1 Application

The use of the proposed methodology is demonstrated on data from the aforementioned HPV-induced transformation study. Primary interest is unravelling temporal and contemporaneous relations among the genes. Analysis and results, alongside with several down-stream explorations, are discussed and presented.

The data stem from an *in vitro* HPV-immortalized cell line experiment. The employed cell line model faithfully mimics cervical cancer development, morphologically and (epi)genetically (Steenbergen et al. [2004]; Wilting et al. [2006]; Henken et al. [2007]). The experiment comprise four cell lines, two with HPV16 and two with HPV18 (Steenbergen et al. [1996]). Over time the cell lines undergo transcriptomic changes. Using oligonucleotide microarrays cell lines are assayed at eight sequential time points in order to assess these changes in mRNA levels. The preprocessing of the resulting gene expression data comprises background correction (using the robust multi-array (RMA) approach of Irizarry et al. [2003]) and between-array normalization (using the robust quantile method proposed by Boldstad et al. [2003]). Next, a variance stabilizing transformation (Huber et al. [2002]) is applied. Then, data of probes interrogating genes mapping to the p53 signalling pathway (as defined by the KEGG repository, Kanehisa and Goto [2000]). Data from probes corresponding to the same gene are averaged. The resulting data set comprises $p = 64$ genes, measured at $\mathcal{T} = 8$ time points in $n = 4$ cell lines.

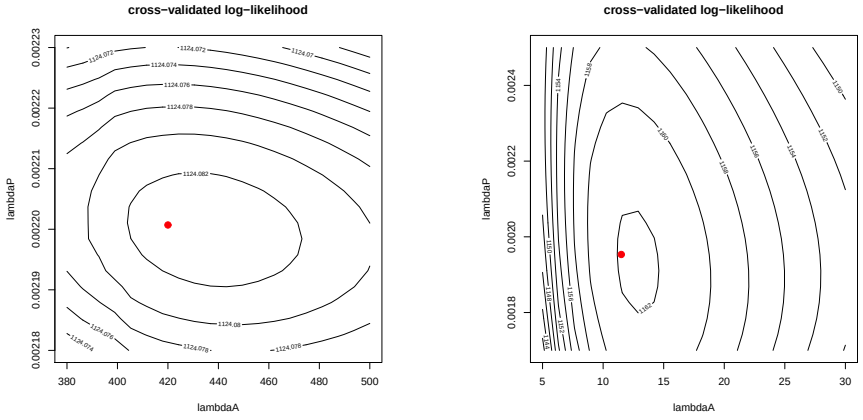

Figure 1: Contour plots of LOOCV log-likelihood of VAR(1) model for p53 signaling pathway. Left panel represent the contour plot of the penalty parameter in the first estimation, while in the right panel the contour plot of the re-estimated parameters

Model parameters are estimated as outlined in the Section 3. First, optimal penalty parameters are determined through maximization of the LOOCV log-likelihood, resulting in $\lambda_a = 420.0000$, $\lambda_\omega = 0.0022$. Figure 1 shows the contour plot of the LOOCV log-likelihood of VAR(1) model. The red dot in Figure 1 represents the penalty parameter choice. According the contour plot the red dot is close to the maximum, implying the employed optimization procedure has converged. With these optimal penalty parameters the ridge penalized

maximum likelihood estimates of $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$ are obtained. These estimates are illustrated as heatmaps in Figure 2.

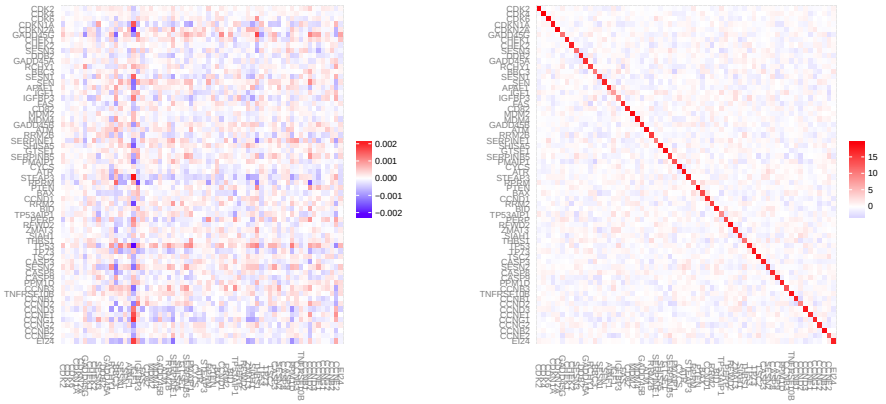

Figure 2: Heat maps of the estimated parameters for p53 signaling pathway using proposed method. Left panel represent estimates for $\mathbf{A}$, while right panel for $\mathbf{\Omega}_{\boldsymbol{\varepsilon}}$. Each element of the heatmap represents, an edge, with intensity of positive or negative effect between two genes. In biological contexts positive(negative) effect mean that one gene activate(repress) another gene.

To facilitate biological insight genes with the strongest temporal and contemporaneous relations are singled out through post-estimation support determination of $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$ (confer Section 4.2). For the selection of nonzero elements, corresponding to edges in the time series chain graph, of $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$ a cut-off of $1 - \widehat{\text{lFDR}} \geq 0.999$ is applied. This high cut-off results in a sparse and interpretable graph. Figure 3 displays the selected edges as a heatmap. Red squares in the heatmap represent (significant) edges between two genes. The heatmap reveals the presence of genes with either a horizontal or vertical red stripy pattern. Genes (e.g. CGKN2A, TP53,TP73, CGKN1A, GADD45G) with the horizontal stripes can be thought of as the 'regulatees' in the pathway, while the vertical ones (e.g. IGF1, IGFBP3, BBC3, CCND2, THBS1) are the 'regulators' of the pathway. This is confirmed when studying node statistics (like in- and out-degree, betweenness, centrality, confer Newman for definition (Newman [2010])) derived from the inferred graph (Table 1).

With knowledge of their support less biased estimates of $\mathbf{A}$ and $\mathbf{\Omega}_\varepsilon$ are obtained by a refitting them taking the support into account. Optimal penalty parameters are re-determined: $\lambda_a = 11.4000$, $\lambda_\omega = 0.0019$ (and confirmed by the contour plot the LOOCV log-likelihood, Figure 1). The re-estimated parameter $\mathbf{A}$ is visualized as a heatmap (Figure 4). The temporal relations among the genes are also visualized in form of a graph (confer Figure 5), where edges represent the effect of the expression levels of a gene at timepoint t to that of another genes at time point $t + 1$.

Employing the re-estimated parameter $\mathbf{A}$, we study the fit ($\hat{\mathbf{Y}}_{*,i,t} = \hat{\mathbf{A}}\mathbf{Y}_{*,i,t-1}$) of the 'regulatees' (genes explained by other genes). Figure 6 shows the expression levels of the up-regulated gene TP73 along with its fit (red line) in the four cell lines (one panel each). Considering the use of a linear model, the fit

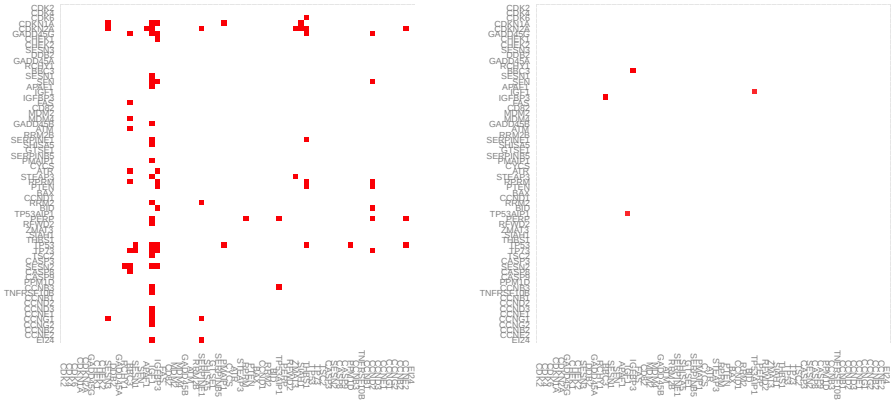

Figure 3: Heat maps of sparsified parameter estimates for p53 signaling pathway using proposed method. Left panel represent estimates for $\mathbf{A}$, while right panel for $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}$.

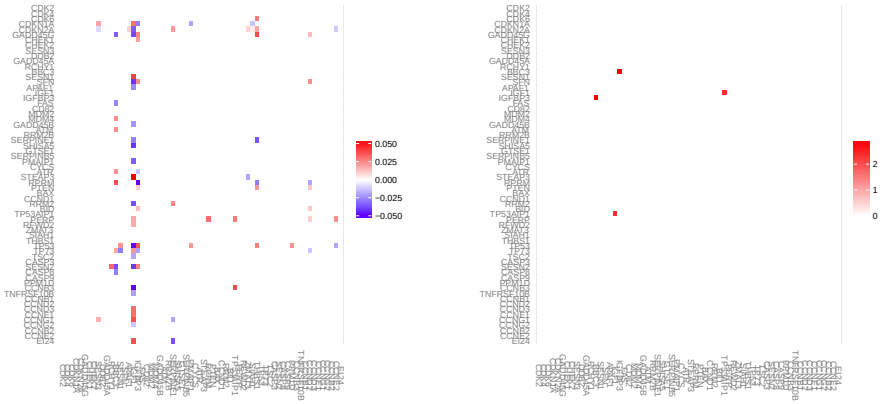

Figure 4: Heat maps of re-fitted parameter estimates for p53 signaling pathway using proposed method incorporating the prior knowledge on the support on $\mathbf{A}$.

captures the salient features of the data. The fit is studied for all genes in the pathway. The result is summarized in Figure 7. It displays the histogram of the Spearman correlations between the fit and the observations, cell line-wise. The histogram in Figure 7 is clearly skewed to the domain $[0, 1]$. This indicates that the fit is generally reasonable.

With final and less-biased estimates of the VAR(1) parameters at hand, we study the quantitative, dynamic implications of the model: what are the downstream effects of a change in expression levels of a gene? This can be done through impulse response analysis (confer Section 4.2 for details). For each gene the column-wise average of the (absolute) impulse response on all other genes at the next time instance is calculated (Table 1). This is a measure of a gene's driving force on the pathway's expression levels. The low and high impulse

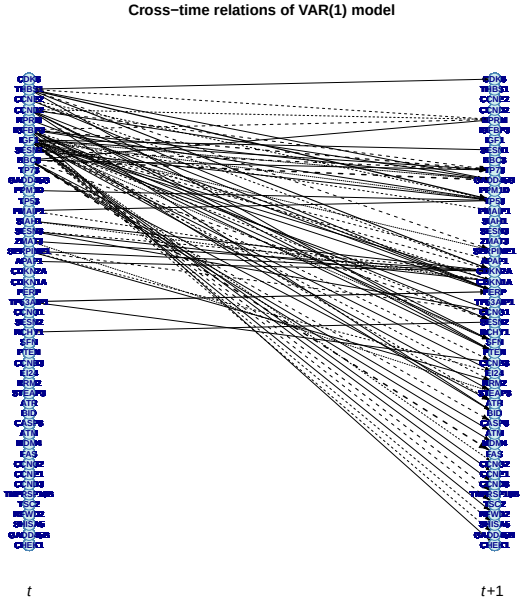

Figure 5: Graphs of the temporal relations among the genes as implied by the parameter estimates of $\mathbf{A}$ for p53 signaling pathway. Solid line represent positive effect, while dashed lines negative effect. Thickness of the lines indicate how strong is the effect. For the sake of the graph clarity, genes which are not connected with edge to another genes are omitted.

responses of the { CGKN2A, TP53,TP73, CGKN1A, GADD45G } and { IGF1, IGFBP3, BBC3, CCND2, THBS1 } genes corroborate with their interpretations of 'regulatees' and 'regulators'. This is supported when evaluating the mutual information between each gene at time point t and the whole pathway at next time point (Table 1).

The downstream effects of a signal may be further elucidated through the decomposition of the covariance between the expression levels of two genes in terms of the paths connecting them in the time-series chain graph (as described in Section 4.3). For illustration purposes consider the regulator-regulatee pair (IGFBP3, TP73) at two contiguous time points. They are connect through two paths: a direct path ($Y_{\text{IGFBP3},t} \rightarrow Y_{\text{TP73},t+1}$) and an indirect one through $Y_{\text{IGFBP3},t} \rightarrow Y_{\text{BBC3},t} \rightarrow Y_{\text{TP73},t+1}$ (depicted in Figure 8). The covariance the (IGFBP3, TP73) gene pair at contiguous time point may now decomposed as:

$$\begin{aligned} \text{Cov}(Y_{\text{IGFBP3,t}}, Y_{\text{TP73,t+1}}) &= (\boldsymbol{\Sigma}_{\varepsilon}\mathbf{A}^{\top})_{\text{IGFBP3,BBC3}} \\ &= -0.003168001 = -0.002483485 - 0.0006845158, \end{aligned}$$

in which the two summands on the right-hand side correspond the direct and indirect path in the time-series chain graph. Based on the paths' contribution the direct one dominates, but being of the same sign the indirect path also contributes in the suppression of the expression levels of the TP73 gene.

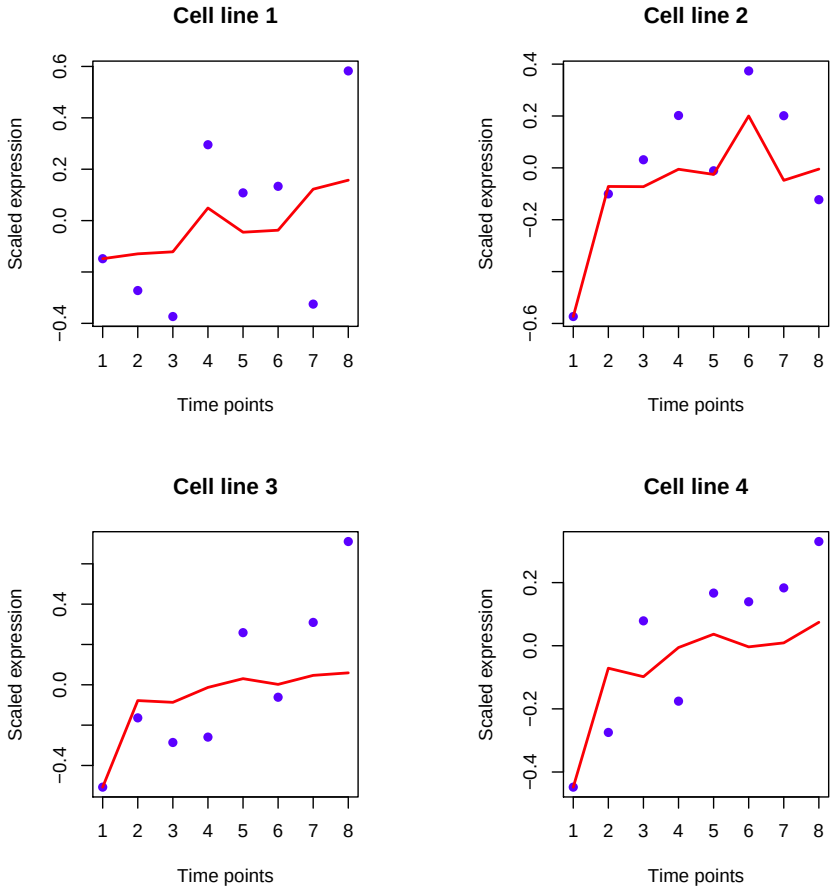

Figure 6: Fit of the expression levels of TP73 over time in all cell lines. Dots represent the expression levels, while solid red lines indicated the fit of the model over time.

Another way to grasp the inferred networks of the p53 signaling pathway is to study its motifs. Motifs are small recurring network patterns and form the building blocks of pathways (Alon [2007]). As the inferred network is very sparse only few (three gene) motifs are found (Figure 8). All these motifs are so-called feedforward loops (FFL), which appear in many gene systems (Alon [2007]). Identified FFL motifs are all of the incoherent type, which connect two genes via two paths that have opposite effects: positive (activating) and negative (repressing). A FFL motif found in the reconstructed time-series chain graph of the p53 signaling pathway is shown in Figure 8. Here, gene IGF1 activates both SESN1 and TP73, while SESN1 represses TP73 gene. IGF1 and SESN1 thus affect TP73 in opposite ways. In the extreme case, when the effect of the SESN1 is equal to that of IGF1, this results (with a slight delay due to the time) in the repression of the TP73, reducing its expression levels to (virtually) zero (confer Alon [2007]). To contrast this, a coherent FFL (e.g. SESN1 affect TP73 in a positive manner) with increased effects in the IGF1 and SESN1 gene,

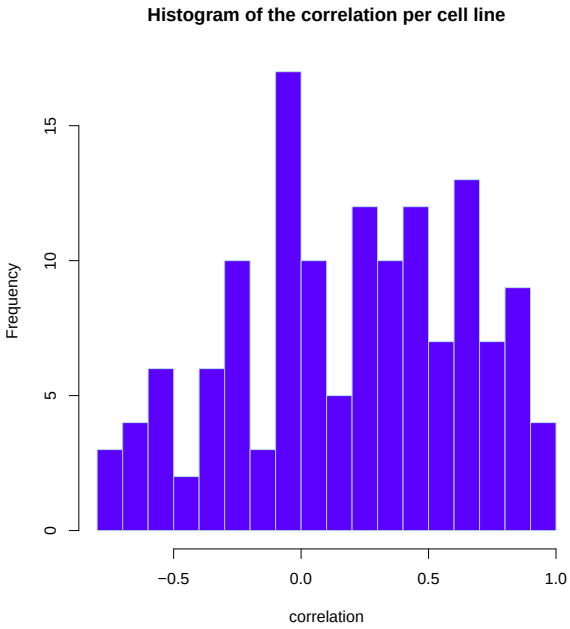

Figure 7: Histogram of the correlation between model fit and observation of data set from the p53 signaling pathway.

propagate delayed amplification of the signal in the TP73 (for more details see SM).

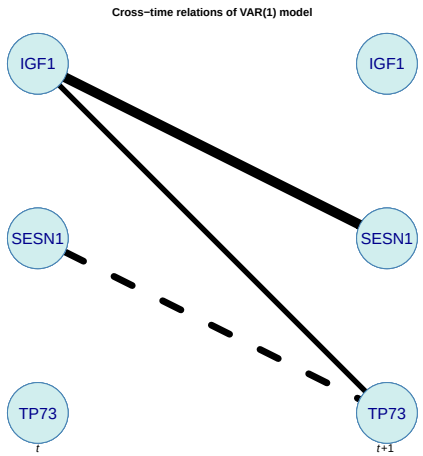

Figure 8: Motifs of p53 signaling pathways.

Table 1: Gene statistics for the top genes(in terms of the number of edges) for p53 signaling pathway; in-degree of A - number of (temporal) edges pointing to each gene; out-degree of A - number of (temporal) edges leaving each gene; centrality - eigenvector centralities of the gene within the graph; impulse response - absolute impulse response in the first time point; mutual information - mutual information in the first time point

	in-deg.	**out-deg.**	**cent.**	**imp. resp.**	**mut. inf.**
IGF1	0	25	1.0000	0.01245	0.0124
IGFBP3	0	11	0.7555	0.0039	0.0027
BBC3	0	9	0.4502	0.0037	0.0034
CCND2	0	7	0.5353	0.0015	0.0010
THBS1	0	7	0.7892	0.0030	0.0044
CDKN2A	8	0	0.3993	0	0
TP53	7	0	0.2595	0	0
TP73	5	0	0.2595	0	0
CDKN1A	5	0	0.2714	0	0
GADD45G	5	0	0.2820	0	0

References

Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

BM Boldstad, RA Irizarry, M Astrand, and TP Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19(2):185–193, 2003.

F E Henken, S M Wilting, R M Overmeer, J G Van Rietschoten, A O H Nygren, A Errami, J P Schouten, C J L M Meijer, P J F Snijders, and R D M Steenbergen. Sequential gene promoter methylation during HPV-induced cervical carcinogenesis. *British Journal of Cancer*, 97:1457–1464, 2007.

Wolfgang Huber, Anja Von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl 1):S96–S104, 2002.

Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, Terence P Speed, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

R D M Steenbergen, D Kramer, B J M Braakhuis, P L Stern, R H M Verheijen, C J L M Meijer, and P J F Snijders. TSLC1 gene silencing in cervical cancer cell lines and cervical neoplasia. *Journal of the National Cancer Institute*, 96: 294–305, 2004.

RD Steenbergen, JM Walboomers, CJ Meijer, EM Van Der Raaij-Helmer, Jacqueline N Parker, Louise T Chow, Thomas R Broker, and PJ Snijders. Transition of human papillomavirus type 16 and 18 transfected human foreskin keratinocytes towards immortality: activation of telomerase and allele losses at 3p, 10p, 11q and/or 18q. *Oncogene*, 13(6):1249–1257, 1996.

S M Wilting, P J F Snijders, G A Meijer, B Ylstra, P R Van den Ijssel, A M Snijders, D G Albertson, J Coffa, J P Schouten, M A Van de Wiel, C J L M Meijer, and R D M Steenbergen. Increased gene copy number at chromosome 20q are frequent in both squamous cell carcinomas and adenocarcinomas of the cervix. *The Journal of Pathology*, 209:220–230, 2006.