A semi-parametric empirical Bayes approach for time-course integrative genomic analysis

Viktorian Miok[1,2], Saskia Wilting[2], Mark van de Wiel[1,3], Annelieke Jaspers[2], Peter Snijders[2], Renske Steenbergen[2], Wessel van Wieringen[1,3]

[1]Department of Epidemiology & Biostatistics and [2]Department of Pathology, VU University Medical Center, [3]Department of Mathematics, VU University, Amsterdam, The Netherlands.

Epidemiology & Biostatistics

VUmc

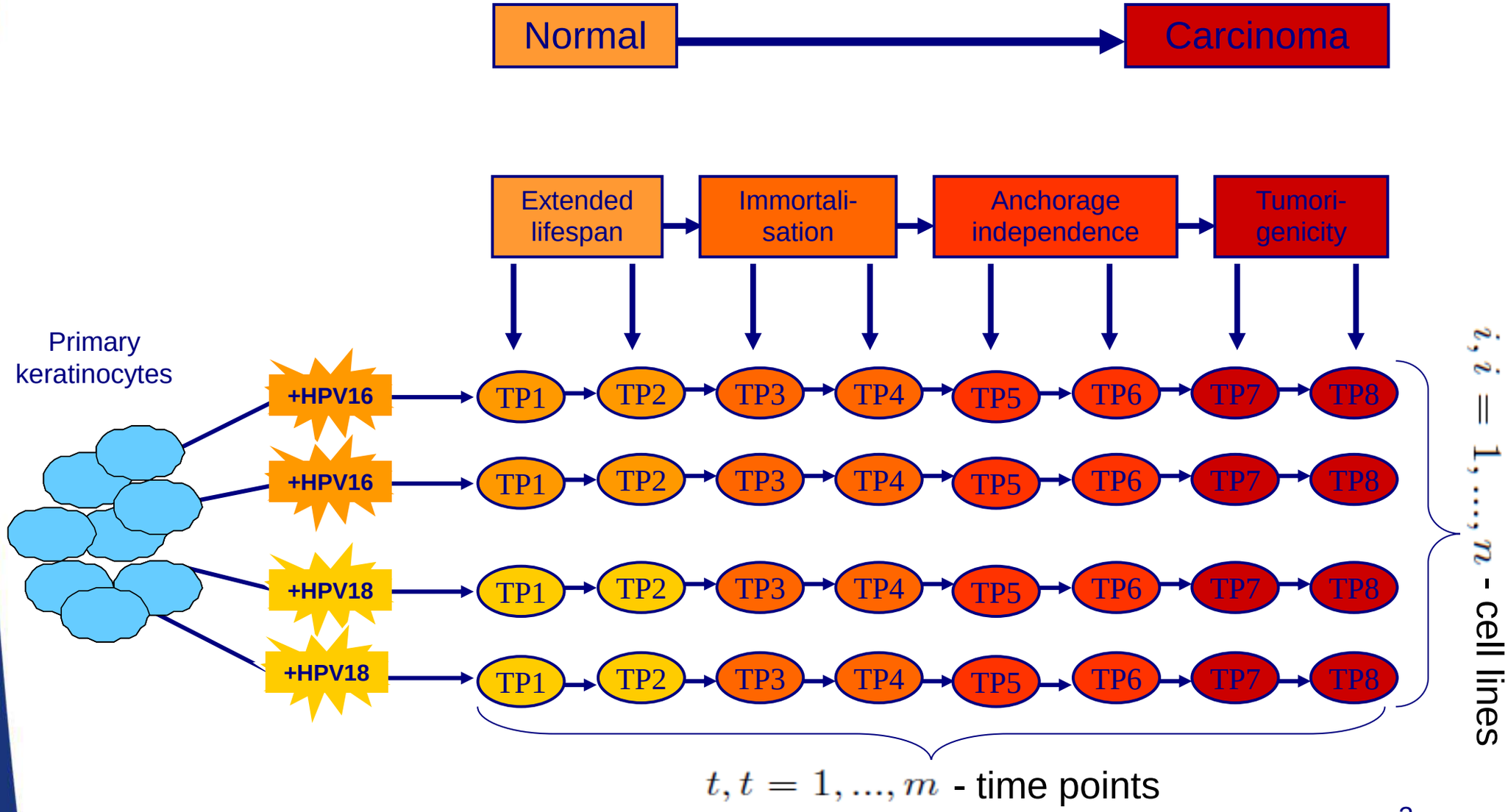
Hexa-Symposium 14-11-2013, Hasselt, Belgium

VU VRIJE UNIVERSITEIT AMSTERDAM

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
Immortali-sation
Anchorage independence
Tumori-genicity
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

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

- **Bayesian GLMM**:

$$\mathbf{Y}_{ijt} \sim \mathcal{N}(\mu_{ijt}, \sigma_j^2)$$

Cell line effect **Time effect**

$$\mu_{ijt} = f_j(v_i; \boldsymbol{\alpha}) + h_j(t; \boldsymbol{\beta}) + \varepsilon_{ijt},$$

$\boldsymbol{\alpha}, \boldsymbol{\beta}$ - Gaussian distribution assumption

Fixed and random effects

Fixed effect:

$$f_{ij}(v_i;\boldsymbol{\alpha}) = \alpha_{ij} v_i,$$

Random effect:

$$h_j(t;\boldsymbol{\beta}) = \sum_{k=1}^{K} \beta_{jk} \left|t - \kappa_k\right|^3$$

Matrix notation: $\mathbf{Y}_{it} = \tilde{\mathbf{V}}\tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_{it}$

$$\tilde{\mathbf{Z}} = \mathbf{Z}_K \boldsymbol{\Omega}_{K\times K}^{-1/2} \qquad \tilde{\boldsymbol{\beta}} = \boldsymbol{\Omega}_{K\times K}^{1/2} \boldsymbol{\beta}$$

Spline basis: $\mathbf{Z}_K = \left\{ \left|t - \kappa_1\right|^3, \ldots, \left|t - \kappa_K\right|^3 \right\}$ **Penalty matrix:** $\boldsymbol{\Omega}_{K\times K}$

Model parameters estimation

$\boldsymbol{\theta}$ - model parameters

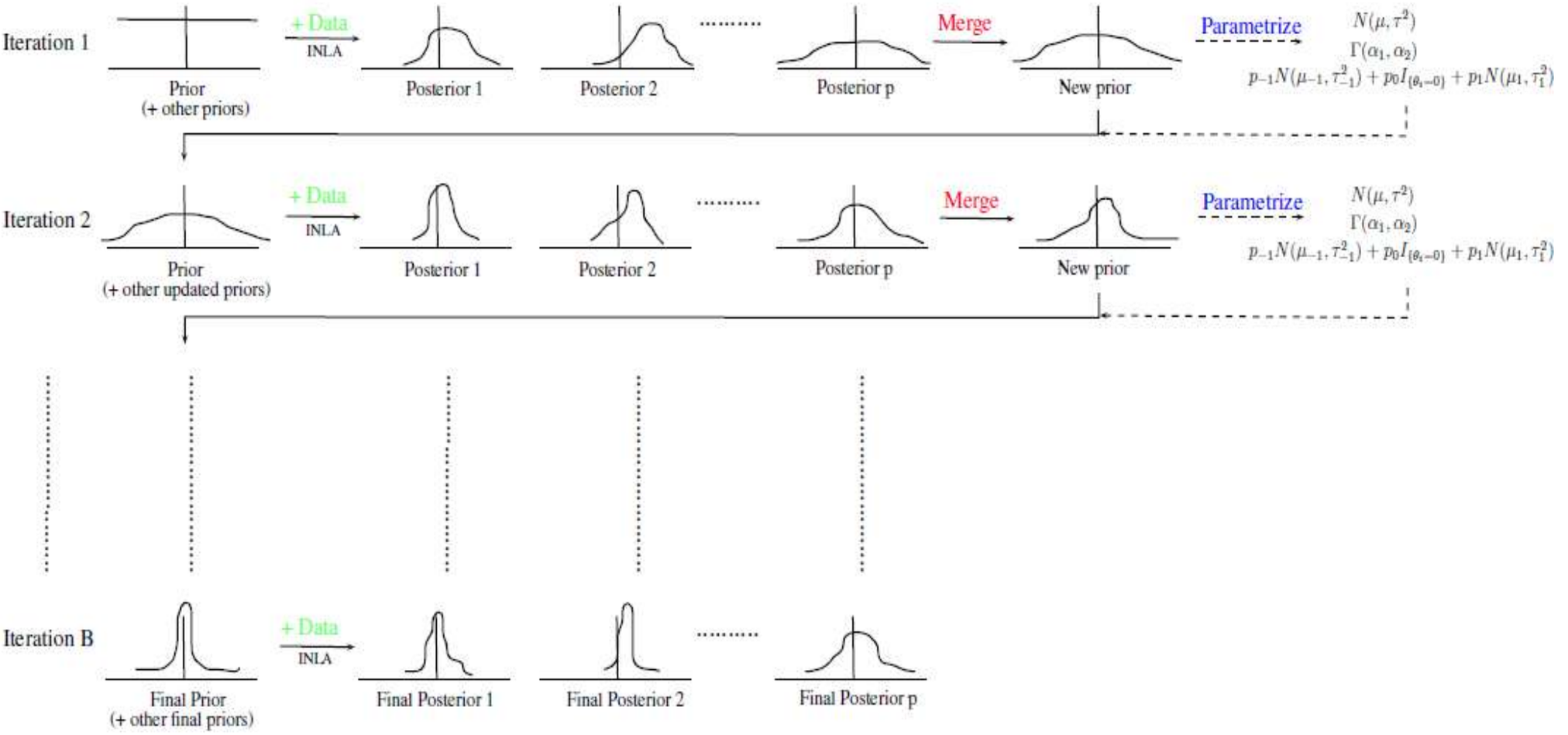
$\boldsymbol{\phi}$ - hyper-parameters

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

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

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

Prior parameter estimation

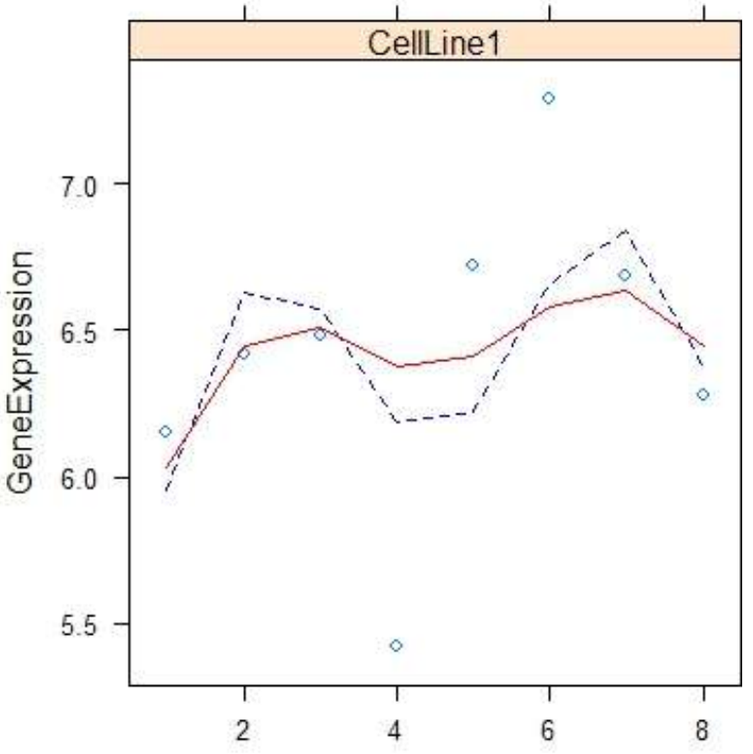


van de Wiel et al. (2013), Biostatistics.

Shrinkage

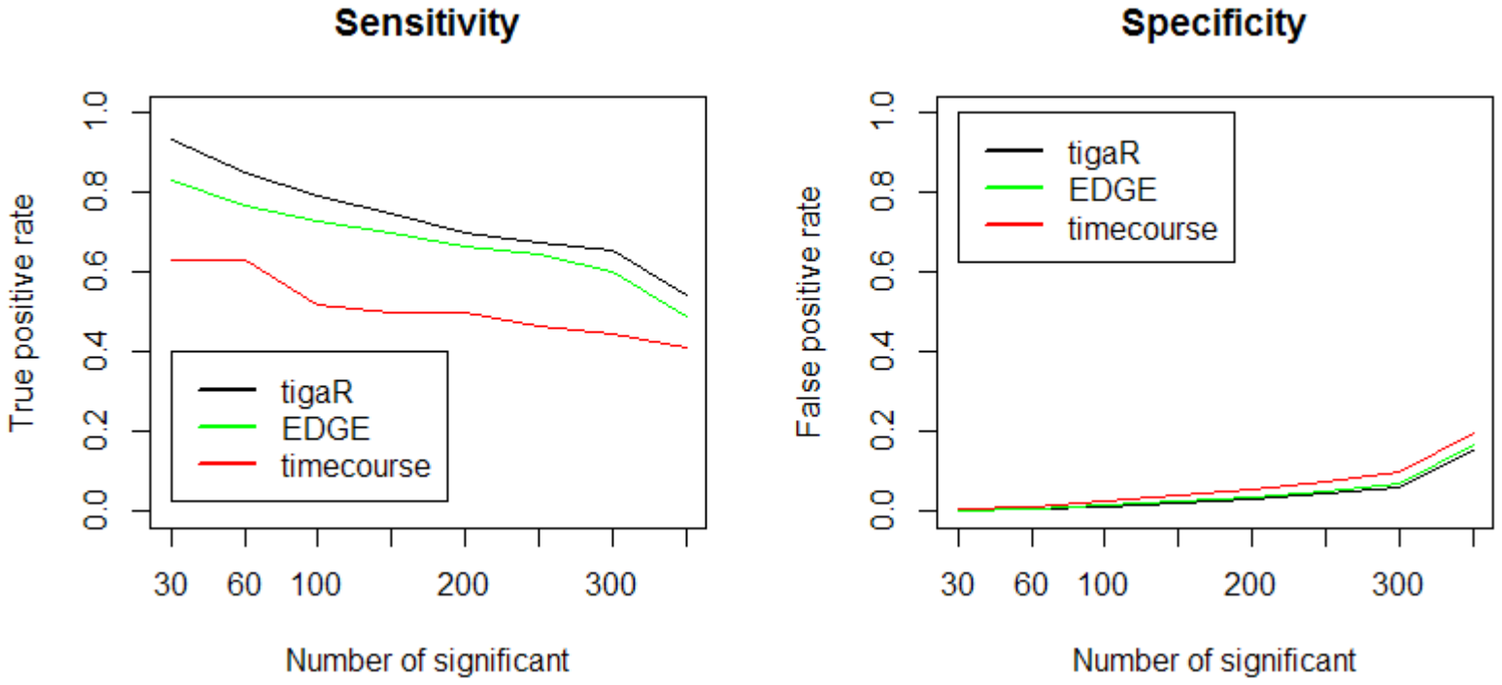
$$\sum_{i=1}^{n}\sum_{t=1}^{m}\left[y_{ijt} - f_{ij}(v_i;\boldsymbol{\alpha}) - h_j(t;\boldsymbol{\beta})\right]^2 + \lambda\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}, \text{ where } \lambda = \frac{\sigma_{\boldsymbol{\epsilon}}^2}{\sigma_{\boldsymbol{\beta}}^2}$$

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

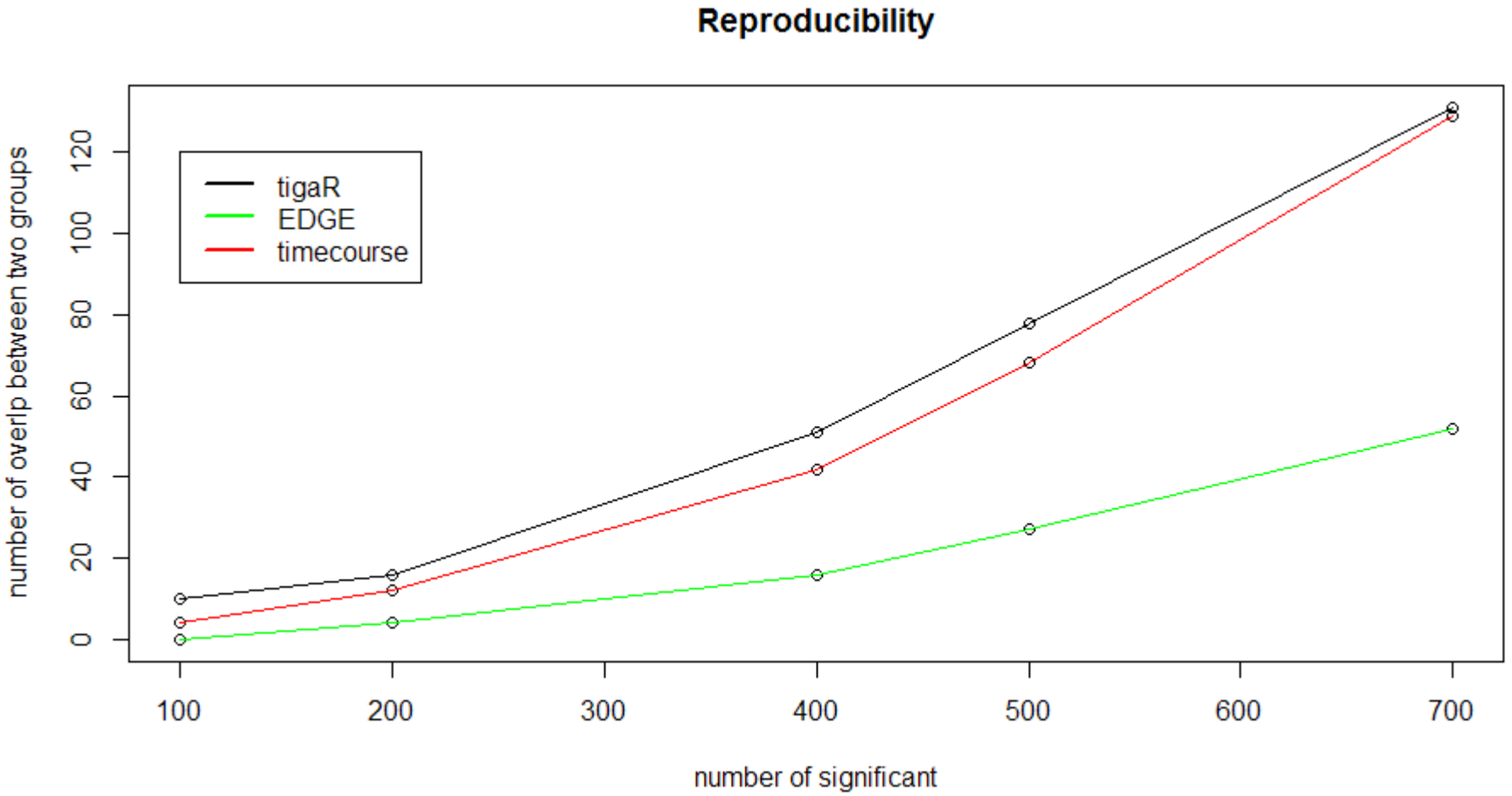


Sensitivity and specificity

- Comparison of methods
  - **tigaR** – temporal integrative genomic analysis in R
  - **EDGE** – Storey et al., PNAS., 2005.
  - **timecourse** – Tai and Speed, Annals of Statistics, 2006.



Reproducibility

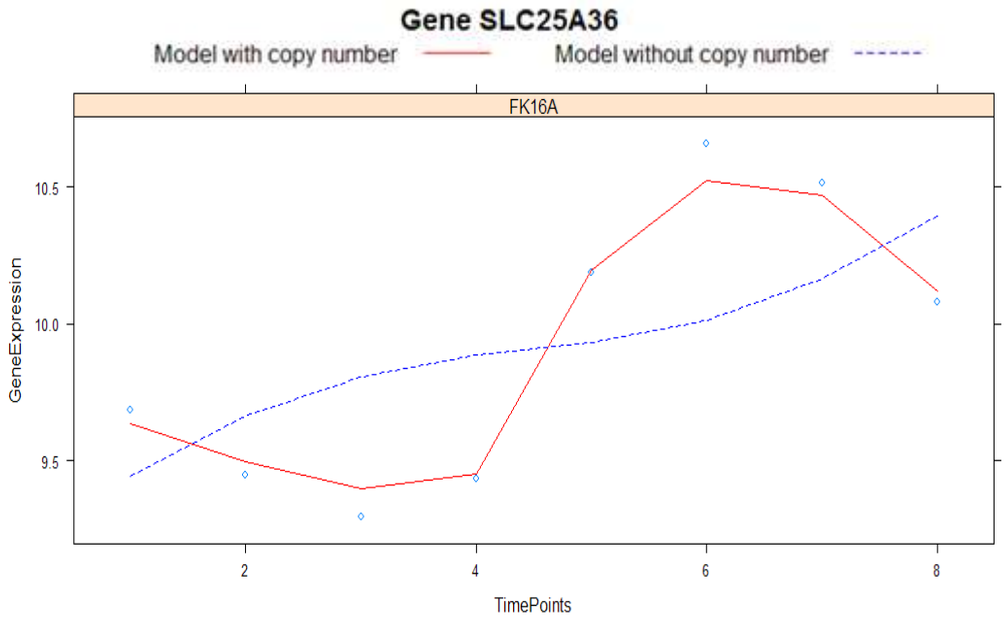


DNA copy number (CN)

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

Cell line effect **CN effect**

Fixed effect: $f_{ij}(v_i, x_{ijt}; \boldsymbol{\alpha}, \gamma) = \alpha_{ij} v_i + \gamma_j x_{ijt}$



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

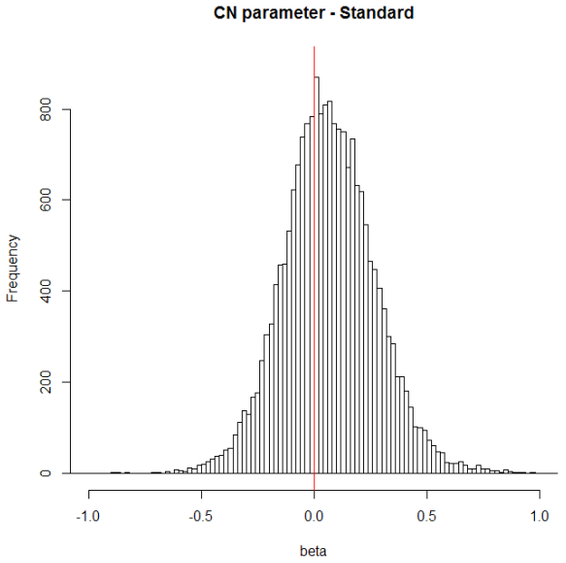
Orthogonalization of splines

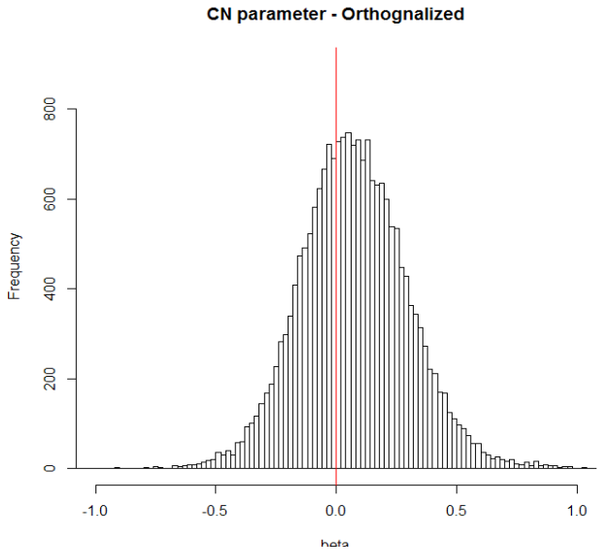
- Time effect consumes CN effect.
- Attribute effect to biological factor rather than to dynamic factor.
- Orthogonalize splines design matrix to CN.

$$\tilde{Z}^{\perp} = \left(I - X(X^T X)^{-1} X^T \right) \tilde{Z}$$

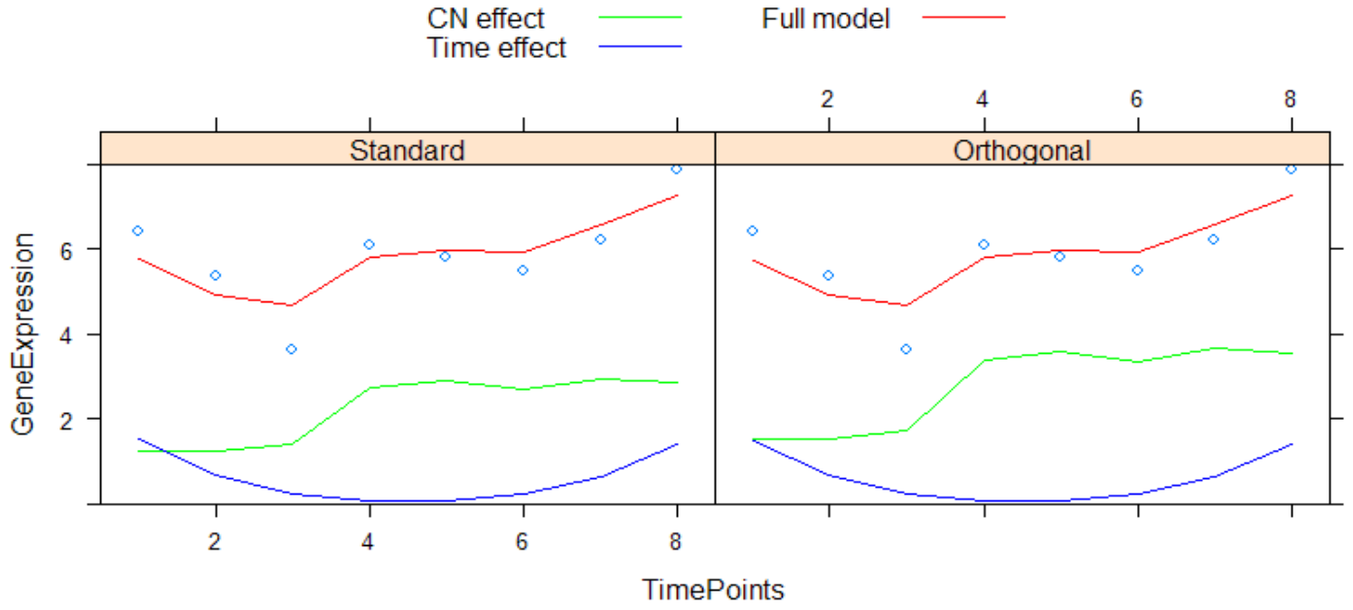
Effect of orthogonalization

Parameter of CN:





Standard vs. orthogonal:

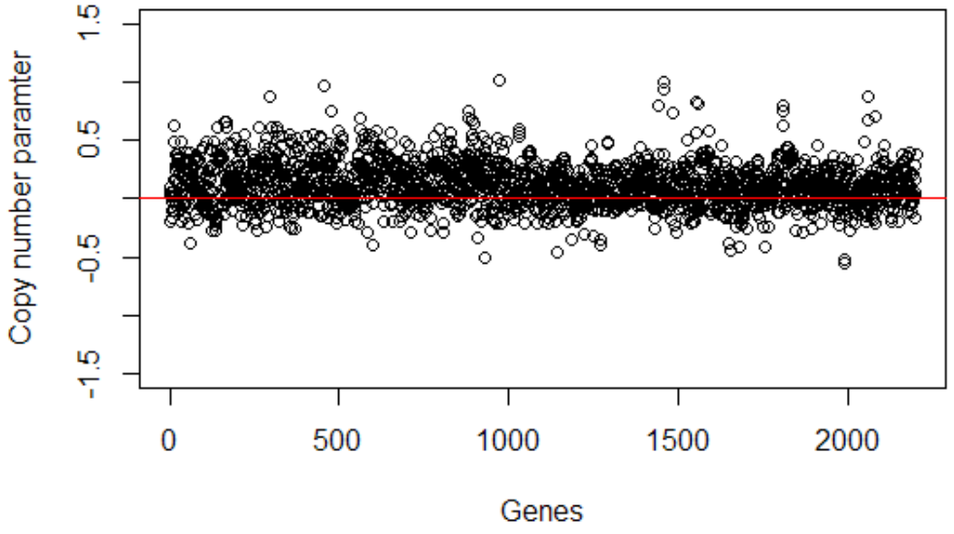


Spatial multivariate prior for CN

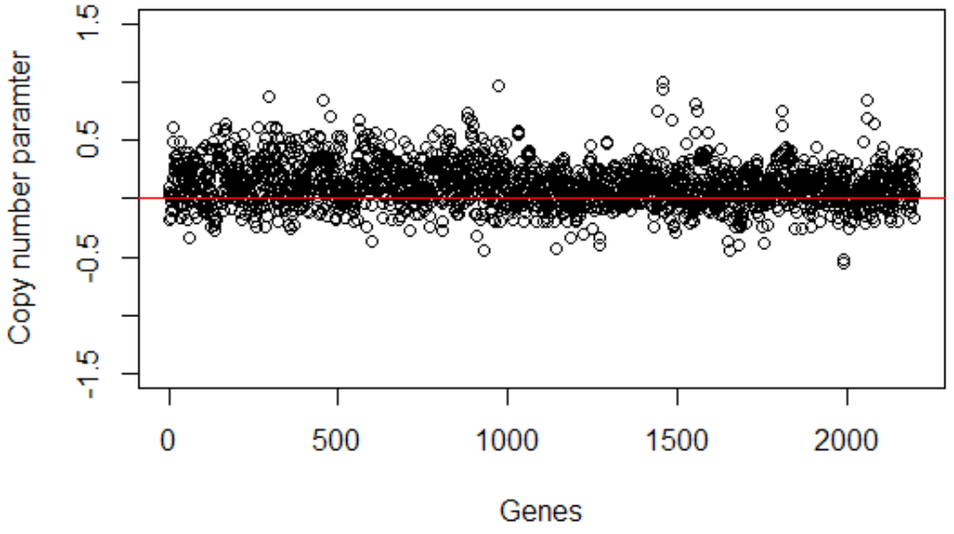
Multivariate prior: $\gamma|\sigma^2_\gamma \sim \mathcal{N}_3(\mathbf{0}, \mathbf{\Sigma}\sigma^2_\gamma)$ $\quad \mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$

- ϕ - hyper-parameters estimated univariate
- ρ - correlation estimated trivariate
- θ - model parameters estimated multivariate per triplets

Univariate parameter estimation

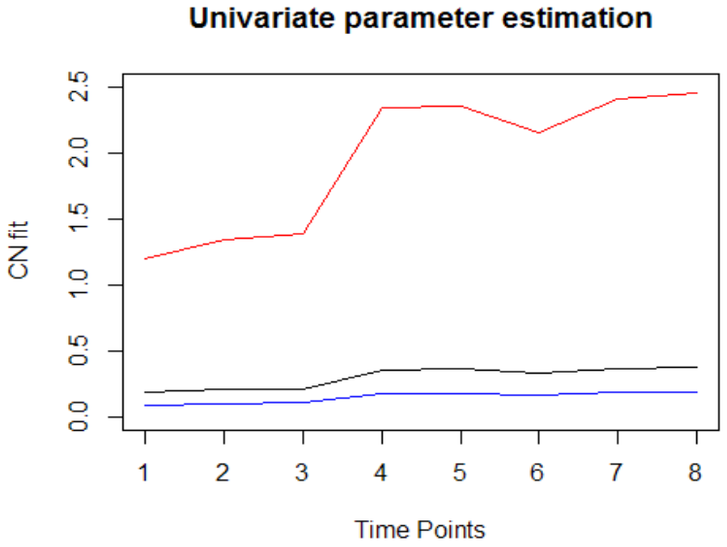


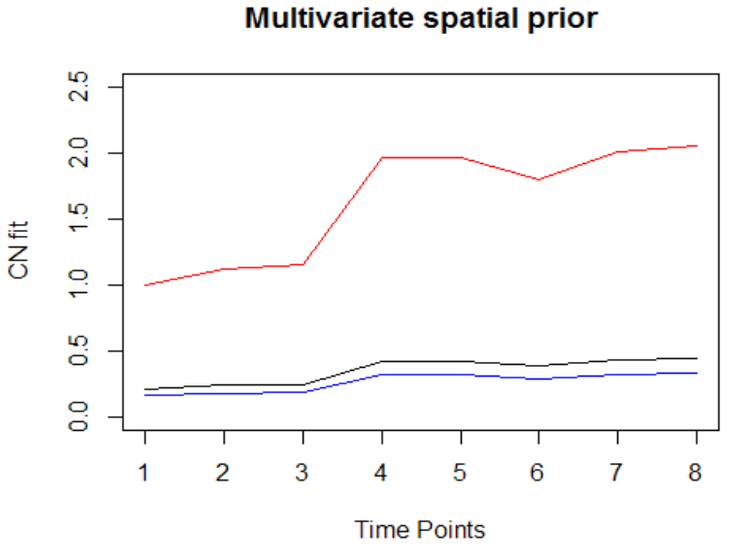
Multivariate spatial prior



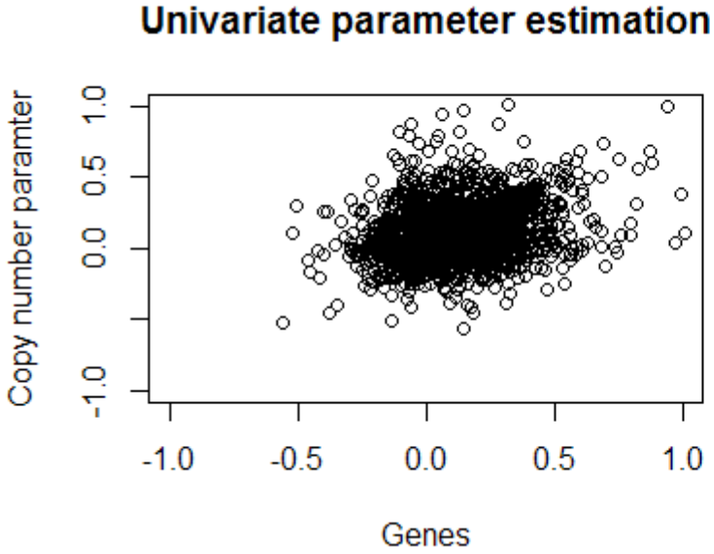
Effect of multivariate spatial prior

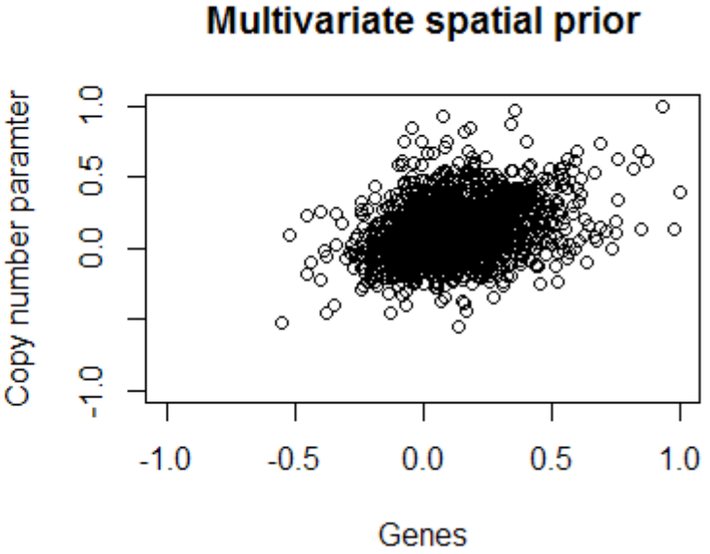
Triplet:





Partial correlation:





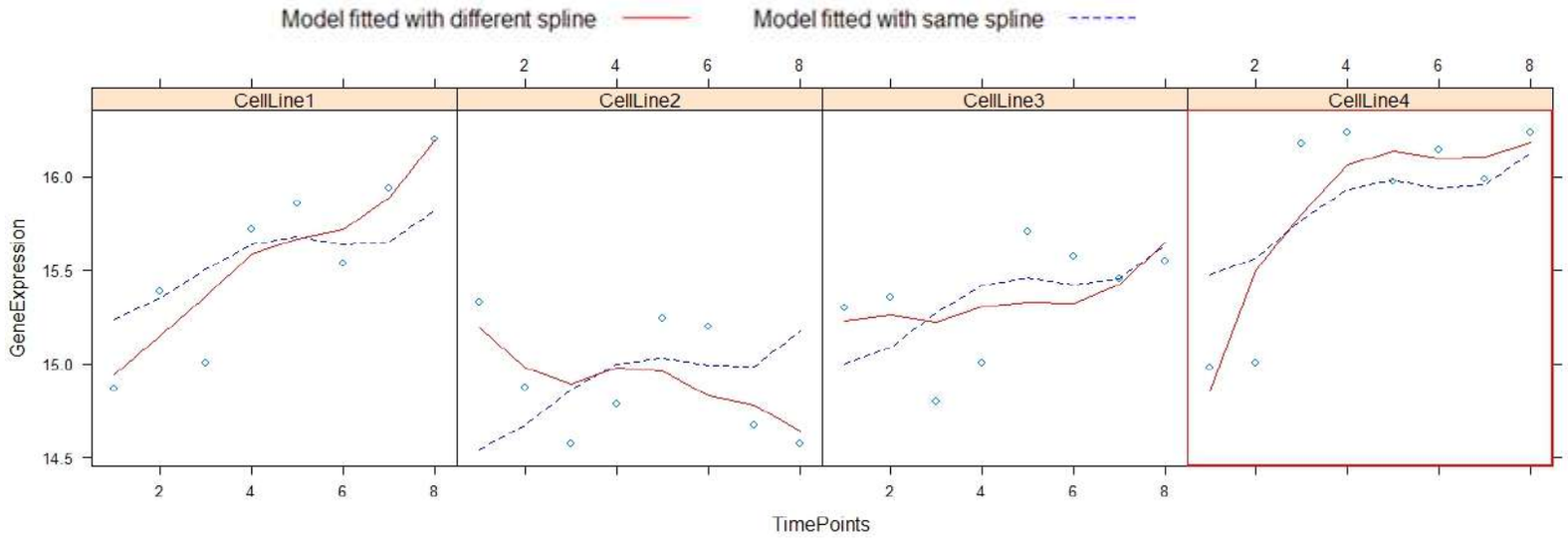
Same and different splines

Same spline – identify up or down regulated genes

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

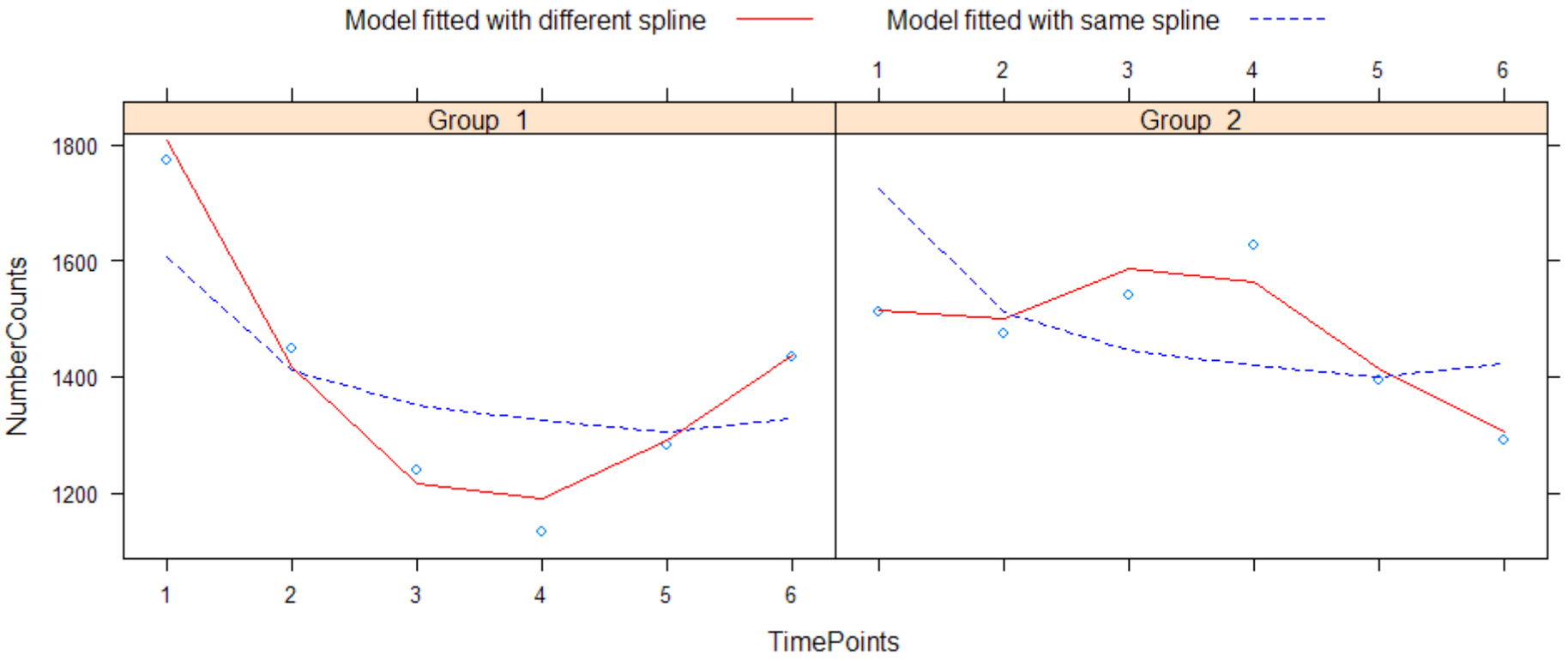
Different spline – allow more flexibility

$$\tilde{Z} = \tilde{Z} \otimes I_{n \times n}$$



RNA-seq data

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

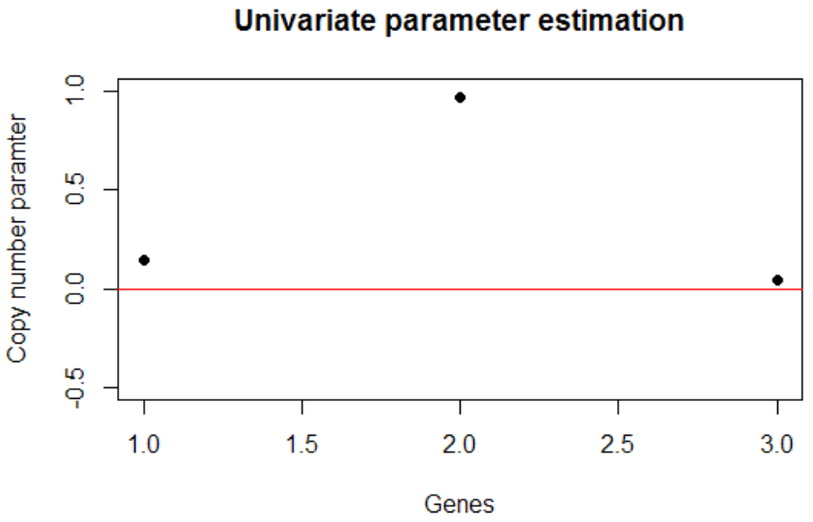


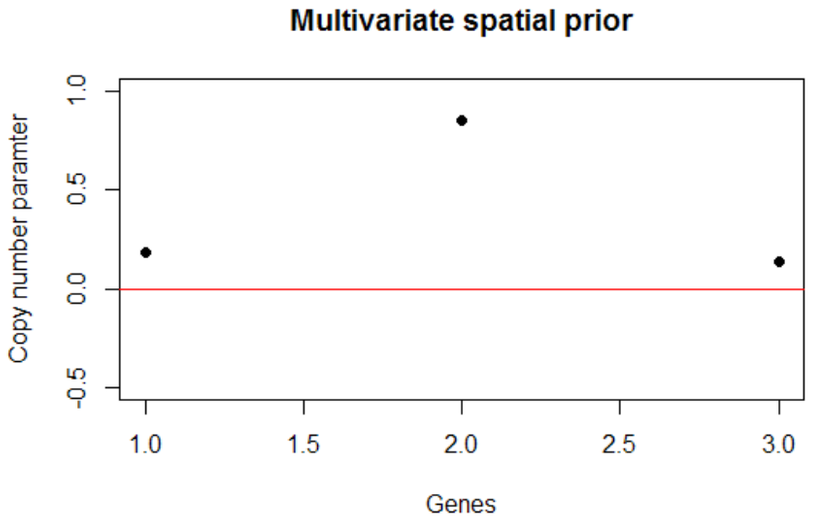
Summary

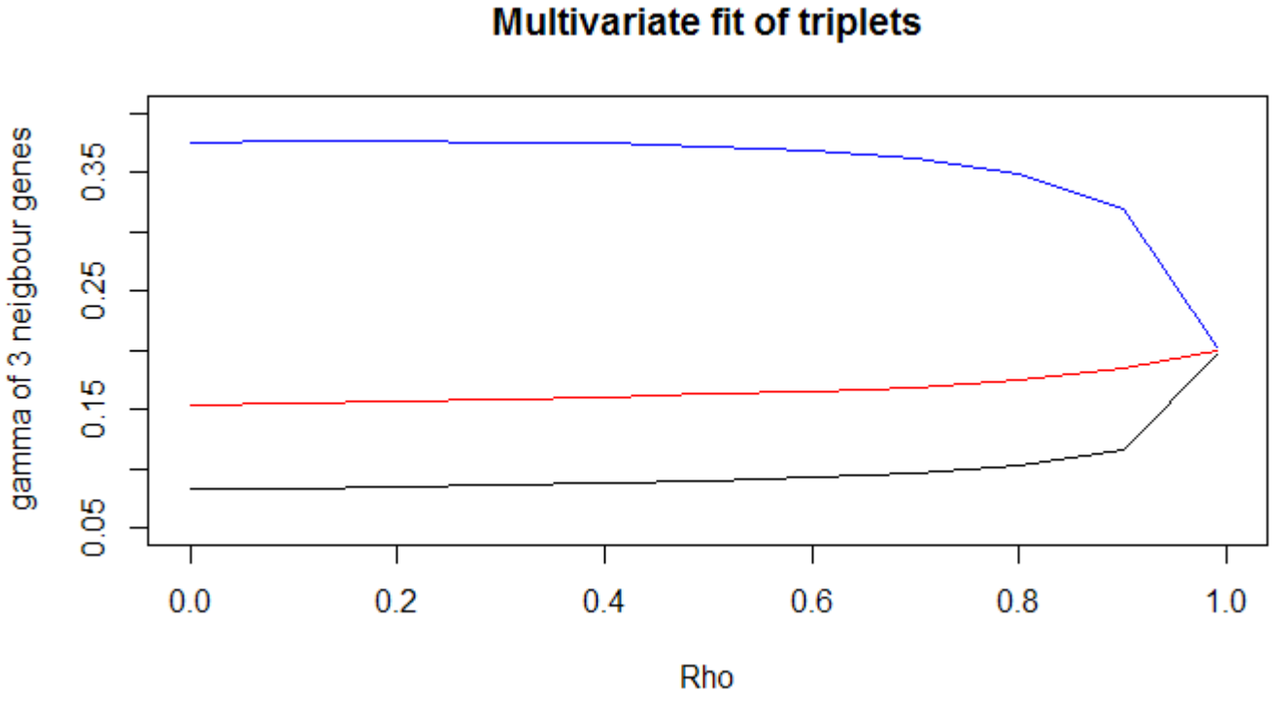
- Improved identification of temporal differential gene expression (TDGE) using penalized splines and empirical Bayes shrinkage.
- Identification of TDGE induced by CN.
- Identification of TDGE in count RNA-seq data.
- Improvement of CN estimates, with orhogonalization and imposing spatial multivariate prior.
- Identification of significant up or down regulated genes.
- As a proof of principle gene **SLC25A36** and **CADM1** are identified.

Thank you for your attention!

Triplet:


Univariate parameter estimation
Copy number paramter
-0.5
0.0
0.5
1.0
1.0
1.5
2.0
2.5
3.0
Genes


Multivariate spatial prior
Copy number paramter
-0.5
0.0
0.5
1.0
1.0
1.5
2.0
2.5
3.0
Genes


Multivariate fit of triplets
gamma of 3 neigbour genes
0.05
0.15
0.25
0.35
0.0
0.2
0.4
0.6
0.8
1.0
Rho