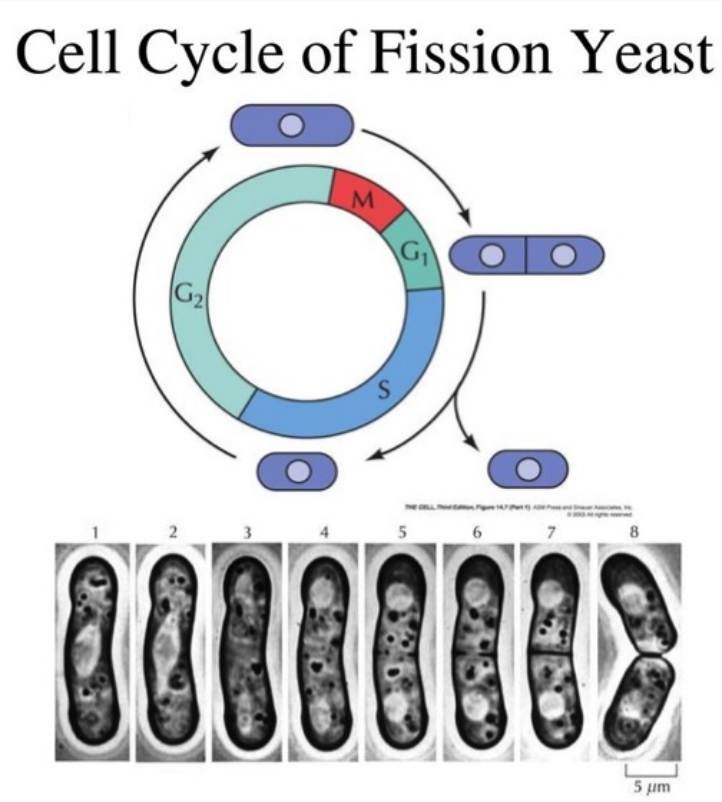
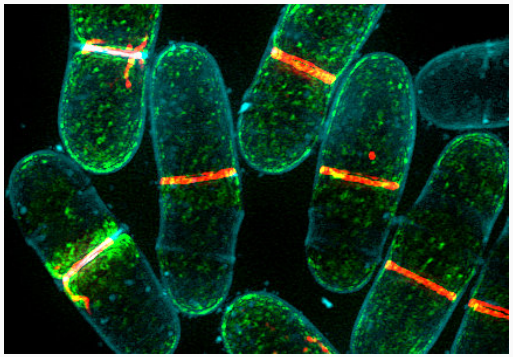
Sparse Envelope Model (Spenv)

Guangyu Zhu

University of Rhode Island

Fission yeast data





- Data: Microarray time-course data [Rustici et al., 2004] on cell-cycle control in the fission yeast
  - Y: expression levels of 396 genes
  - X: 10 equally spaced time points of the cell cycle
  - n: 178 sample hybridizations
- Purpose: study how gene expression levels change in a cell cycle

Envelope Method

- Under the Multivariate linear regression model

$$Y = \beta X + \varepsilon, \quad \varepsilon \sim (0, \mathbf{\Sigma})$$

where $Y \in \mathbb{R}^r$, $X \in \mathbb{R}^p$, and $\beta \in \mathbb{R}^{r \times p}$

- Given an orthogonal matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r \times r}$, where $\mathbf{\Gamma} \in \mathbb{R}^{r \times u}$, then

$$Y = (P_{\mathbf{\Gamma}} + Q_{\mathbf{\Gamma}})Y = P_{\mathbf{\Gamma}}Y + Q_{\mathbf{\Gamma}}Y$$

where $P_{\Gamma} = \mathbf{\Gamma}\mathbf{\Gamma}^T$ and $Q_{\mathbf{\Gamma}} = I_r - P_{\mathbf{\Gamma}} = \mathbf{\Gamma}_0\mathbf{\Gamma}_0^T$ are the projection matrices.

- **Goal of envelope:** $P_{\mathbf{\Gamma}}Y$ contains all the information in Y that can help us estimate β.

Envelope Method

- Model assumptions:

(1) $Q_\Gamma Y | X \sim Q_\Gamma Y \quad \Leftrightarrow \quad \beta = \Gamma\eta$

(2) $\mathrm{Cor}(P_\Gamma Y, Q_\Gamma Y \mid X) = 0 \quad \Leftrightarrow \quad \Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$

- Coordinate form of the envelope model:

$$Y = \Gamma\eta X + \varepsilon, \quad \Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$$

Response feature selection

- Response feature selection
  - Y_i is called **static** if $\beta_i = 0$, otherwise called **dynamic**
  - Y_i is called **inactive** if $\Gamma_i = 0$, otherwise called **active**

- Relationship:
  - Because $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, **inactive** responses are **static**. However, an **active** response may also be **static**
  - **Proposition**: If Y_i is **active** and **static**, then Y_i must be connected with **dynamic** response (in the covariance graph)
  - This indicates that if an **active** response is **static**, it still offers information in estimating the non-zero regression coefficients.

Sparse Envelope Method

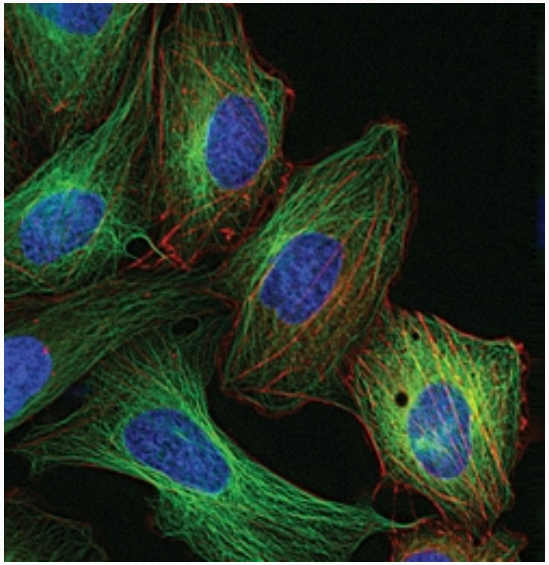
- So we try to identify inactive responses.

- Sparse Envelope Model [Su et al., 2016]

$$Y = \Gamma\eta X + \varepsilon, \quad \Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \quad \Gamma = \begin{pmatrix} \Gamma_{\mathcal{A}} \\ 0 \end{pmatrix}$$

- Asymptotic Properties:
  - **Selection consistency**: probability of selection inactive responses goes to 1
  - **Asymptotic efficiency**: estimator has the same asymptotic covariance matrix as the oracle estimator

Fission yeast data



- Spenv identified 25 inactive responses (genes whose intensities do not change in a cell cycle).

- Some of them are also confirmed by other researchers: gene cdc20 was also identified by [Gilks et al., 2005] to have "very little cell-cycle activity"

References


Gilks, W. R., Tom, B. D., and Brazma, A. (2005).
Fusing microarray experiments with multivariate regression.
Bioinformatics, 21(suppl 2):ii137–ii143.


Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bähler, J. (2004).
Periodic gene expression program of the fission yeast cell cycle.
Nature genetics, 36(8):809–817.


Su, Z., Zhu, G., Chen, X., and Yang, Y. (2016).
Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression.
Biometrika, 103(3):579–593.