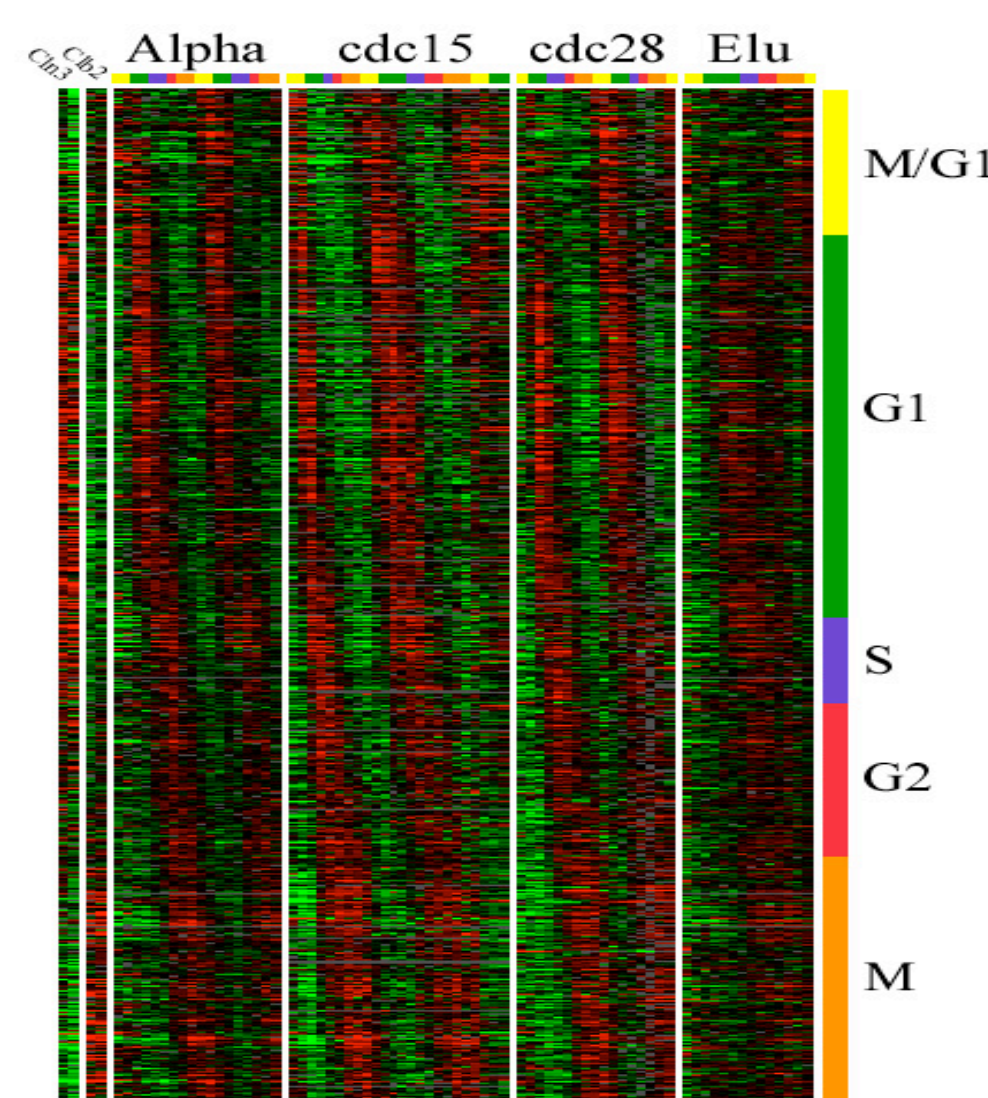




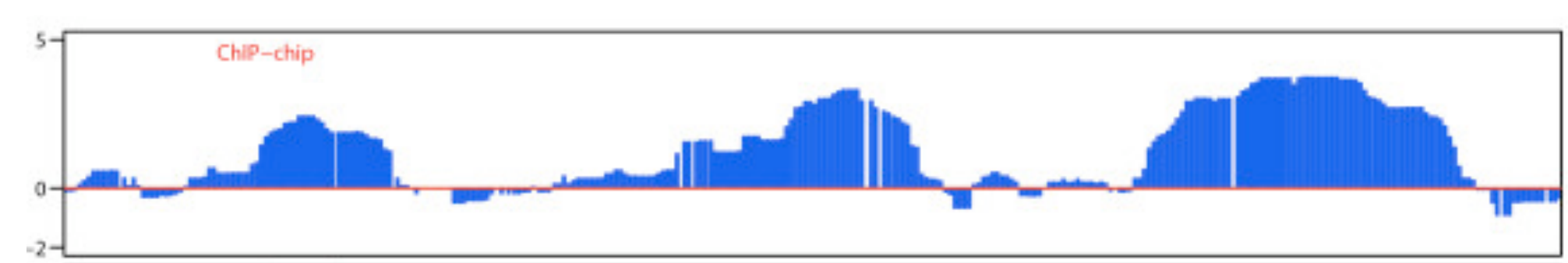
Problem: Variable selection with high dimensional and multi-colinear genomic data.



Motivating data

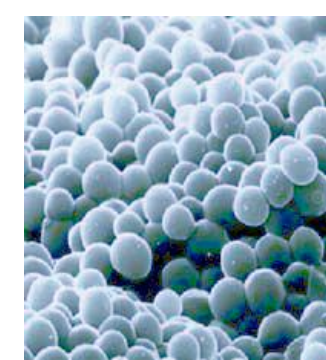
- Yeast cell cycle gene expression data (Spellman et al 1998): mRNA levels of approximately 800 cell cycle related genes from alpha factor based synchronization experiments measured at every 7 mins for 119 mins with total of 18 measurements covering two cell cycle periods.
- Yeast chromatin immunoprecipitation (ChIP-chip) data (Lee et al 2002): binding measurements of 106 TFs which elucidate how these yeast transcriptional regulators bind to promoter sequences of the genes across the genome.

precipitation (ChIP-chip) data (Lee et al 2002): binding measurements of 106 TFs which elucidate how these yeast transcriptional regulators bind to promoter sequences of the genes across the genome.



Aim of the Study

- Activities of TFs are highly **correlated** and only a **subset** of these may be related to cell cycle regulation among 106 TFs.
- Our goal is to **identify cell cycle related TFs** as well as to **infer transcription factor activities (TFA)** of yeast via integrative analysis of expression and binding data.
- We propose a novel regression method: Sparse Partial Least Squares (SPLS) regression. SPLS capitalize on both **dimension reduction** method as well as **variable selection**.



Methods: Partial Least Squares

- Partial Least Squares (PLS) is a dimension reduction method based on a latent component construction.
- X : matrix of covariates.; Y : matrix of responses.
- PLS is a generalization of principal component analysis (PCA) and multiple regression.

- Find orthogonal linear combinations (c_1, c_2, \dots, c_K) of the original predictors that are also relevant to the response: **generalization of PCA**.

$$c_k = \operatorname{argmax}_{\{c \in \Sigma c_j = 0\}_{j=1}^{k-1}, c^T c = 1} \operatorname{cor}^2(Y, Xc) \operatorname{var}(Xc),$$

where Σ is covariance matrix of X .

- Build a multiple linear regression model for Y on constructed orthogonal components: **multiple linear regression**.
- PLS does not promote variable selection and estimate of the direction vector becomes inconsistent with large number of noise variables (Chun and Keleş 2007).

Methods: Sparse Partial Least Squares (SPLS) regression

- Formulation for finding SPLS direction vectors:

$$\begin{aligned} \text{Let } M &= X^T Y Y^T X, \\ \min_{\alpha, c} & -\kappa \alpha^T M \alpha + (1 - \kappa)(c - \alpha)^T M (c - \alpha) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2 \\ \text{s.t.} & \alpha^T \alpha = 1. \end{aligned}$$

- Promote exact zero property by imposing L_1 penalty and direction constraint onto separate vectors while keeping these vectors close to each other.
- L_1 penalty is for imposing sparsity on c , L_2 penalty is for handling potential multi-collinearity that may arise during solving for c .
- α plays the role of the original PLS direction vector, and $\kappa \in (0, 0.5]$ aims to reduce the effect of concavity.

- SPLS solution:

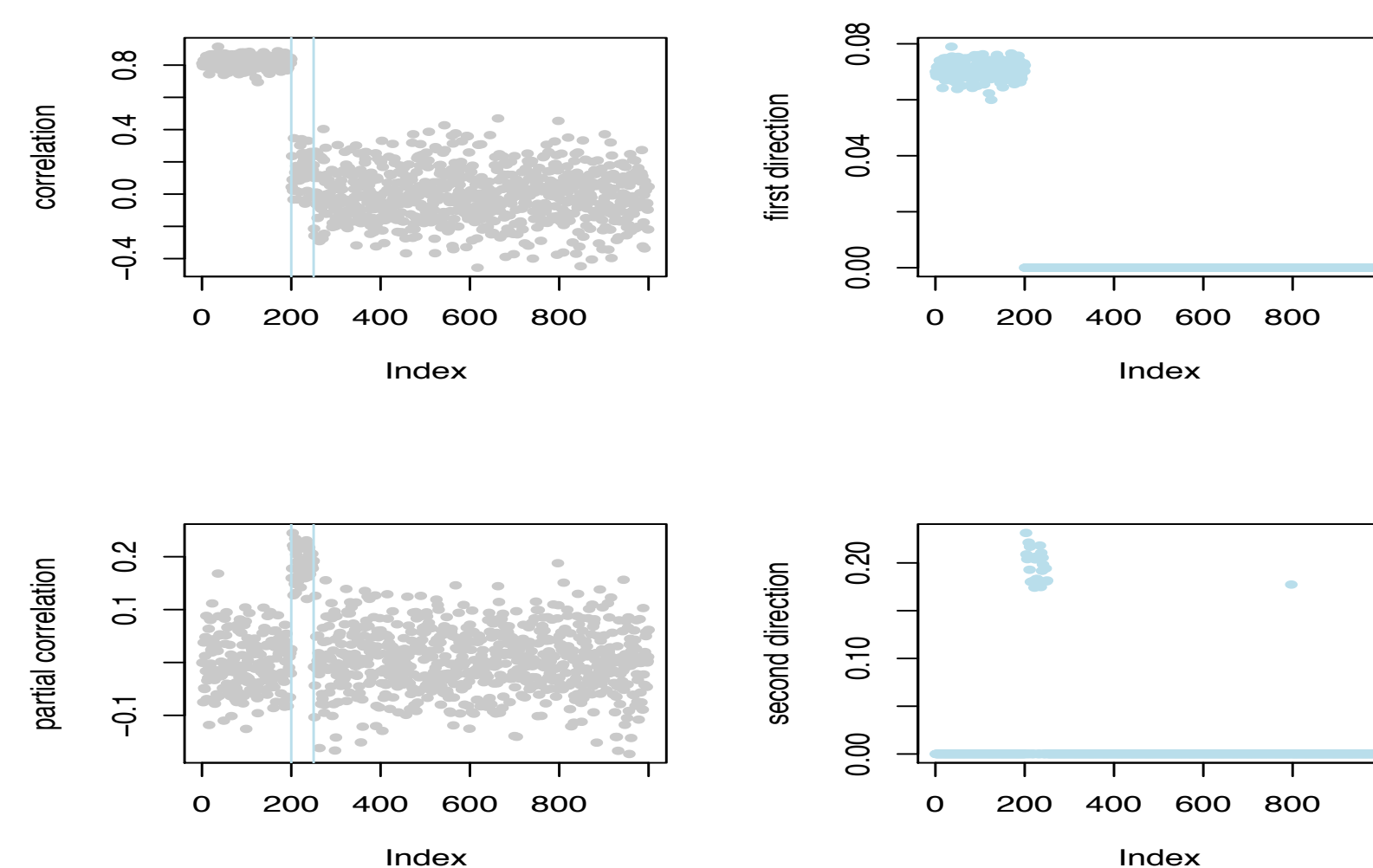
- For univariate Y : the solution can be found by

$$\hat{c} = \left(\frac{Z}{\|Z\|_2} - \frac{\lambda_1}{2} \right)_+ \operatorname{sign}(Z),$$

where $Z = X^T Y$.

- For multivariate Y : the solution can be found by using iterations between soft thresholding (for c) and constrained least squares problem (for α).

Methods: How does SPLS work?

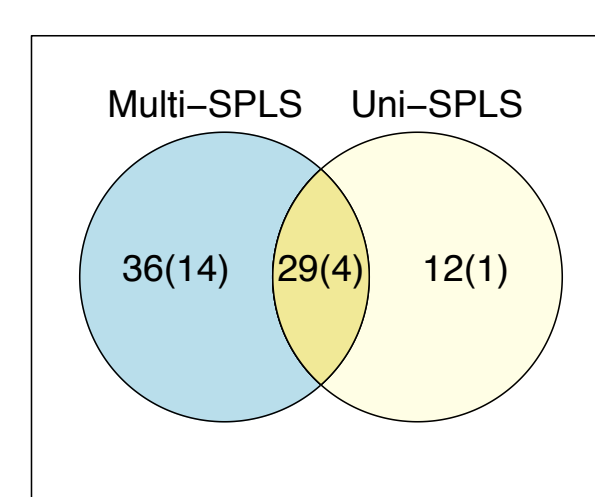


- For univariate Y , it first looks into **correlations** and selects the variables which have the largest correlations to derive first direction vector.
- Next, it looks into **partial correlations** after removing the variations explained by first direction vector, and selects variables which show the largest partial correlations.

Results: # of Selected TFs.

Comparisons of methods: #of selected TFs (# of experimentally verified TFs) are reported.

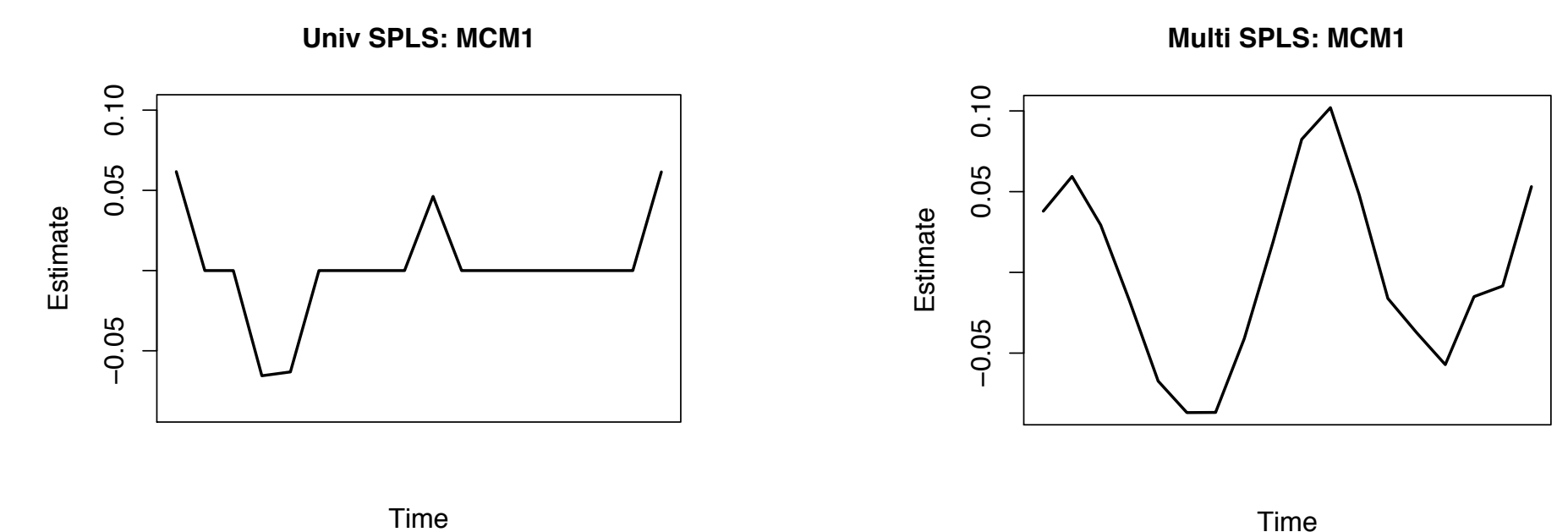
- Multi-SPLS: 48(15)
- Uni-SPLS : 65(18)
- Super PC : 65(18)
- LASSO : 102(18)
- Total : 106(21)



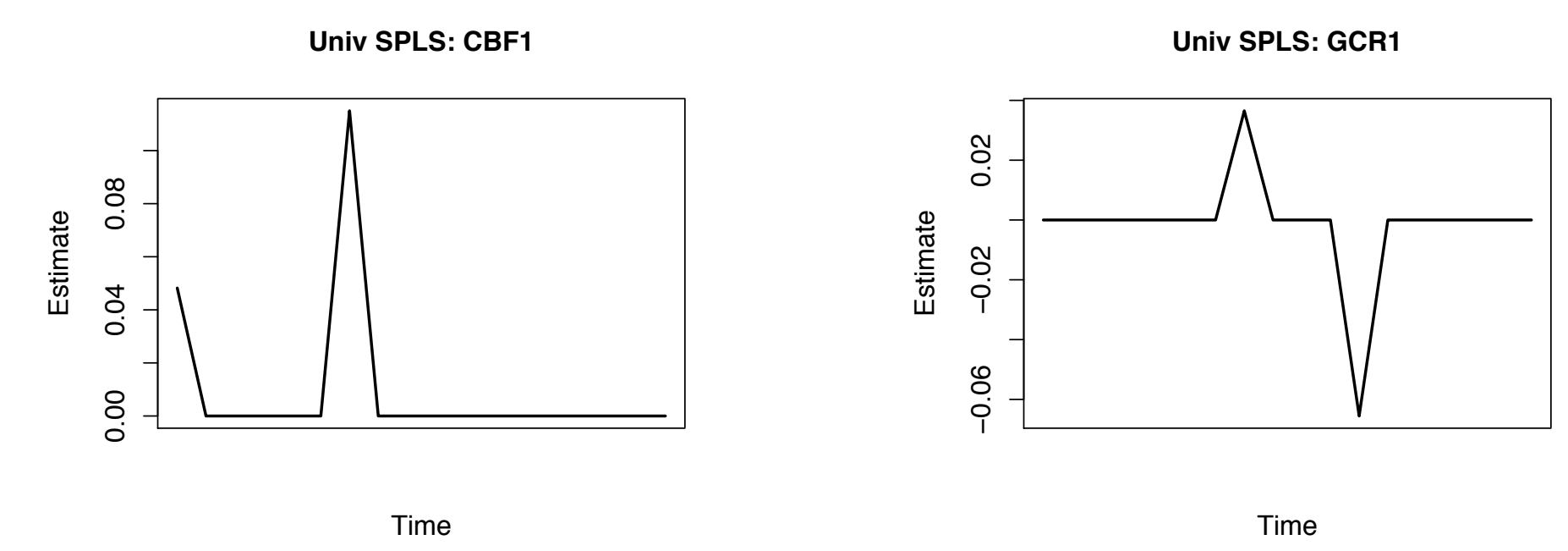
The numbers inside parenthesis indicate the numbers of experimentally verified TFs in the set.

Results: Examples of estimated TFAs

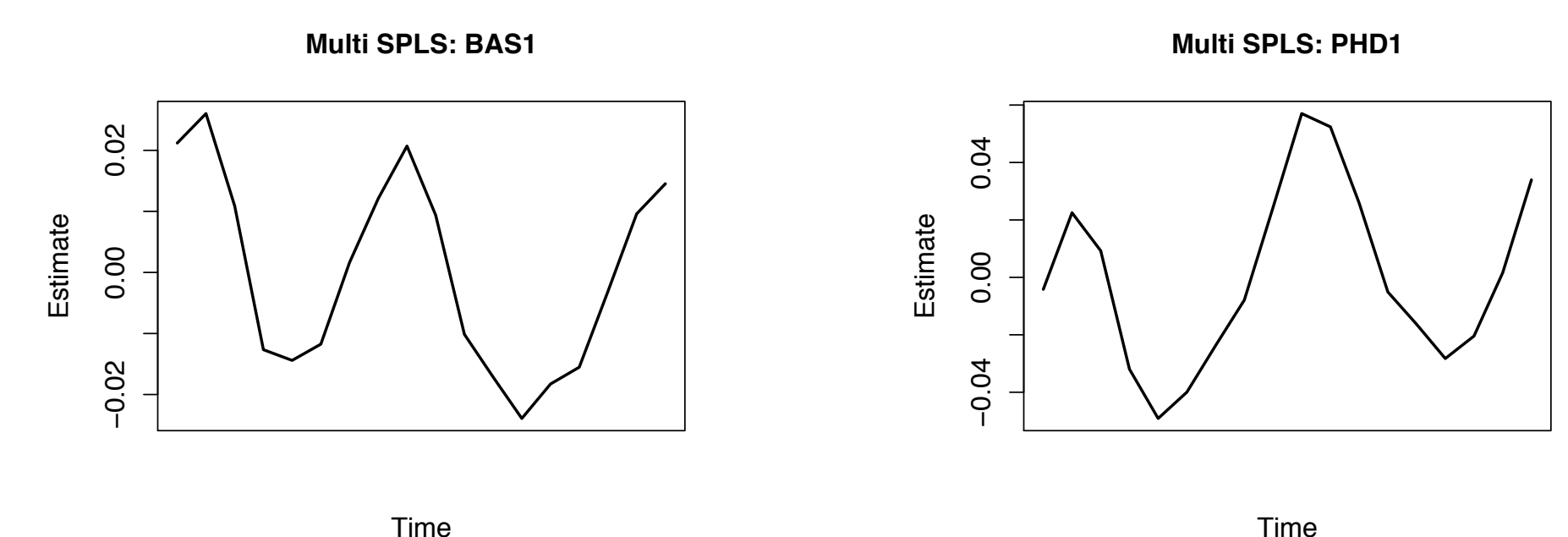
- Factors selected by both methods.
Multivariate SPLS gives smooth and periodic estimates.



- Factors selected only by univariate SPLS.
Univariate SPLS yields non-smooth estimates and does not always exhibit periodicity in the estimates.



- Factors selected only by multivariate SPLS.
Multivariate SPLS can capture small but smooth and periodic coefficients.



Discussion

- SPLS regression is a new methodology based on the sparsity of the PLS direction vector and promotes dimension reduction as well as variable selection.
- It is applicable even when the number of relevant variables are larger than sample size.
- Simulation studies (results not shown) show that SPLS performs well when sample size is small and covariates are correlated.
- Estimated cell cycle related TFAs from multivariate SPLS regression exhibit smoothness and periodicity.

References

Spellman et al., (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* 9: 3273-3297.
Lee et al., (2002). Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.
Chun and Keleş (2007). Sparse Partial Least Squares with an Application to Genome Scale Transcription Factor Analysis. Submitted to JASA.
Source of figures: Spellman et al (1998 Molecular Biology), Shouyong Peng et al (2007 BMC bioinformatics), Budding Yeast Cell Cycle homepage.

Acknowledgement

This research has been supported in part by NIH grants 1-R01-HG03747-01(S.K.), 4-R37-GM038660-20(S.K., H.C.).