

Misaligned Principal Component Analysis: Application to Gene Expression Time Series Analysis.

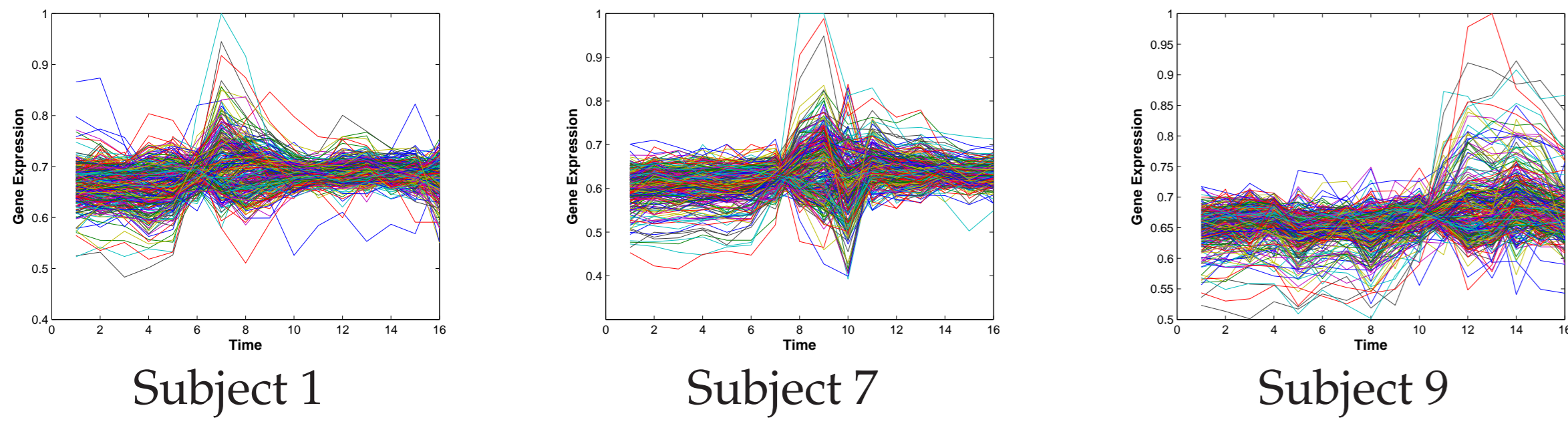
Arнау Tibau-Puig*, Ami Wiesel†, Raj Rao Nadakuditi*, Alfred O. Hero III*

*Dept. of EECS, University of Michigan, †School of Computer Science and Engineering, The Hebrew University of Jerusalem



High-dimensional Misaligned Time Series

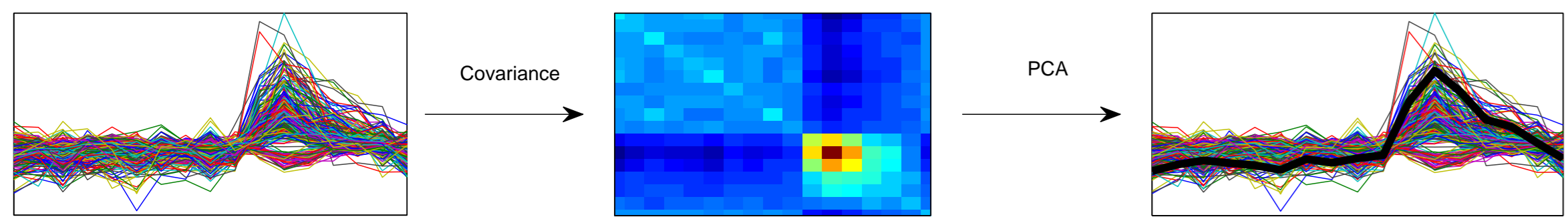
p_n : number of time points
 n : number of samples
 , with $c = \lim_{n \rightarrow \infty} \frac{p_n}{n} > 0$



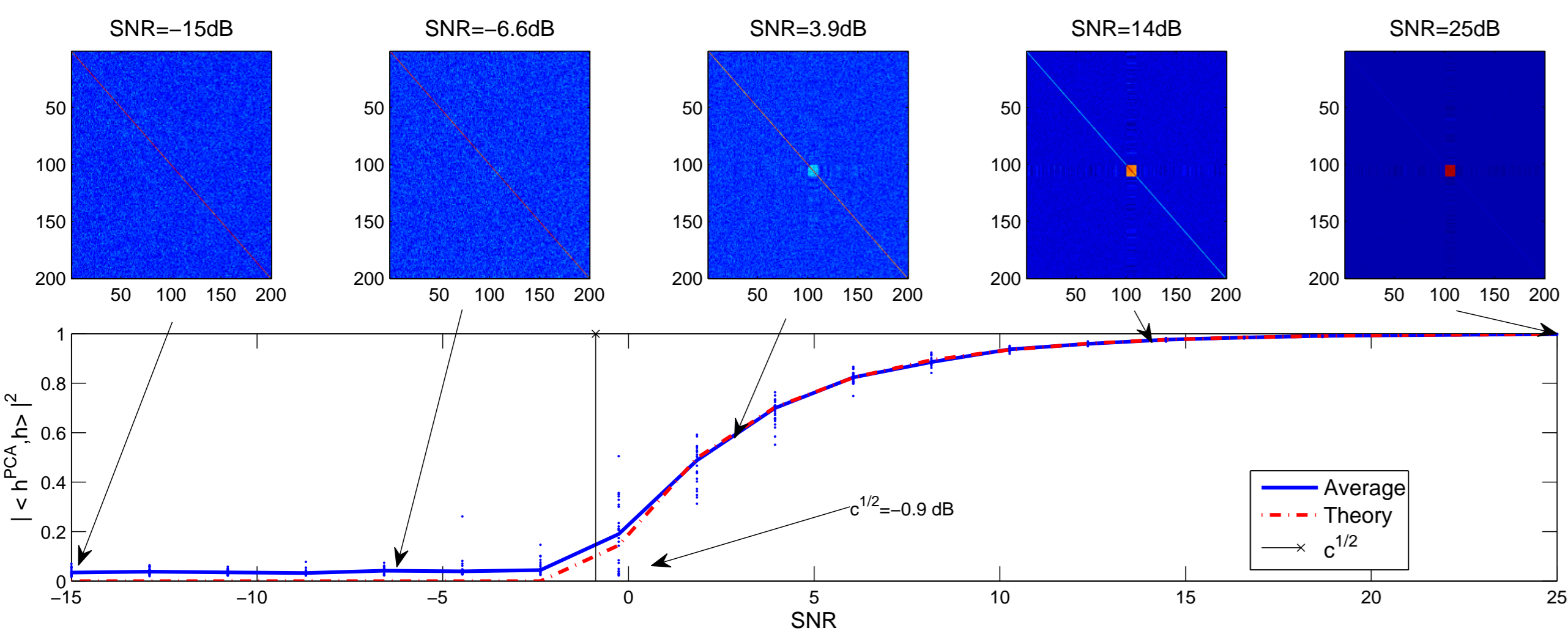
Principal Component Analysis

Given the sample covariance $\mathbf{S} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$,

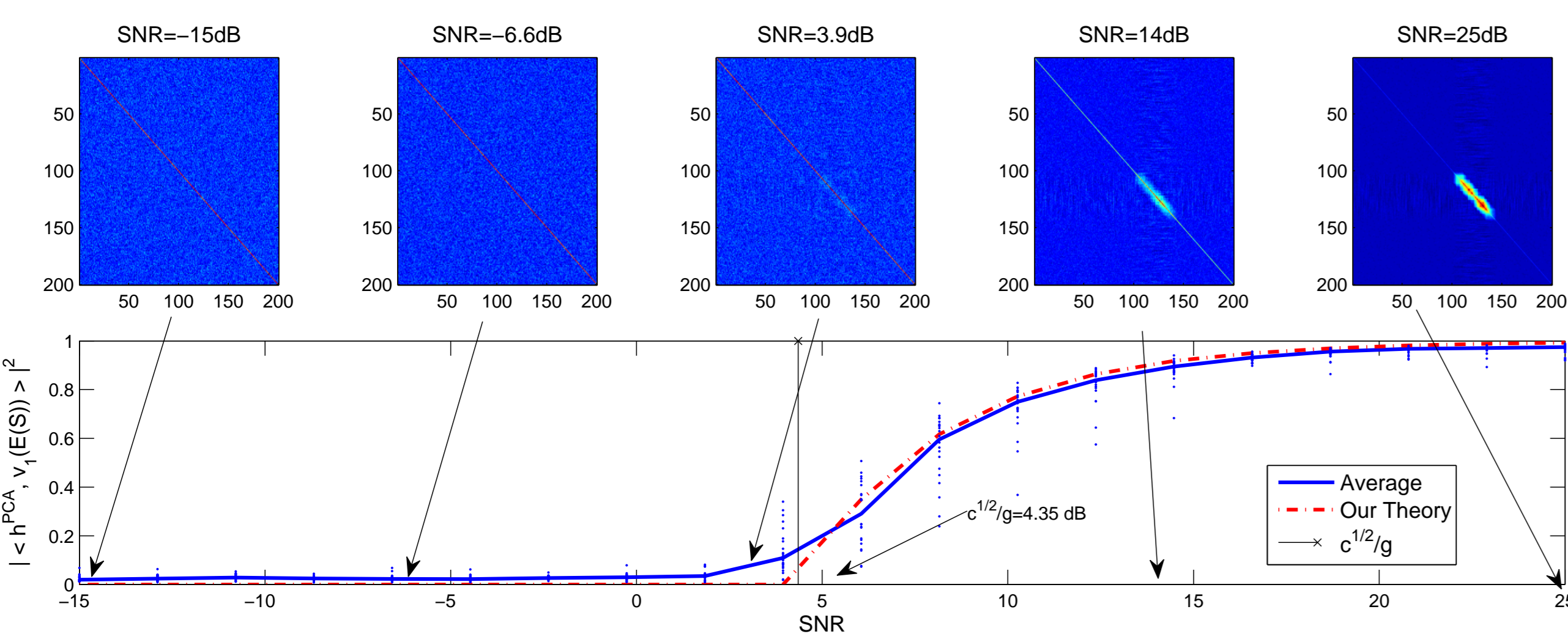
$$\mathbf{h}^{\text{PCA}} := \arg \max_{\|\mathbf{h}\|_2=1} \mathbf{h}^T \mathbf{S} \mathbf{h} \quad (\text{PCA})$$



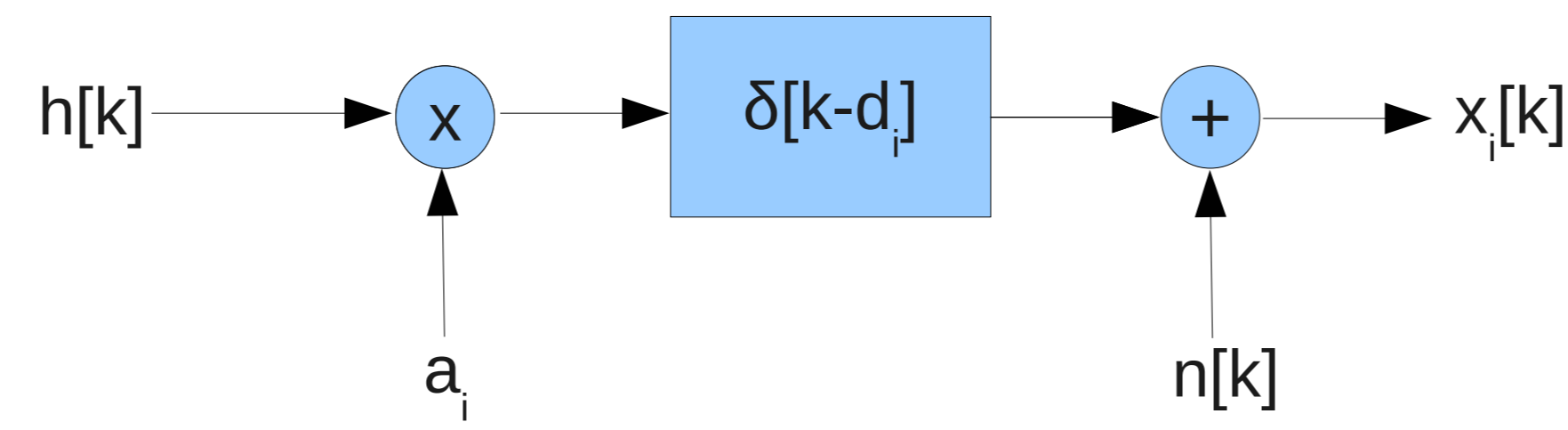
$$|\langle \mathbf{h}^{\text{PCA}}, \mathbf{v}_1(E[\mathbf{S}]) \rangle|^2 \xrightarrow{\text{a.s.}} \begin{cases} \frac{\text{SNR}^2 - c}{\text{SNR}^2 + c \text{SNR}}, & \text{SNR} > \sqrt{c} \\ 0, & \text{SNR} \leq \sqrt{c} \end{cases} \quad [\text{Paul 2007}]$$



PCA with misalignment



Misaligned rank-1 linear factor model



$$\mathbf{x}_i = a_i \mathbf{C}_{d_i} \mathbf{h} + \mathbf{n}_i, \quad \text{with } \begin{cases} d_i, \mathbf{h}: & \text{deterministic} \\ a_i, \mathbf{n}_i: & \text{random.} \end{cases}$$

$$\Sigma_i := E[\mathbf{x}_i \mathbf{x}_i^T] = \text{SNR} \mathbf{C}_{d_i} \mathbf{h} \mathbf{h}^T \mathbf{C}_{d_i}^T + \mathbf{I} \quad (\text{"Spiked" covariance})$$

MLE of \mathbf{h}, \mathbf{d} : The MisPCA problem

$$\mathbf{d}^{\text{MisPCA}} = \arg \max_{\tau \in \{0, \dots, d_{\max}\}^n} \lambda_1(\mathbf{S}(\tau))$$

$$\mathbf{h}^{\text{MisPCA}} = \mathbf{v}_1(\mathbf{S}(\mathbf{d}^{\text{MisPCA}}))$$

$$\mathbf{S}(\tau) = \frac{1}{n} \sum_{i=1}^n \mathbf{C}_{\tau_i}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{C}_{\tau_i} = \text{"Aligned" sample covariance}$$

Approximations:

- $\mathbf{d} = \mathbf{0}$: $\mathbf{h}^{\text{PCA}} = \mathbf{v}_1(\mathbf{S}(\mathbf{0}))$ (PCA)
- Alternating-MisPCA, iteration t :

$$\mathbf{d}_t^{\text{A-MisPCA}} = \arg \max_{\tau \in \{0, \dots, d_{\max}\}^n} \mathbf{h}_{t-1}^{\text{A-MisPCA}T} \mathbf{S}(\tau) \mathbf{h}_{t-1}^{\text{A-MisPCA}}$$

$$\mathbf{h}_t^{\text{A-MisPCA}} = \mathbf{v}_1(\mathbf{S}(\mathbf{d}_t^{\text{A-MisPCA}}))$$

Theorem: Asymptotic eigenstructure of $\mathbf{S}(\tau)$

$$\lambda_1(\mathbf{S}(\tau)) \xrightarrow{\text{a.s.}} \begin{cases} (\text{SNR}\gamma + 1) \left(1 + \frac{c}{\text{SNR}\gamma}\right) & \text{SNR} > \frac{\sqrt{c}}{\gamma} \\ (1 + \sqrt{c})^2 & \text{otherwise.} \end{cases}$$

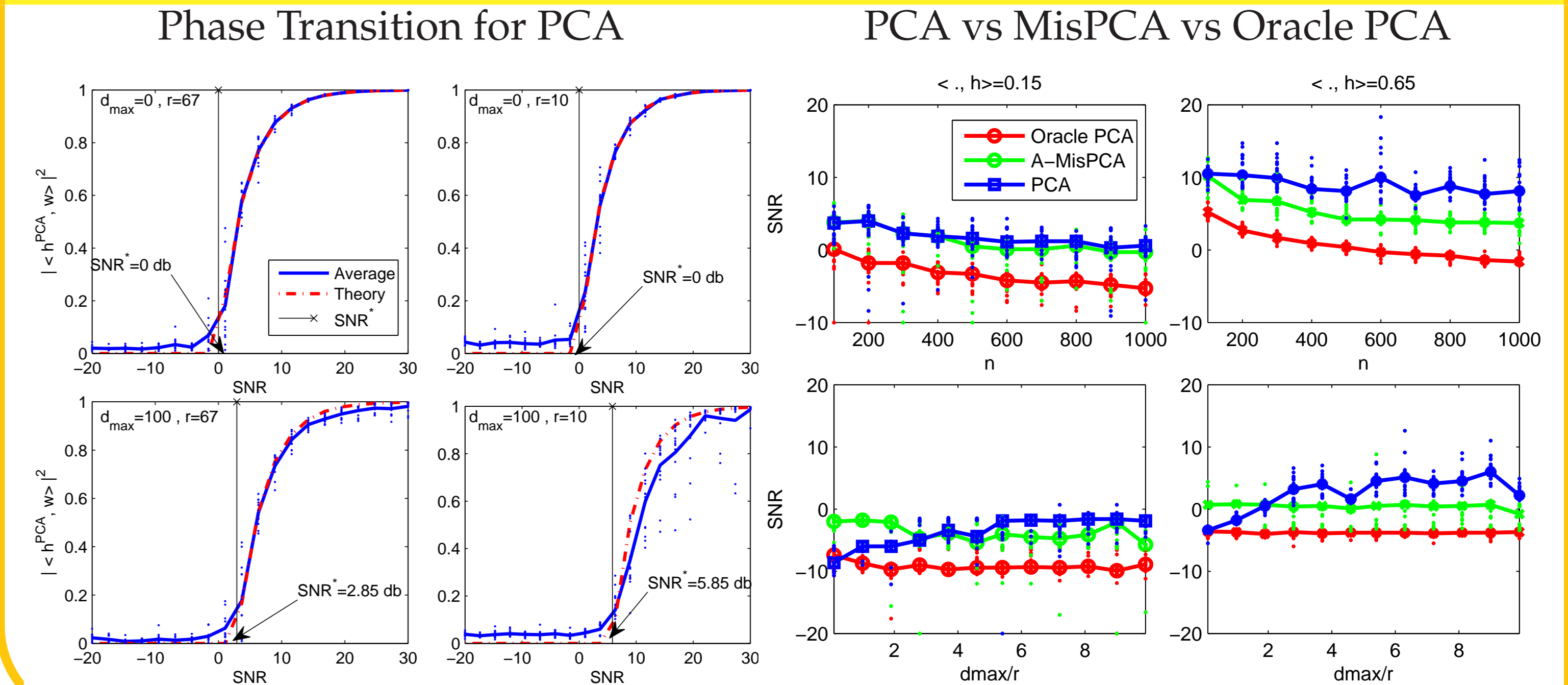
$$|\langle \mathbf{v}_1(\mathbf{S}(\tau)), \mathbf{v}_1(E[\mathbf{S}(\tau)]) \rangle|^2 \xrightarrow{\text{a.s.}} \begin{cases} \frac{(\text{SNR}\gamma)^2 - c}{(\text{SNR}\gamma)^2 + c \text{SNR}\gamma} & \text{SNR} > \frac{\sqrt{c}}{\gamma} \\ 0 & \text{otherwise.} \end{cases}$$

Loss/gain due to misalignment:

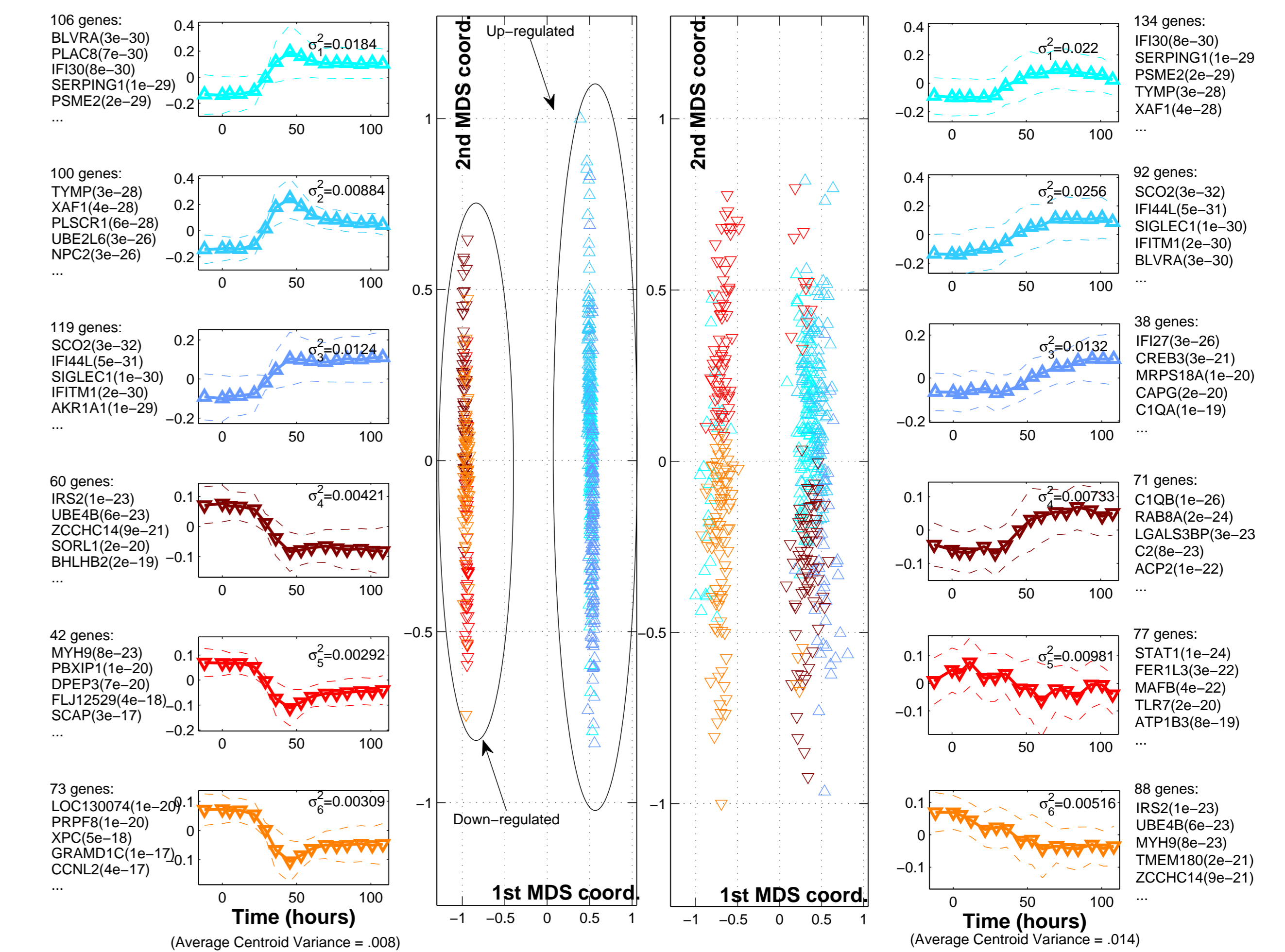
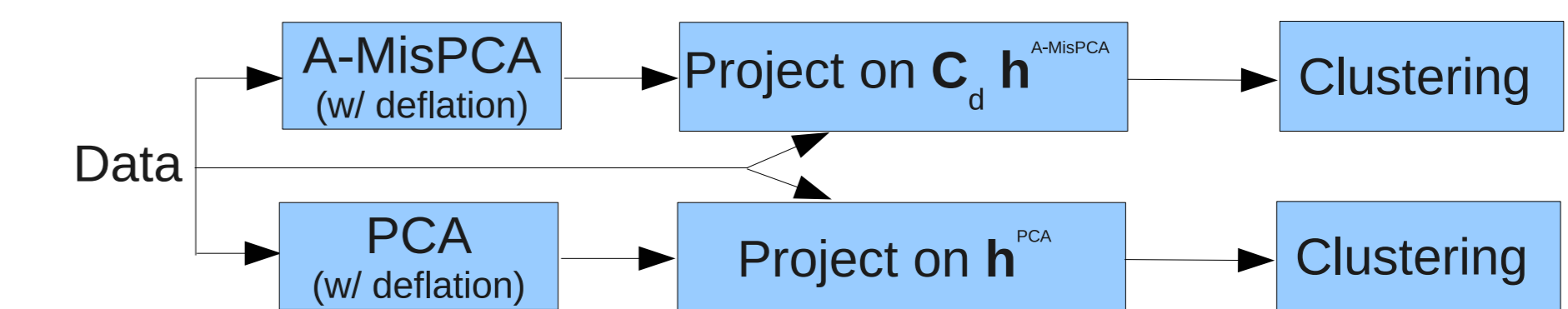
$$\gamma = \lambda_1 \left(\mathbf{D}(\mathbf{d} - \tau)^{\frac{1}{2}} \mathbf{R}_h \mathbf{D}(\mathbf{d} - \tau)^{\frac{1}{2}} \right),$$

- \mathbf{R}_h : Autocorrelation matrix of \mathbf{h}
- $\mathbf{D}(\mathbf{x})$: diagonal, $[\mathbf{D}(\mathbf{x})]_{i,i} = \frac{|j \in \{1, \dots, n\} : x_j = i|}{n}$.

Numerical Experiments



Gene Expression Data Analysis



References

- [Paul 2007] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," Statistica Sinica, Oct. 2007.
- [Huang 2011] Y. Huang et al., "Temporal Dynamics of Host Molecular Responses Differentiate Symptomatic and Asymptomatic Influenza A Infection," PLoS Genetics, Sept. 2011
- [Tibau-Puig 2011] A. Tibau-Puig et al. "Order-preserving factor analysis - application to longitudinal gene expression," IEEE TSP, Sept. 2011