marginal posterior densities of $\boldsymbol{\theta}$ and ϕ integrating over hyper-parameters of posteriors $\pi(\phi|\boldsymbol{y})$ and $\pi(\theta_l|\phi, \boldsymbol{y})$.

Model parameters estimation



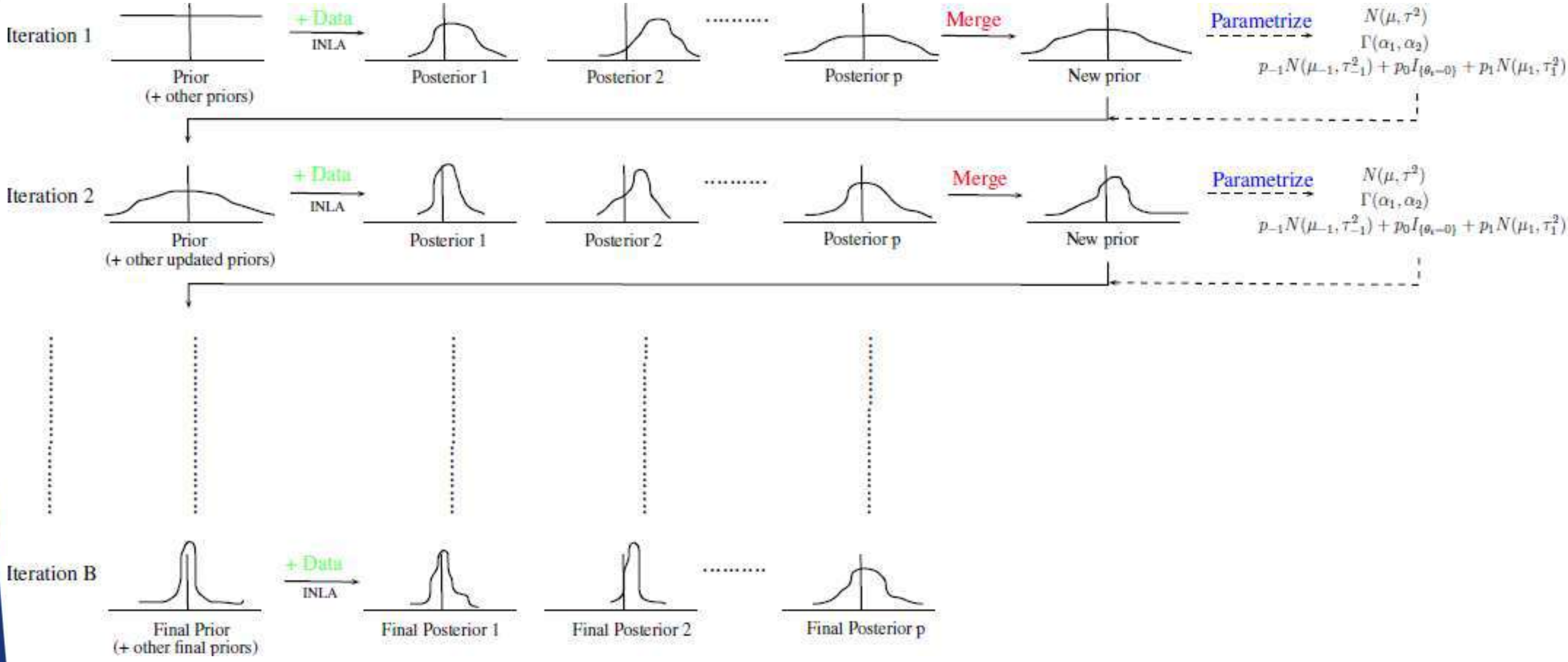
- Procedure start with flat prior
- Using data and INLA poster distributions are estimated
- Merge the posterior distribution – borrowing of the information

Model parameters estimation


Iteration 1
Prior
+ Data
INLA
Posterior 1
Posterior 2
....
Posterior p
Merge
New prior
Parametrize
$N(\mu, \tau^2)$
$\Gamma(\alpha_1, \alpha_2)$
$p_{-1}N(\mu_{-1}, \tau^2_{-1}) + p_0 I_{\{\theta_i = 0\}} + p_1 N(\mu_1, \tau^2_1)$

- Merged posterior distribution is used as prior distribution
- New prior distribution is parametrized - shrinkage

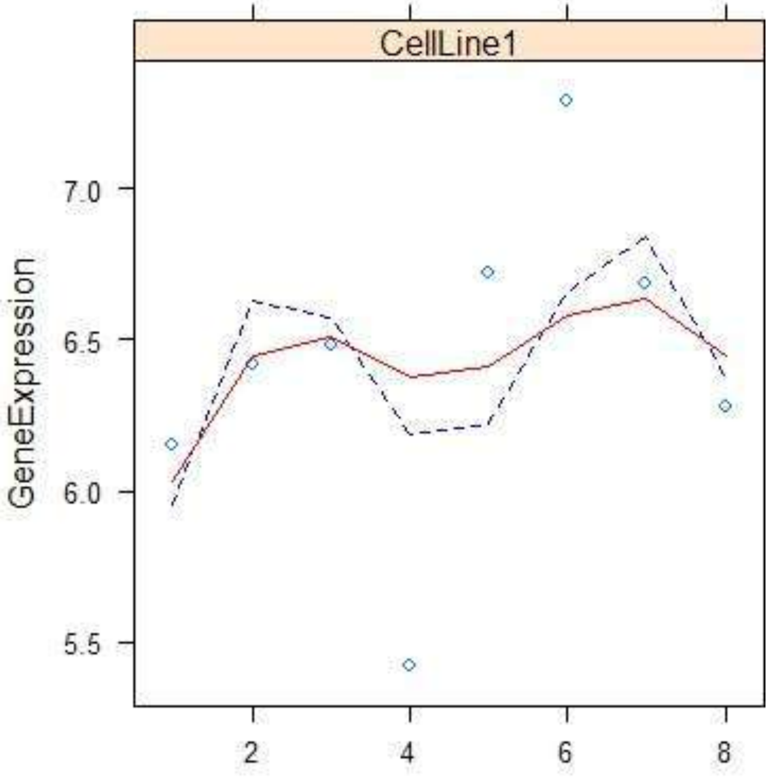
Model parameters estimation



van de Wiel et al. (2013), Biostatistics.

Shrinkage

- borrowing information across the genes
- better control of false positives
- leads to more stable estimates
- improvement of reproducibility



Comparison of the methods

- **tigaR** – Miok et al., BMC Bioinformatics, 2014.
- **EDGE** – Storey et al., PNAS., 2005.
- **timecourse** – Tai and Speed, Annals of Statistics, 2006.
- **BATS** – Angelini et al., Stat. Appl. Genet. Mol. Biol., 2007.

Comparison set-up

Data

- Real data from the experiment

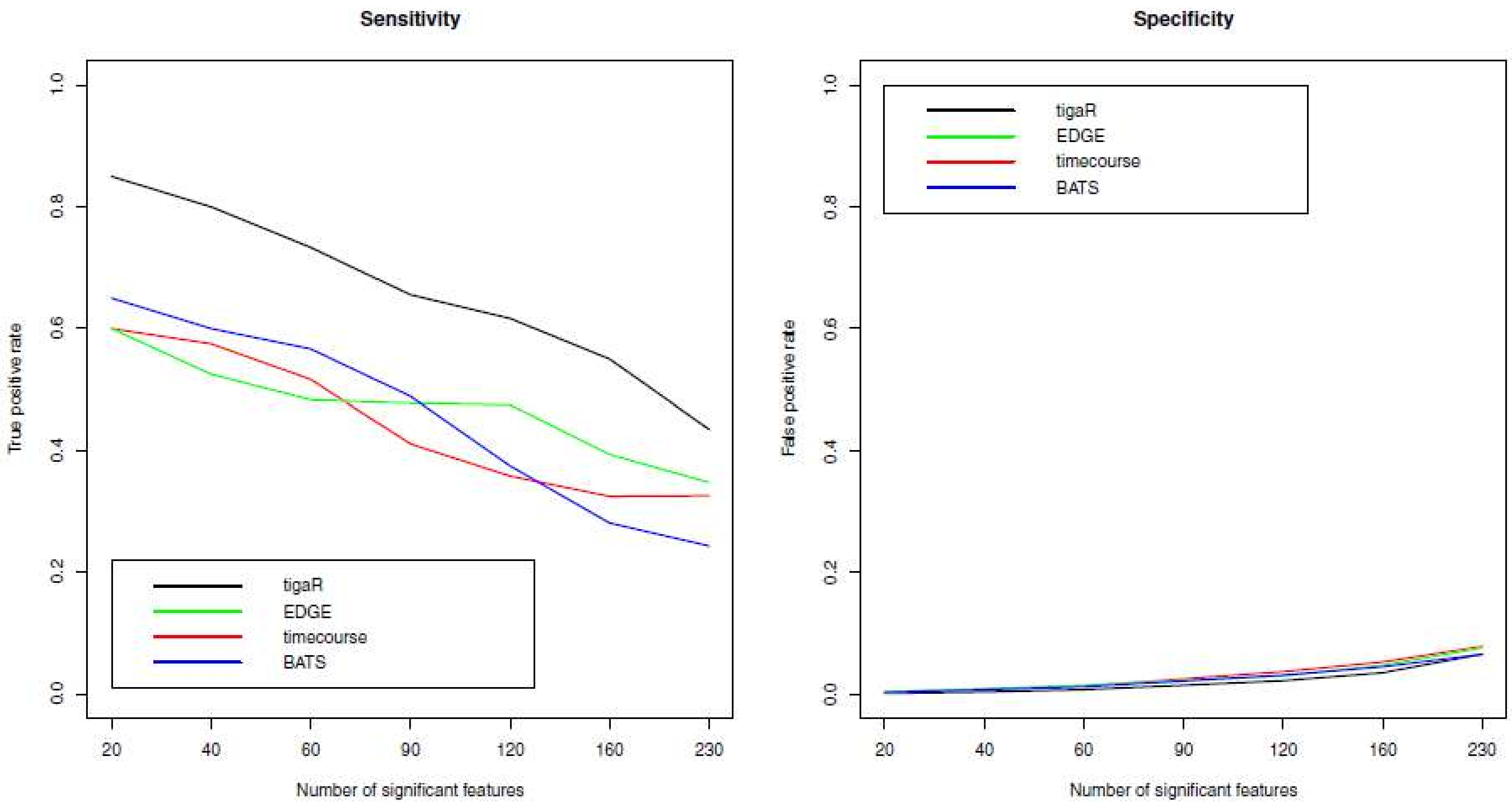
Sensitivity and specificity

- Truth – significant genes among methods
- Calculate true and false positive rate.

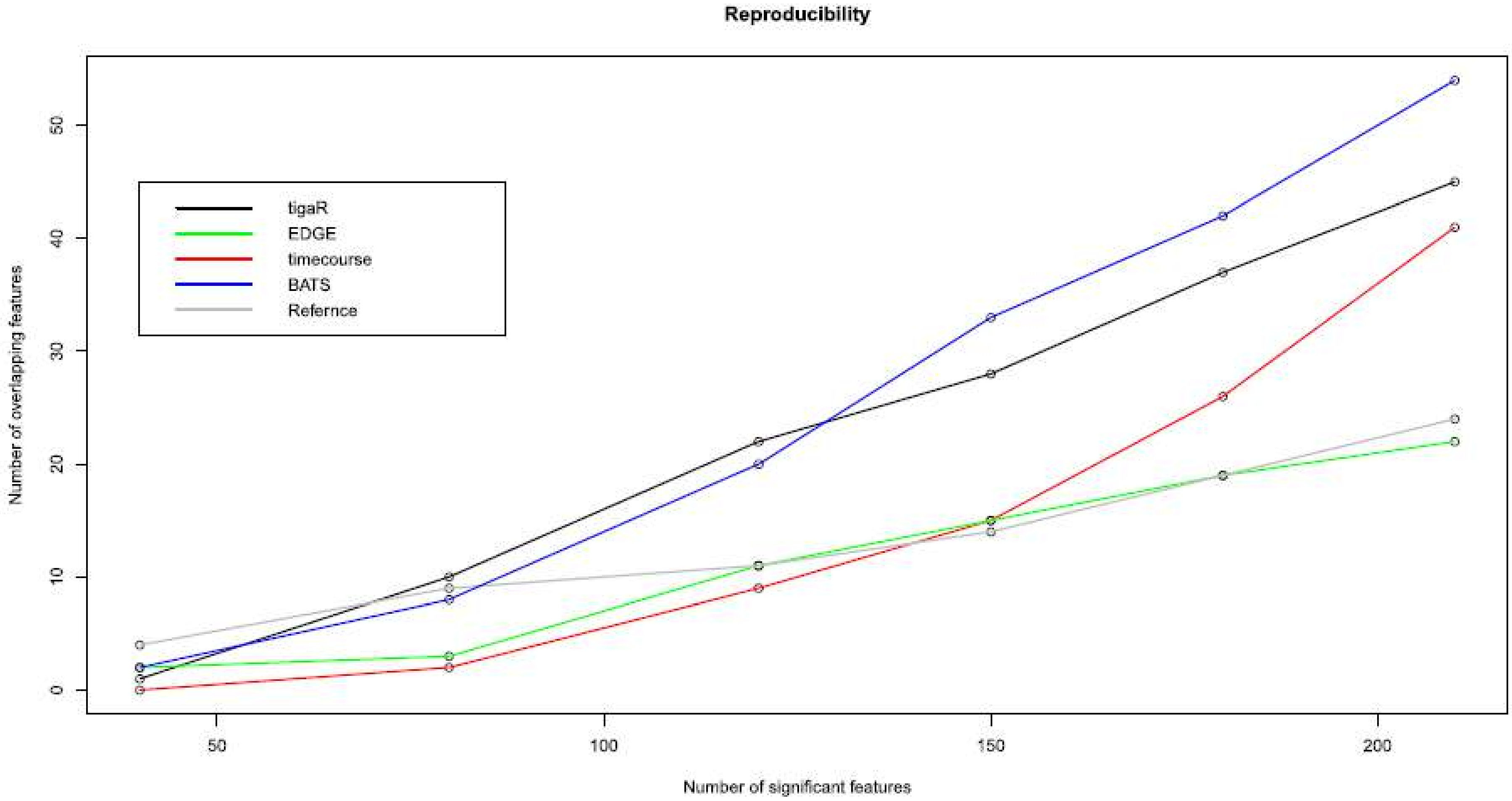
Reproducibility

- Equally divided data set in two groups
- Methods applied on the groups and calculate number of overlaps

Sensitivity and specificity



Reproducibility

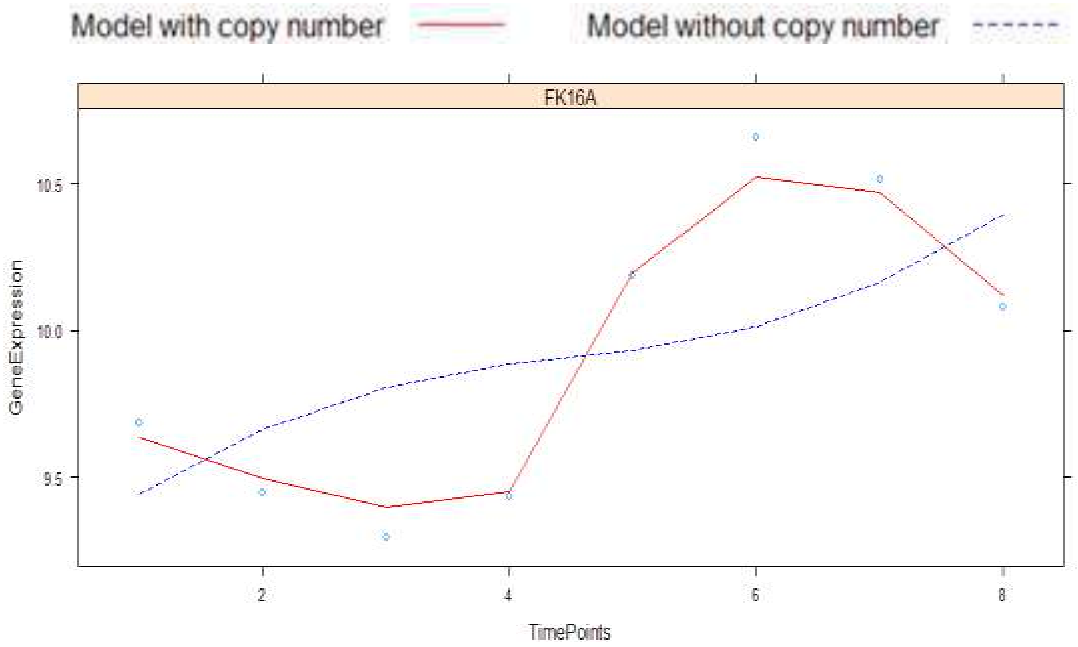


DNA copy number (CN)

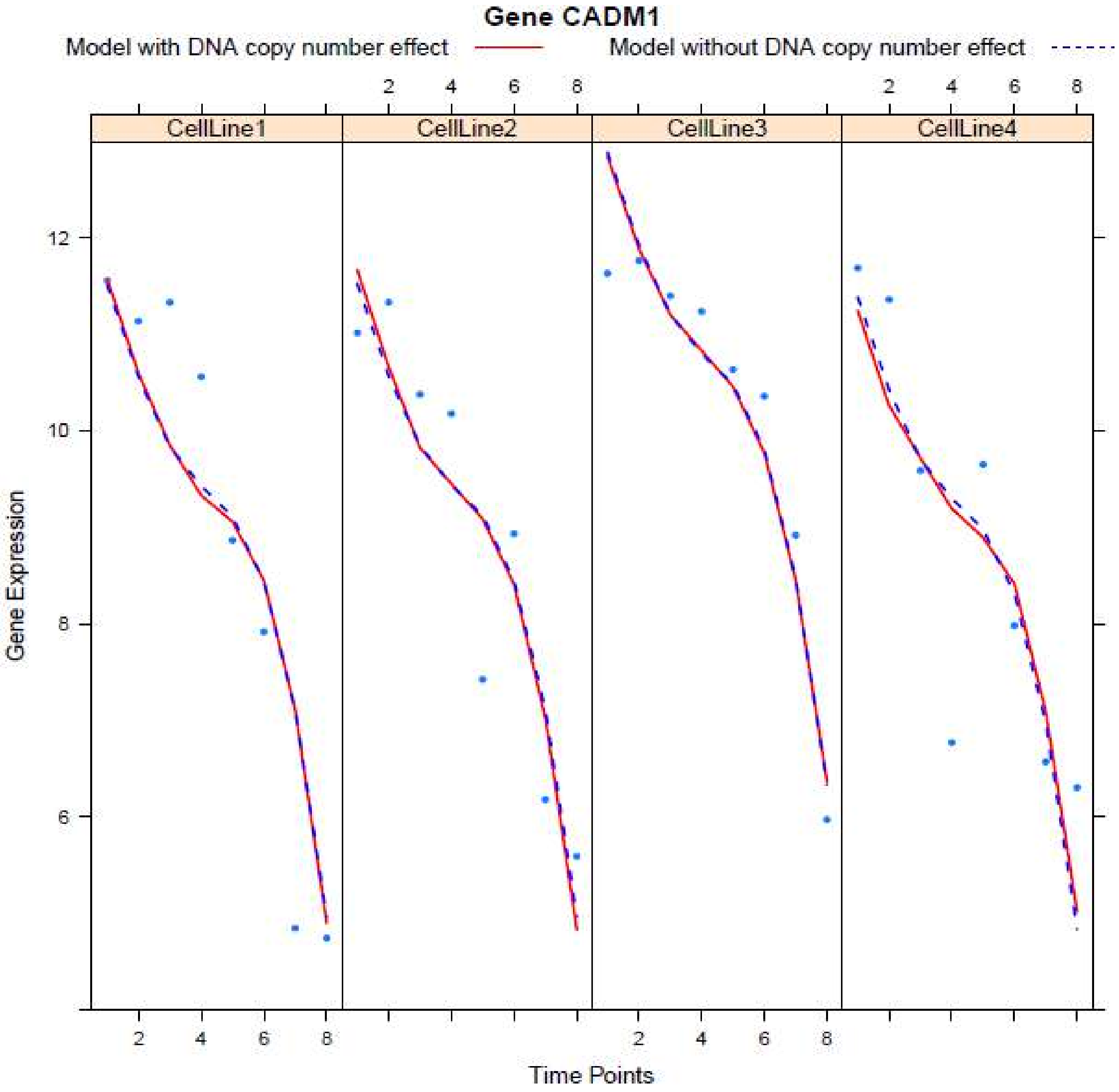
$$\mathbf{X}_{*,*,t} = (\mathbf{X}_{1,*,t}, \ldots, \mathbf{X}_{n,*,t})$$ - CN observations

$$Y_{i,j,t} = \overbrace{\alpha_{i,j}}^{\text{Cell line}} + \overbrace{\beta_j\, x_{i,j,t}}^{\text{CN}} + \overbrace{\tilde{\mathbf{Z}}_t \tilde{\boldsymbol{\gamma}}_j}^{\text{Time}} + \overbrace{\varepsilon_{i,j,t}}^{\text{Error}}$$

Gene GSTM3:

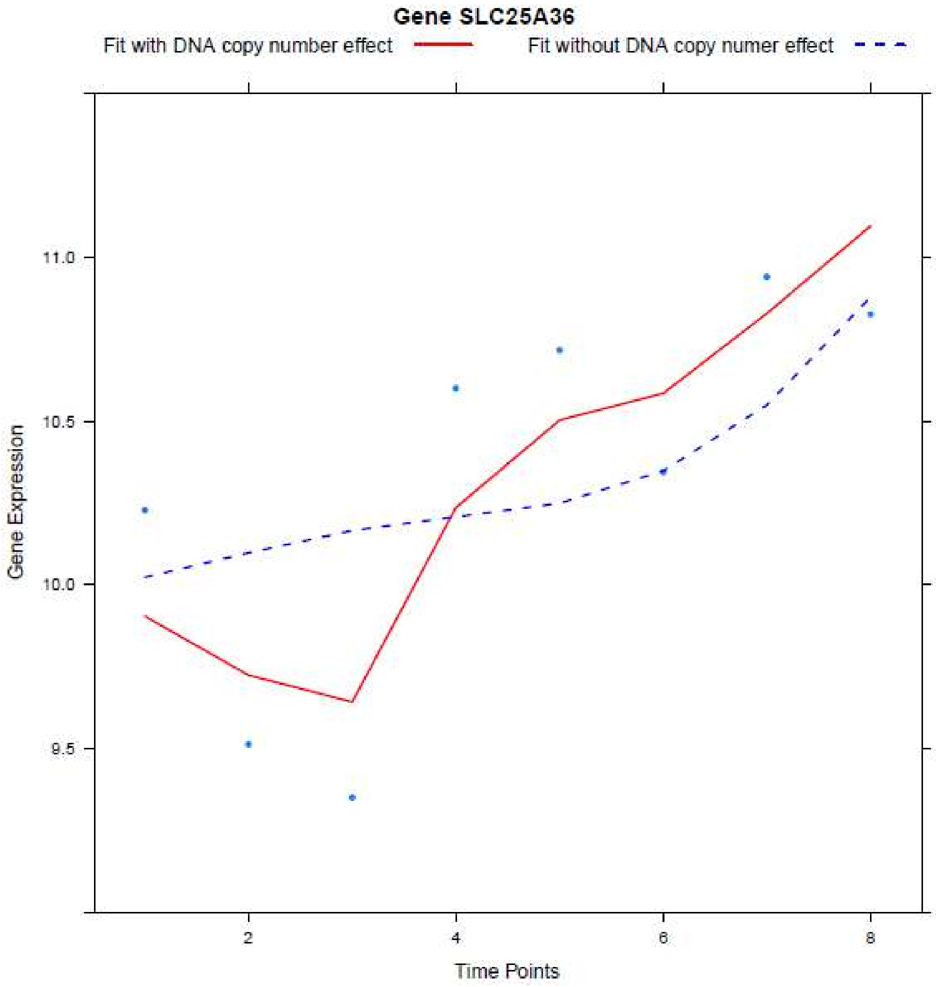


CADM1- gene without CN effect



Overmeer et al., J Pathol., 2008.

SLC25A36 – gene with CN effect



Wilting et al., Genes, Chromosomes and Cancer, 2008.

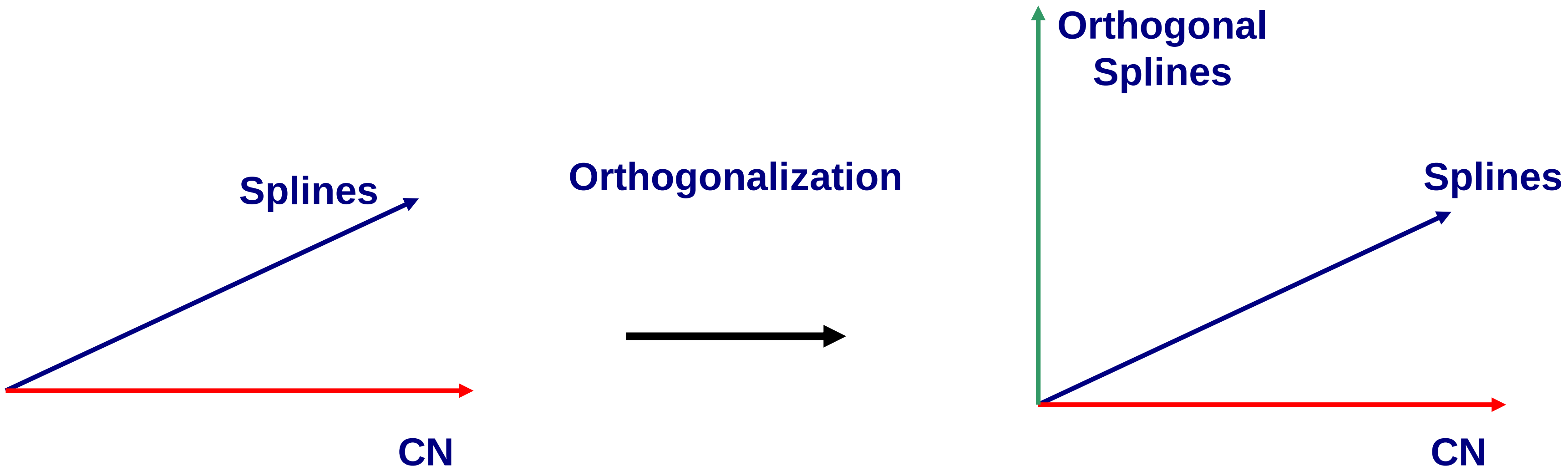
Time vs. CN effect

Probelem:

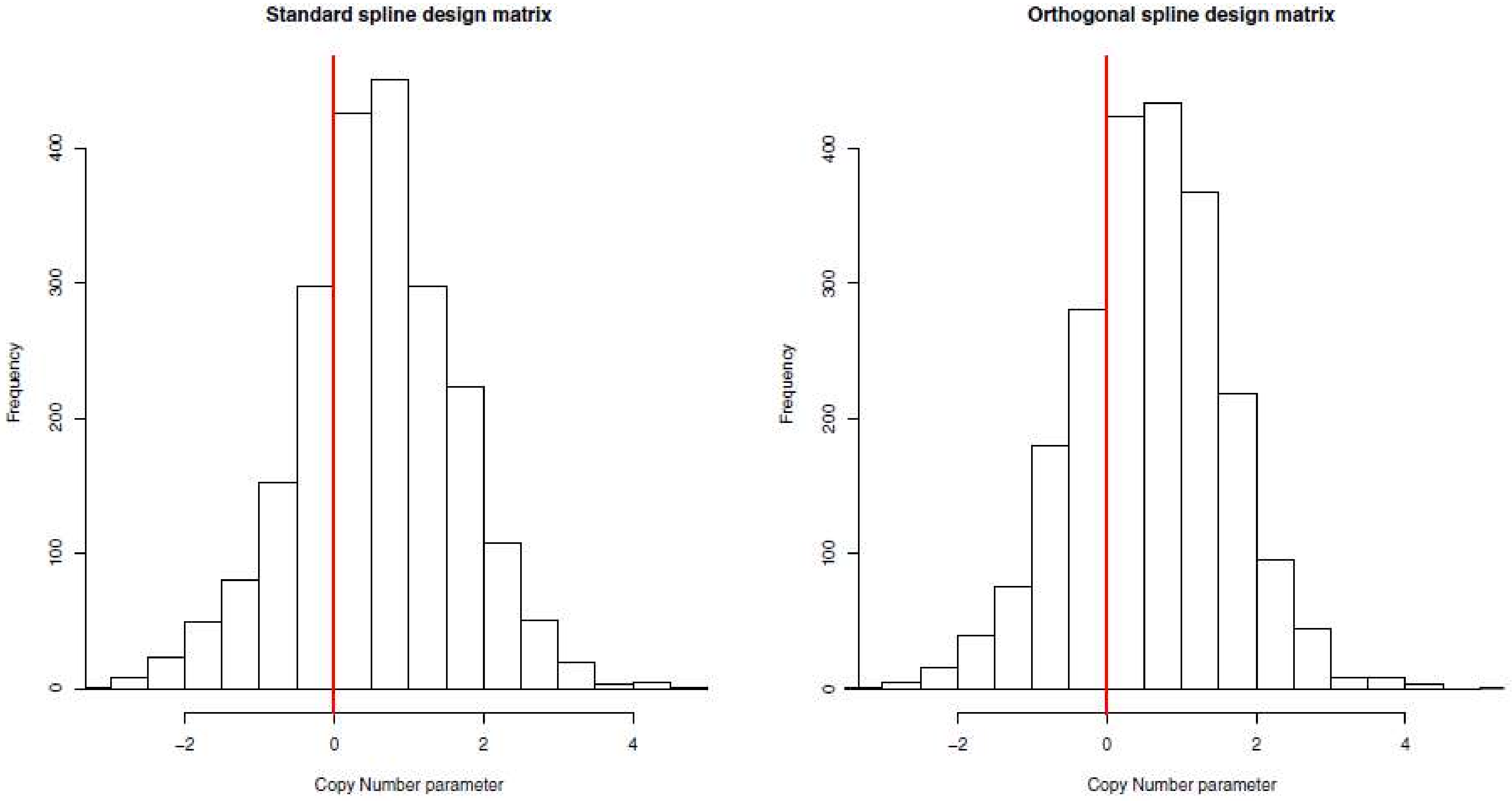
- Flexibility of the splines(time) consumes effect of CN

Potential solution

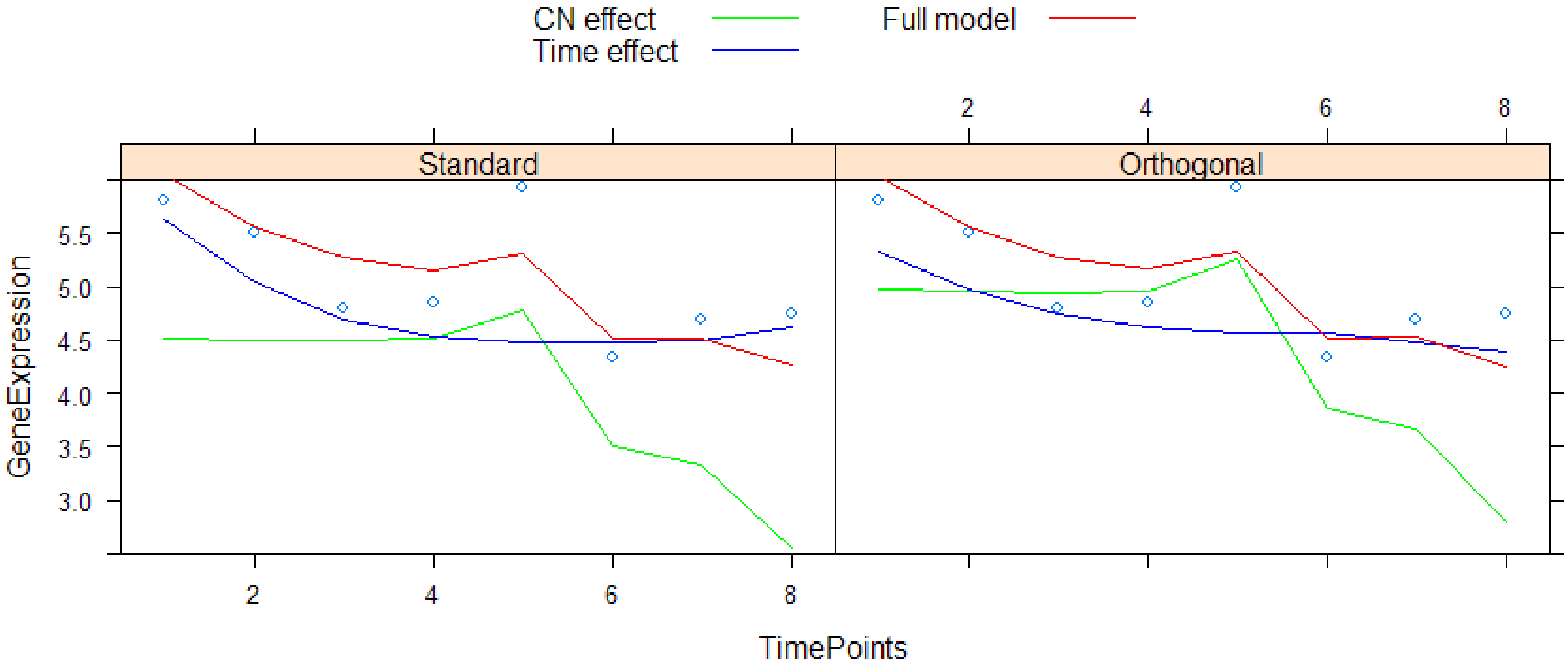
- Orthgonalization of the splines onto CN design matrix



CN parameter



Fit of the model



Spatial multivariate prior for CN

β_j follow the first-order autoregressive process along the genome:

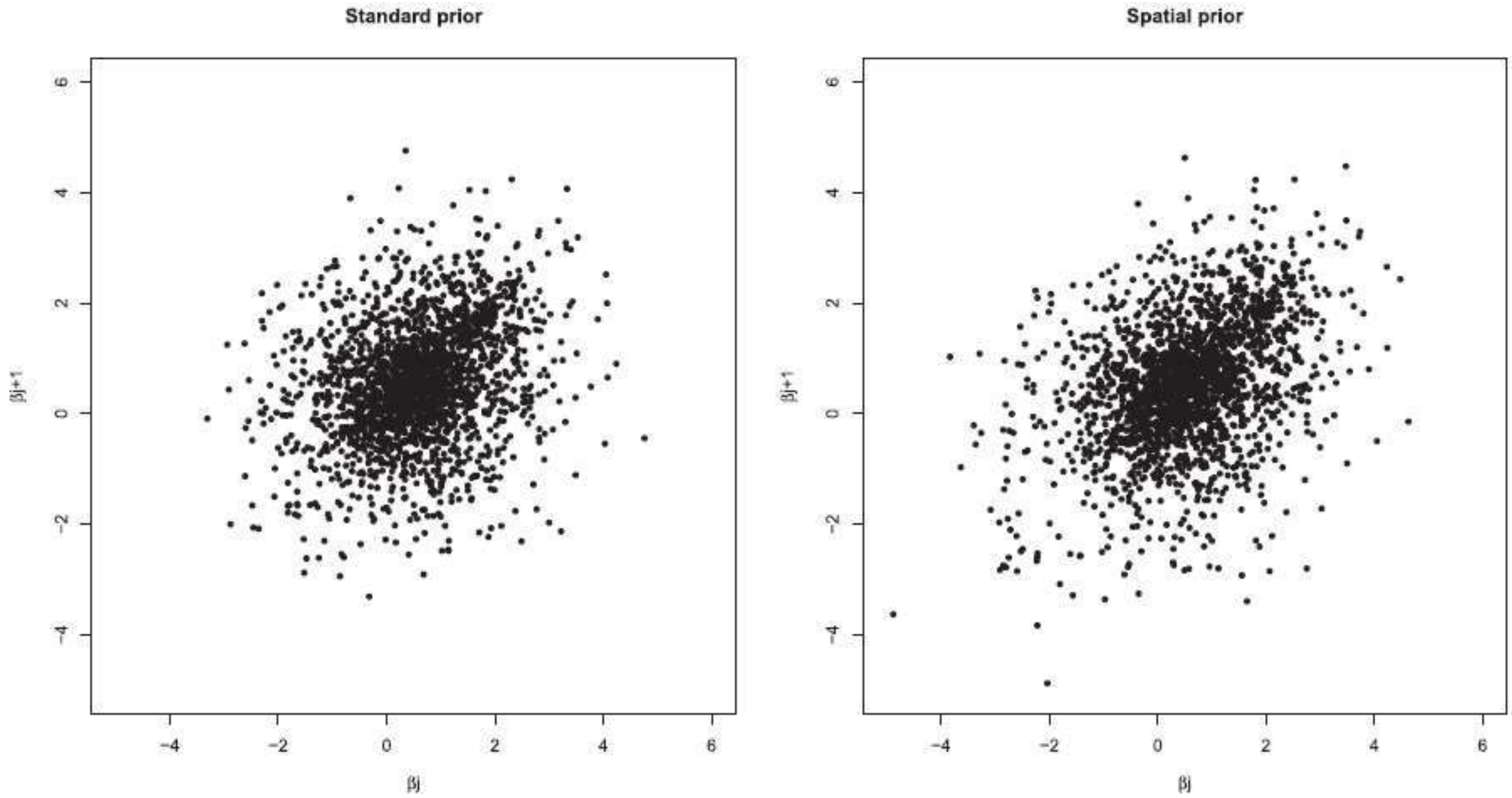
$$\beta_j = \rho\beta_{j-1} + \varepsilon_j$$

For each triplet trivariate normal prior is assumed:

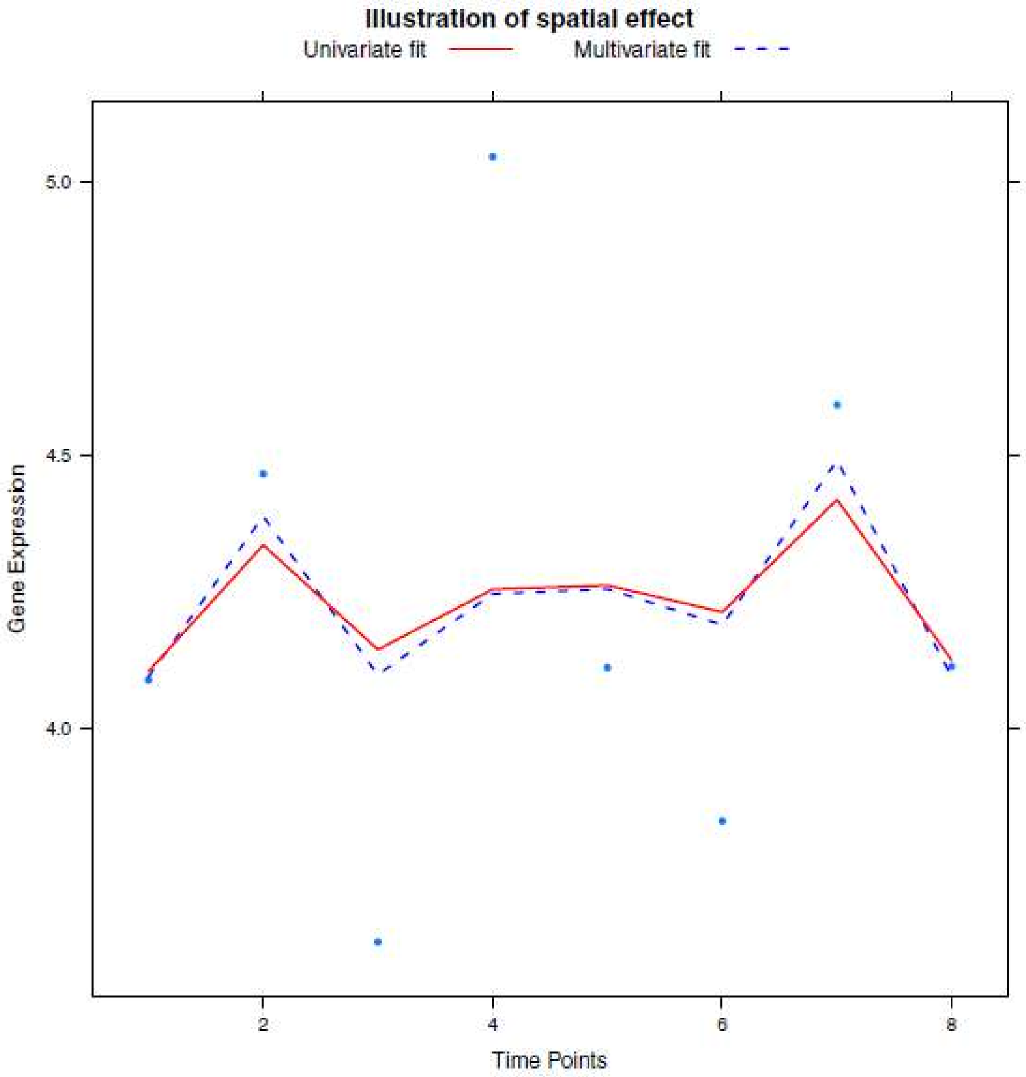
$$\begin{pmatrix} \beta_{j-1} \\ \beta_j \\ \beta_{j+1} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{j-1}^2 & \sigma_{j-1}\sigma_j\rho & \sigma_{j-1}\sigma_{j+1}\rho^2 \\ \sigma_{j-1}\sigma_j\rho & \sigma_j^2 & \sigma_j\sigma_{j+1}\rho \\ \sigma_{j-1}\sigma_{j+1}\rho^2 & \sigma_j\sigma_{j+1}\rho & \sigma_{j+1}^2 \end{pmatrix} \right)$$

CN parameters

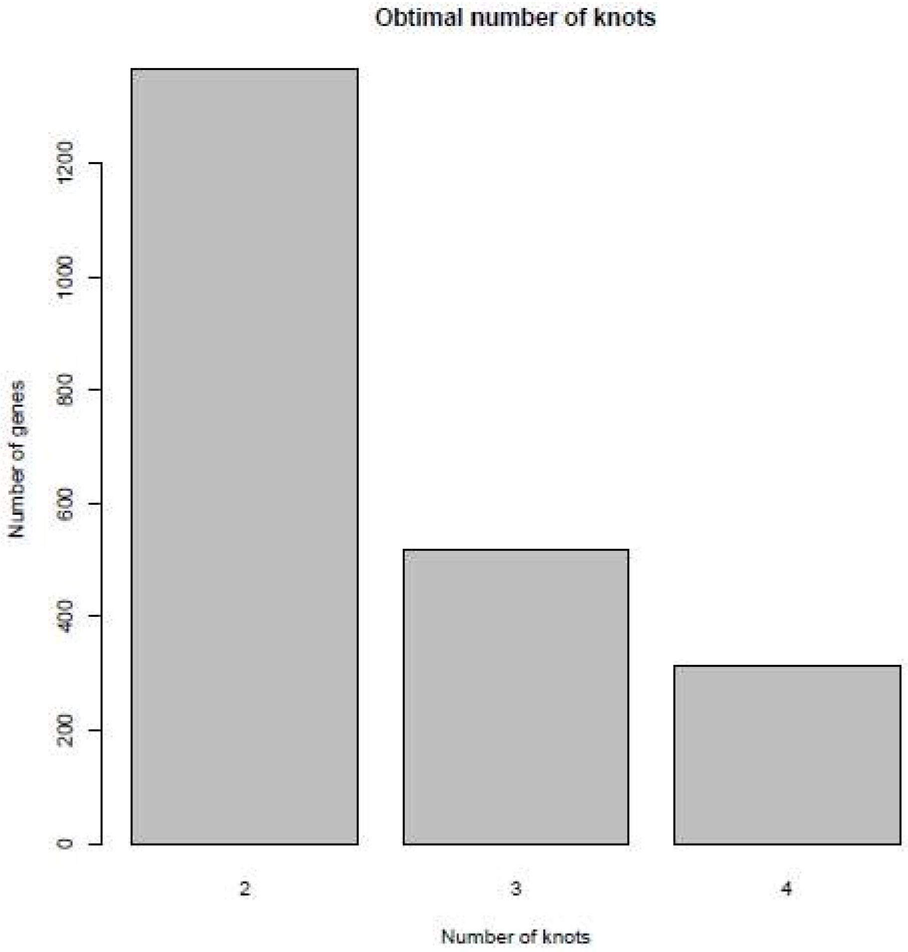
Partial correlation of CN parameter:



Fit of the model



Optimal number of knots for splines



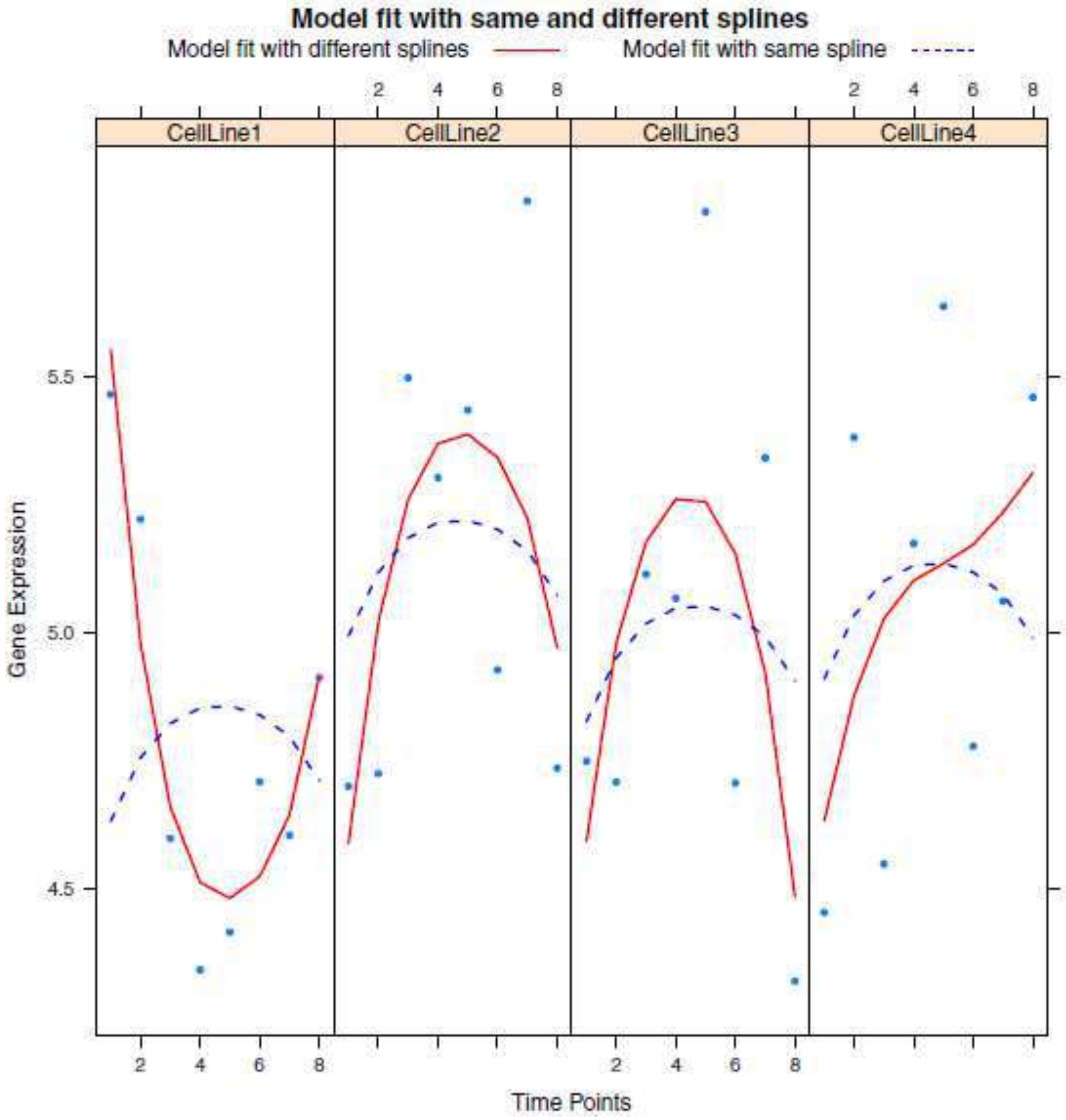
Splines flexibility

Same spline – up/down regulated genes

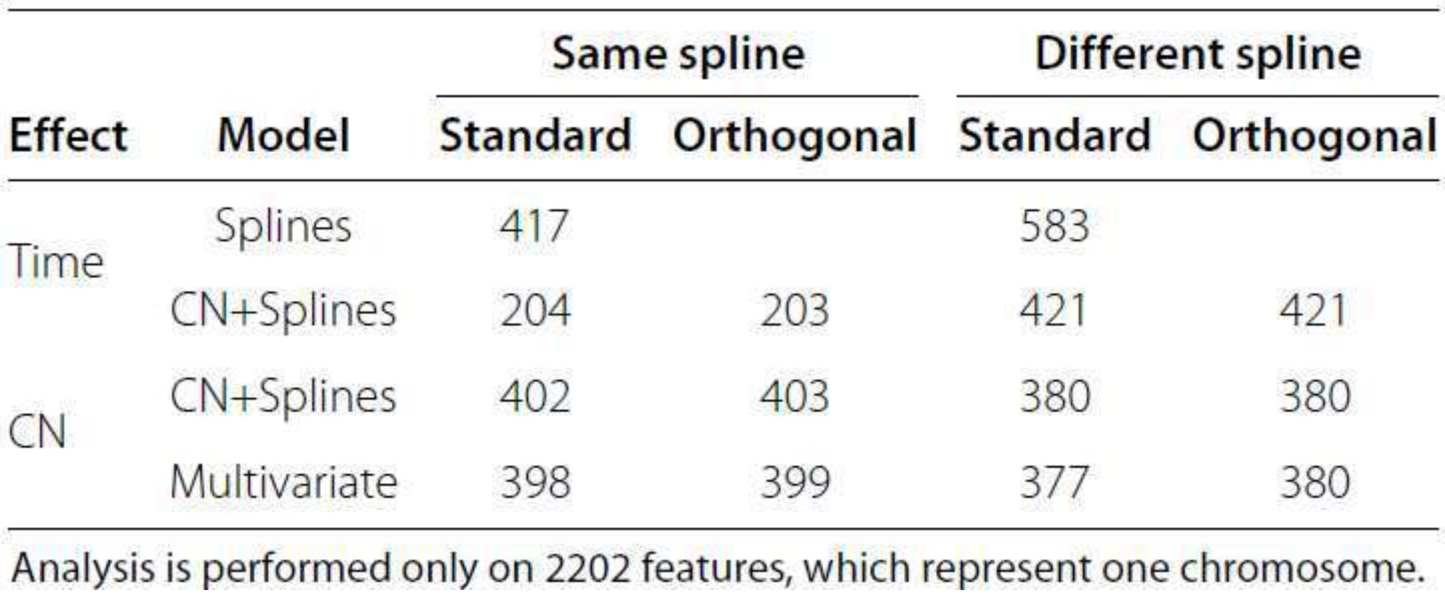
$$\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}} \otimes \mathbf{1}_{n\times n}$$

Different spline – allow more flexibility

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}} \otimes \mathbf{I}_{n\times n}$$



Application



		Same spline		Different spline	
Effect	Model	Standard	Orthogonal	Standard	Orthogonal
Time	Splines	417		583	
	CN+Splines	204	203	421	421
CN	CN+Splines	402	403	380	380
	Multivariate	398	399	377	380

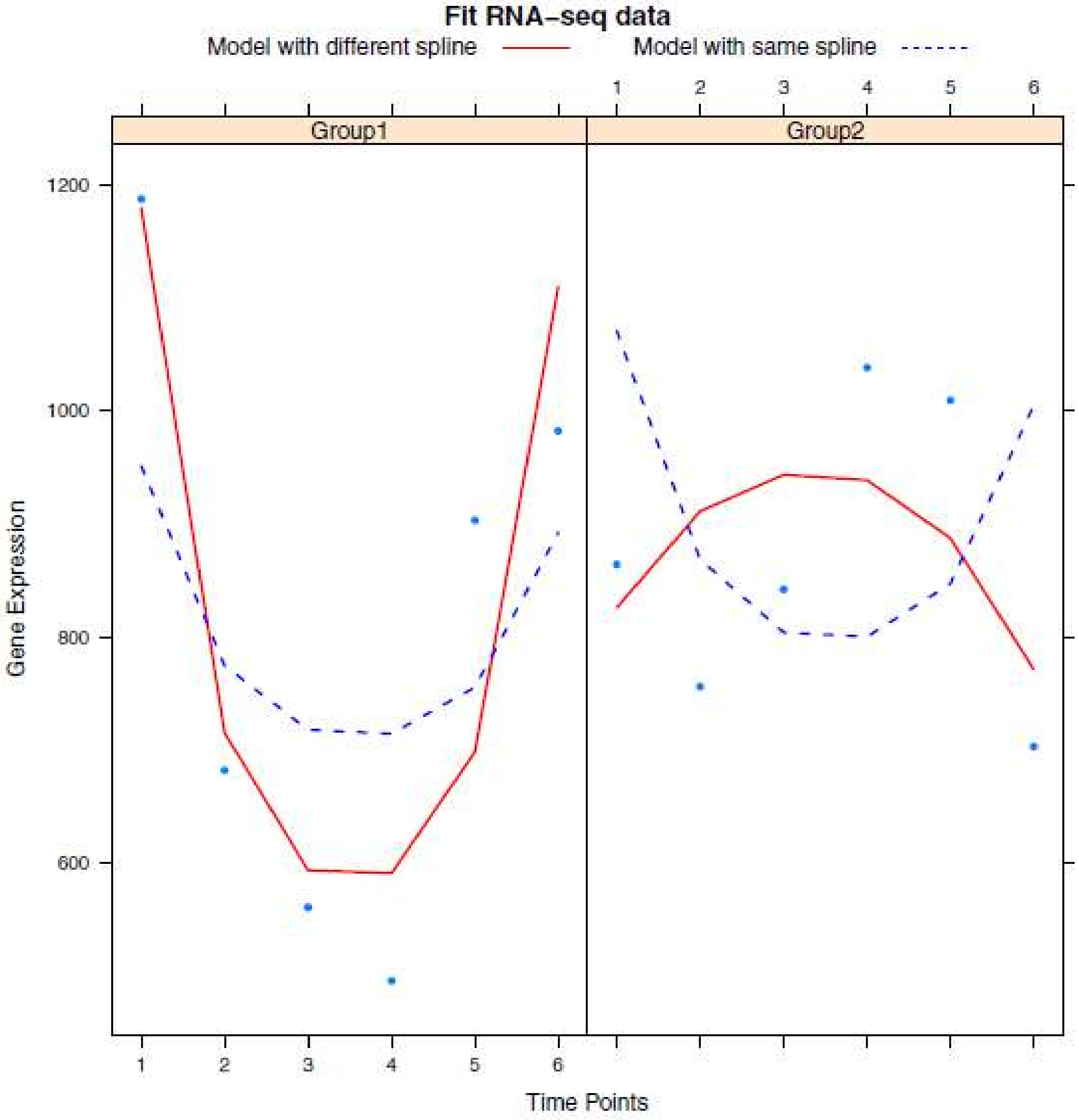
Analysis is performed only on 2202 features, which represent one chromosome.

Method identify genes with time and CN effect allowing for:

- flexibility in modeling of time effect
- additional stability of CN parameters

RNA-seq data

- Changing link function method can deal with count data.
- Two group time-course RNA-seq data.



Summary

- Improved identification of temporal differential gene expression (TDGE) using penalized splines and empirical Bayes shrinkage.

- Identification of TDGE induced by CN.

- Identification of TDGE in count RNA-seq data.

- Improvement of CN estimates, with orhogonalization and imposing spatial multivariate prior.

- Identification of significant up or down regulated genes.

- As a proof of principle biologically relevant genes **SLC25A36** and **CADM1** are identified.

Thank you for your attention!