Phase transition in PCA with missing data: Reduced signal-to-noise ratio, not sample size!

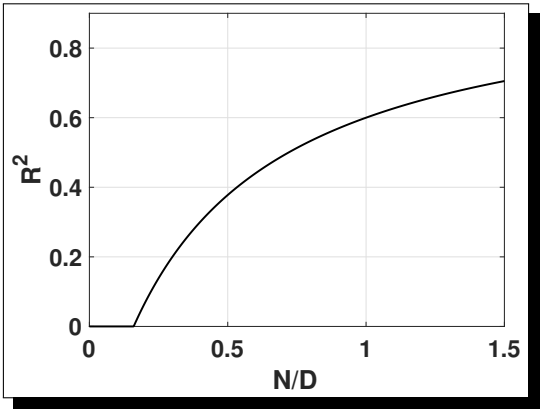
Niels Bruun Ipsen, Lars Kai Hansen

Technical University of Denmark

nbip@dtu.dk

June 7, 2019

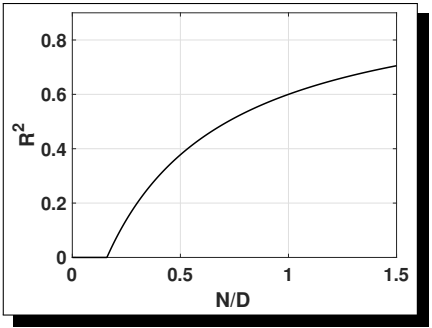
$$R^2 = (\hat{\mathbf{a}}^\top \cdot \mathbf{a}_0)^2$$



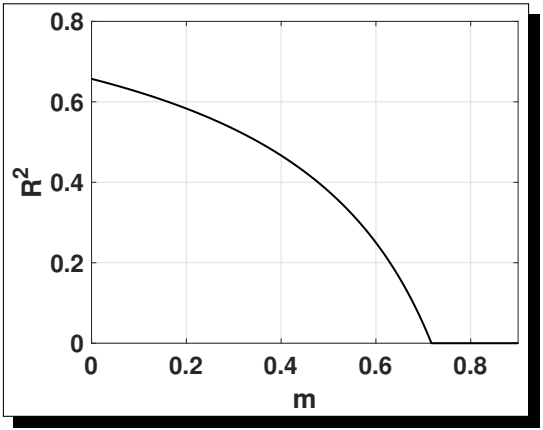
R^2, alignment
N, number of samples
D, number of dimensions

[Biehl, M., & Mietzner, A. (1993)], [Hoyle, D. C., & Rattray, M. (2007)]

$$\langle R^2 \rangle_{\mathcal{X}} = \begin{cases} 0 & \frac{N}{D} S^2 < 1, \\ \frac{\frac{N}{D} S^2 - 1}{S + \frac{N}{D} S^2} & \frac{N}{D} S^2 \geq 1. \end{cases}$$



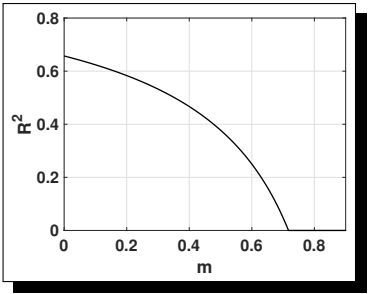
R^2, alignment
N, number of samples
D, number of dimensions
S, signal-to-noise ratio



R^2, alignment
m, missing rate

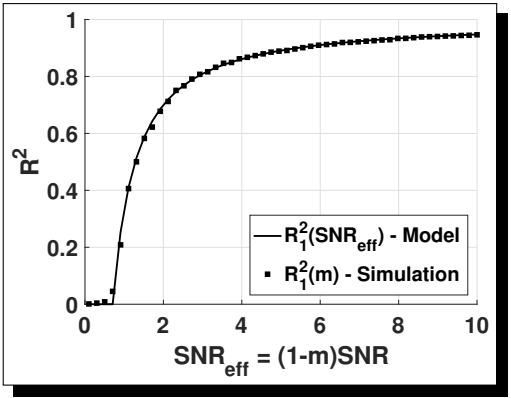
Learning curves, including missing

$$\langle R^2 \rangle_{\mathcal{X}} = \begin{cases} 0 & \frac{N}{D}((1-m)S)^2 < 1, \\ \frac{\frac{N}{D}((1-m)S)^2 - 1}{(1-m)S + \frac{N}{D}((1-m)S)^2} & \frac{N}{D}((1-m)S)^2 \geq 1. \end{cases}$$

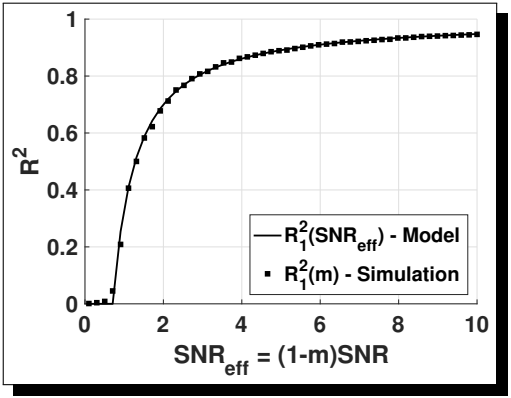


R^2, alignment
m, missing rate
S, signal-to-noise ratio
N, number of samples
D, number of dimensions

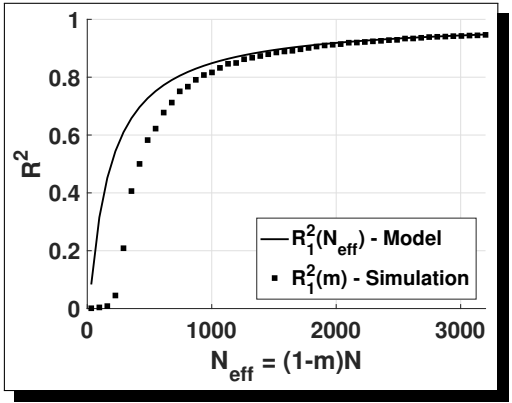
Effective SNR

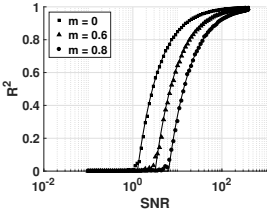


Effective SNR

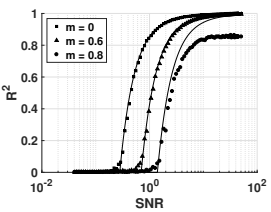


Effective sample size?

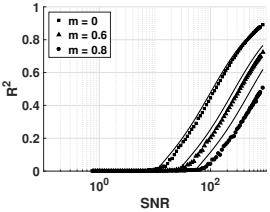




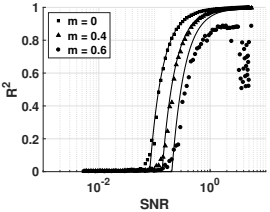
(a) Wild Faces



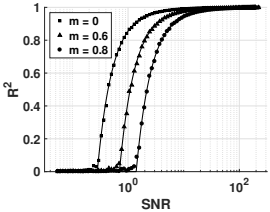
(b) MNIST



(c) NCI60



(d) Food Pairing



(e) Fashion

Thank You

Poster 199 Pacific Ballroom, 6:30 - 09:00 pm

Hoyle, D. C., & Rattray, M. (2007)
Statistical mechanics of learning multiple orthogonal signals: Asymptotic theory and fluctuation effects.
Physical Review E 75.1 (2007): 016101.

Biehl, M., & Mietzner, A. (1993)
Statistical mechanics of unsupervised learning.
EPL (Europhysics Letters) 24.5 (1993): 421.