Unsupervised learning of features from data: a statistical physics approach

R. Monasson
Laboratory of Physics
CNRS & Ecole Normale Supérieure, Paris

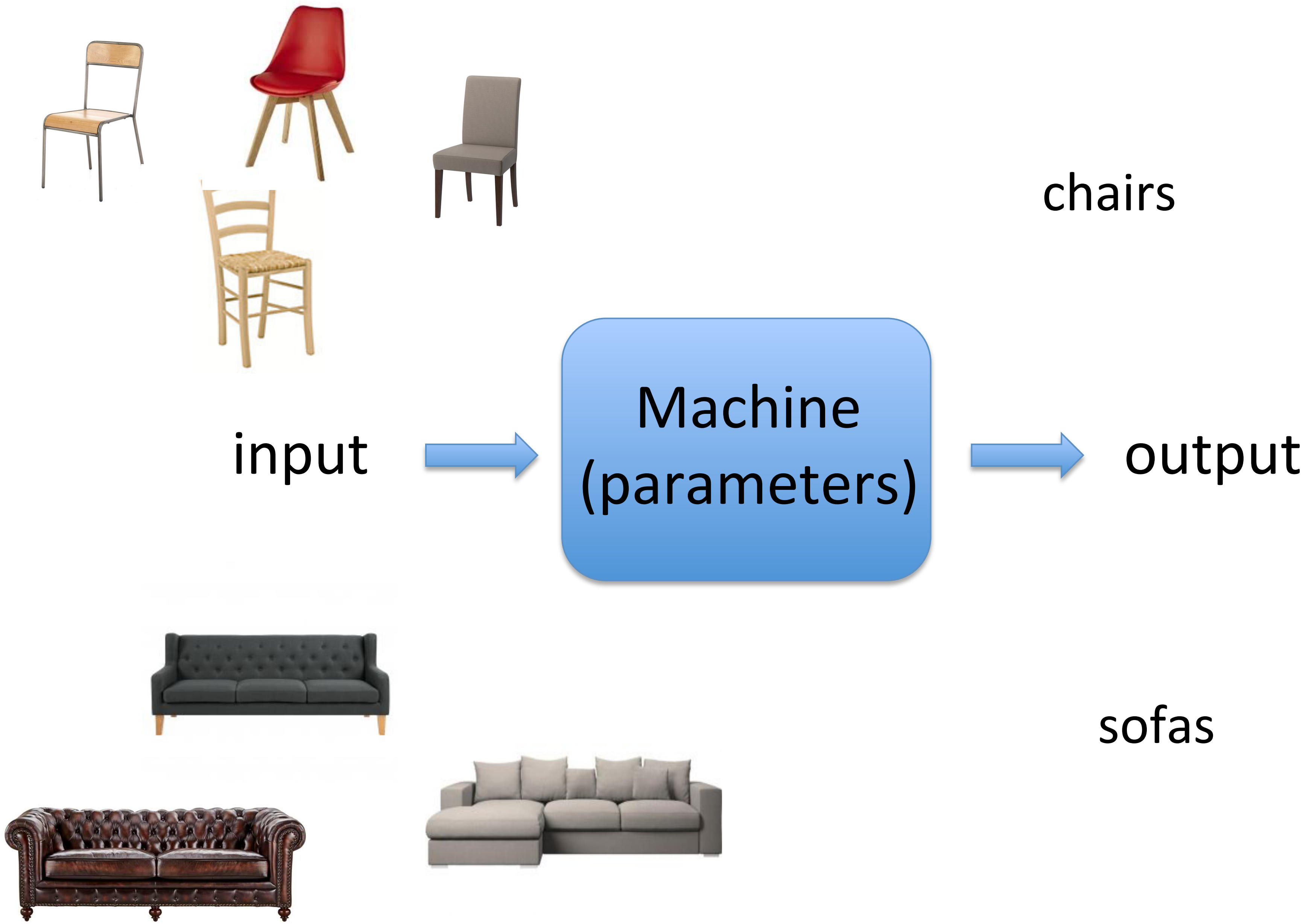
Model-guided data science, Como, September 2019

Learning from data

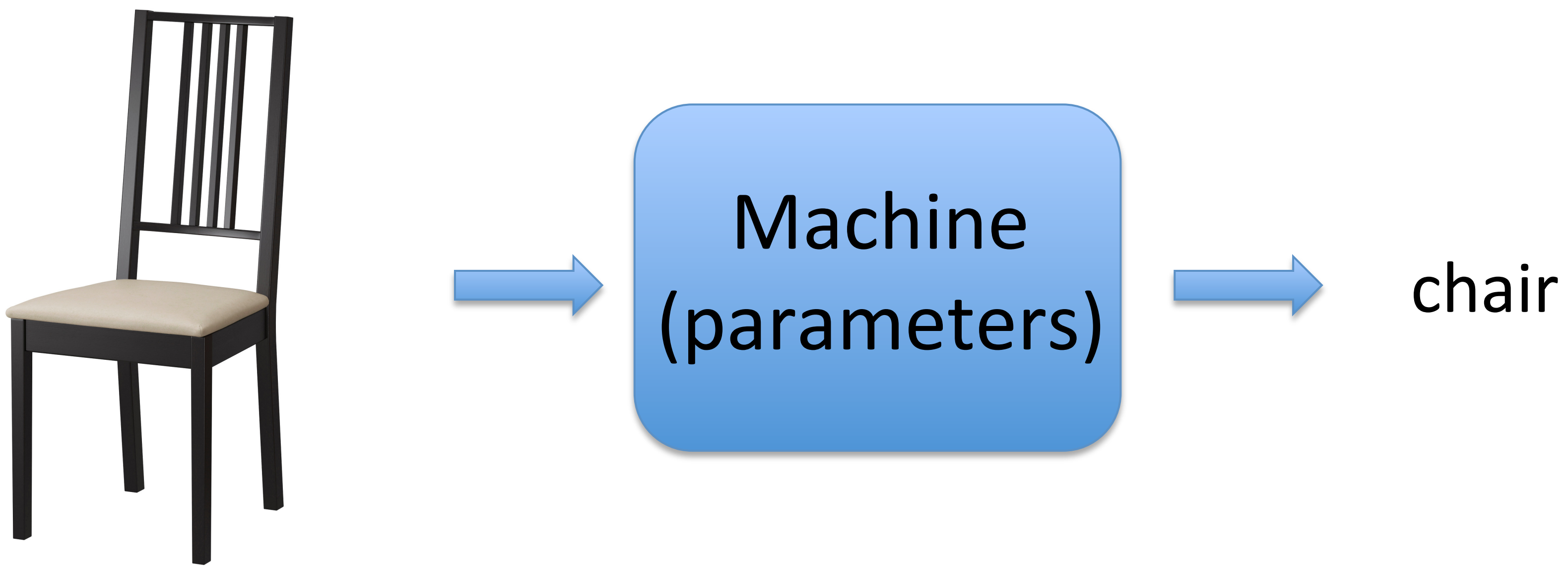
chairs

sofas

Supervised learning from data



Supervised learning from data



Supervised learning from data

Machine
(parameters)

chair!

Supervised learning: fit of input-output relation from examples
(in high dimensions)

Unsupervised learning from data

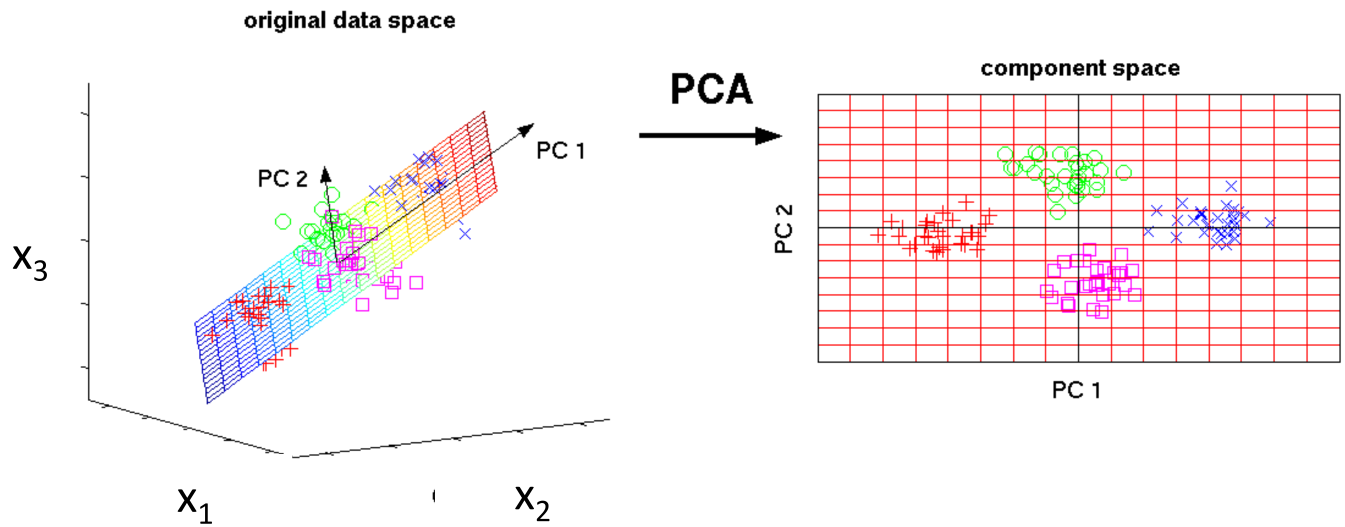
Machine
(parameters)

→ Representations

What is unsupervised learning about?

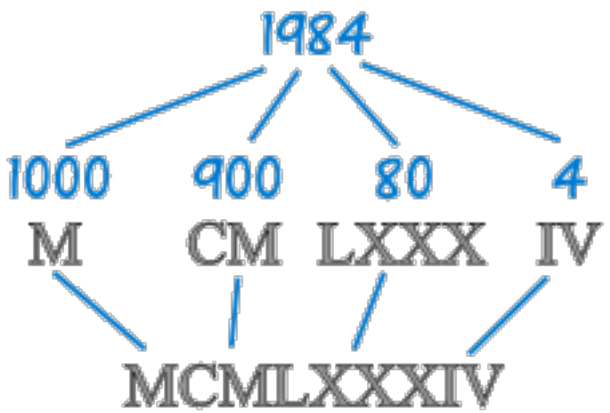
- Find statistical features of data: clustering, dimensional reduction, …

- Find adequate **representations** to interpret,



or process

data

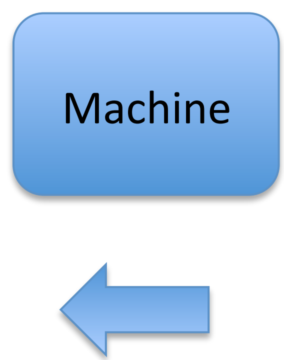


What is unsupervised learning about?

- Find statistical features of data:
  clustering, dimensional reduction, …

- Find adequate **representations** (to interpret or process data)

- Generate new data



Representation
generating model

What is unsupervised learning about?

- Find statistical features of data:
  clustering, dimensional reduction, … Today!

- Find adequate **representations** (to interpret or process data)

- Generate new data

Representation generating model

Plan of the lectures

1. **Bayesian Inference and dimensional reduction: phase transition in principal component analysis**

2. Representations: auto-encoders, Restricted Boltzmann Machines & sparse feature learning

3. Restricted Boltzmann Machines: connections with graphical models, phase transitions & applications

Bayesian inference

(in a few slides ..)

Data
(observations σ)

Model
(parameters τ)

new data
(predictions)

Bayesian inference

Data
(observations σ)

Model
(parameters τ)

new data
(predictions)

Probabilistic description: joint distribution of σ & τ

$$p(\sigma,\tau) = p(\sigma|\tau) \times p(\tau)$$

Prior distribution over model parameters

Likelihood of model parameters

$$= p(\tau|\sigma) \times p(\sigma)$$

Proba. data are generated by models τ

Posterior distribution over model parameters
(can be sampled, maximized, ...)

Bayes inference formula:

$$p(\tau|\sigma) = \frac{p(\sigma|\tau) \times p(\tau)}{p(\sigma)}$$

A historical example: Laplace birth rate problem

Historical example: Laplace's « proof » that boys and girls have ≠ birth rates

Data: Nbs. of girls born in Paris from 1745 to 1770 : 245,945

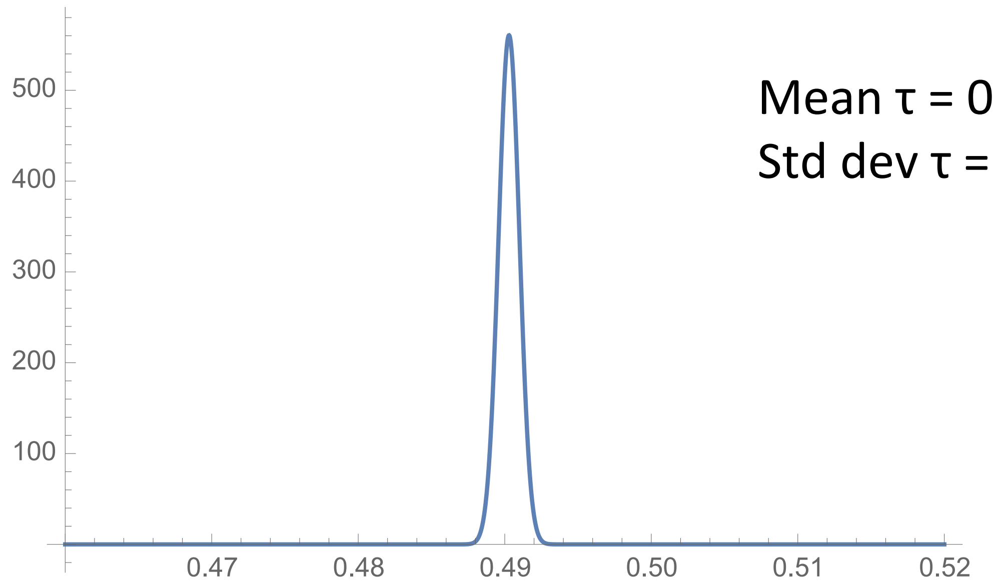
... boys ... : 251,527

Model: σ = nb. of female births, n = nb. nirths, τ = girl birth probability

- likelihood: $$p(\sigma|\tau) = \binom{n}{\sigma} \tau^{\sigma}(1-\tau)^{n-\sigma}$$
- prior: uniform over τ in [0;1]
- Bayes: $$p(\tau|\sigma) = \frac{\tau^{\sigma}(1-\tau)^{n-\sigma}}{\int_0^1 d\tau' \, \tau'^{\sigma}(1-\tau')^{n-\sigma}}$$

A historical example: Laplace birth rate problem

Posterior distribution:

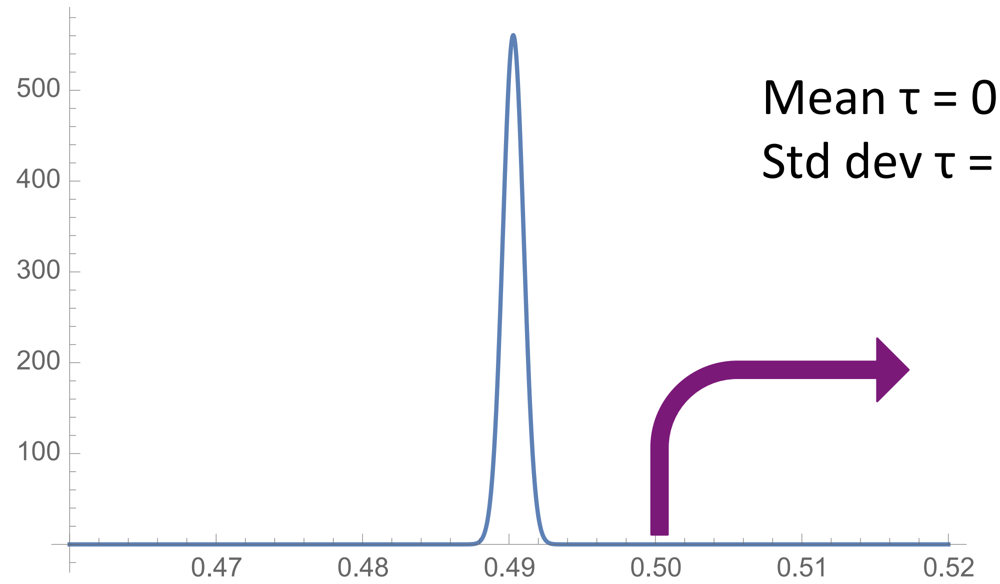


Mean τ = 0.490291

Std dev τ = 0.007117

A historical example: Laplace birth rate problem

Posterior distribution:

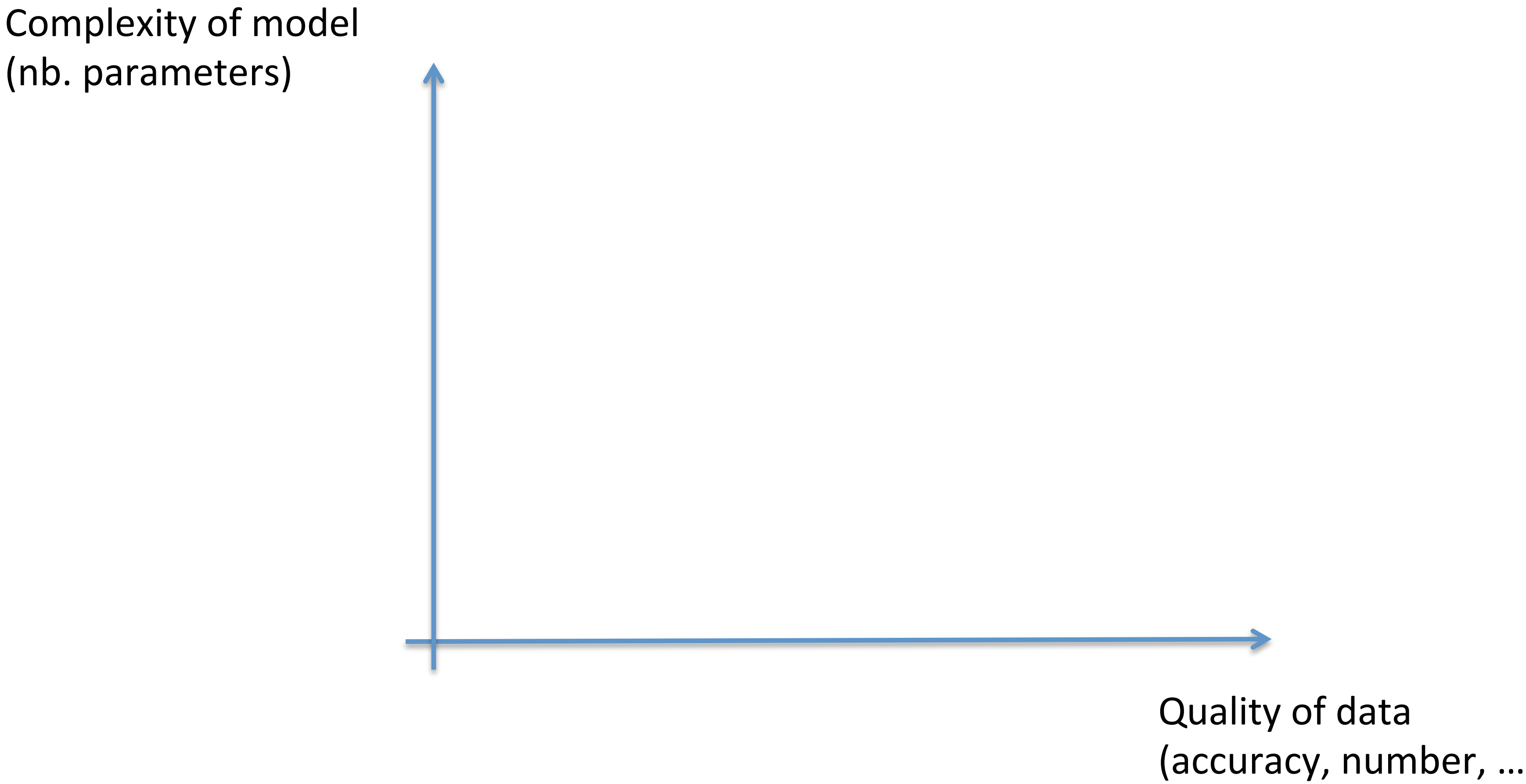


Mean τ = 0.490291
Std dev τ = 0.007117

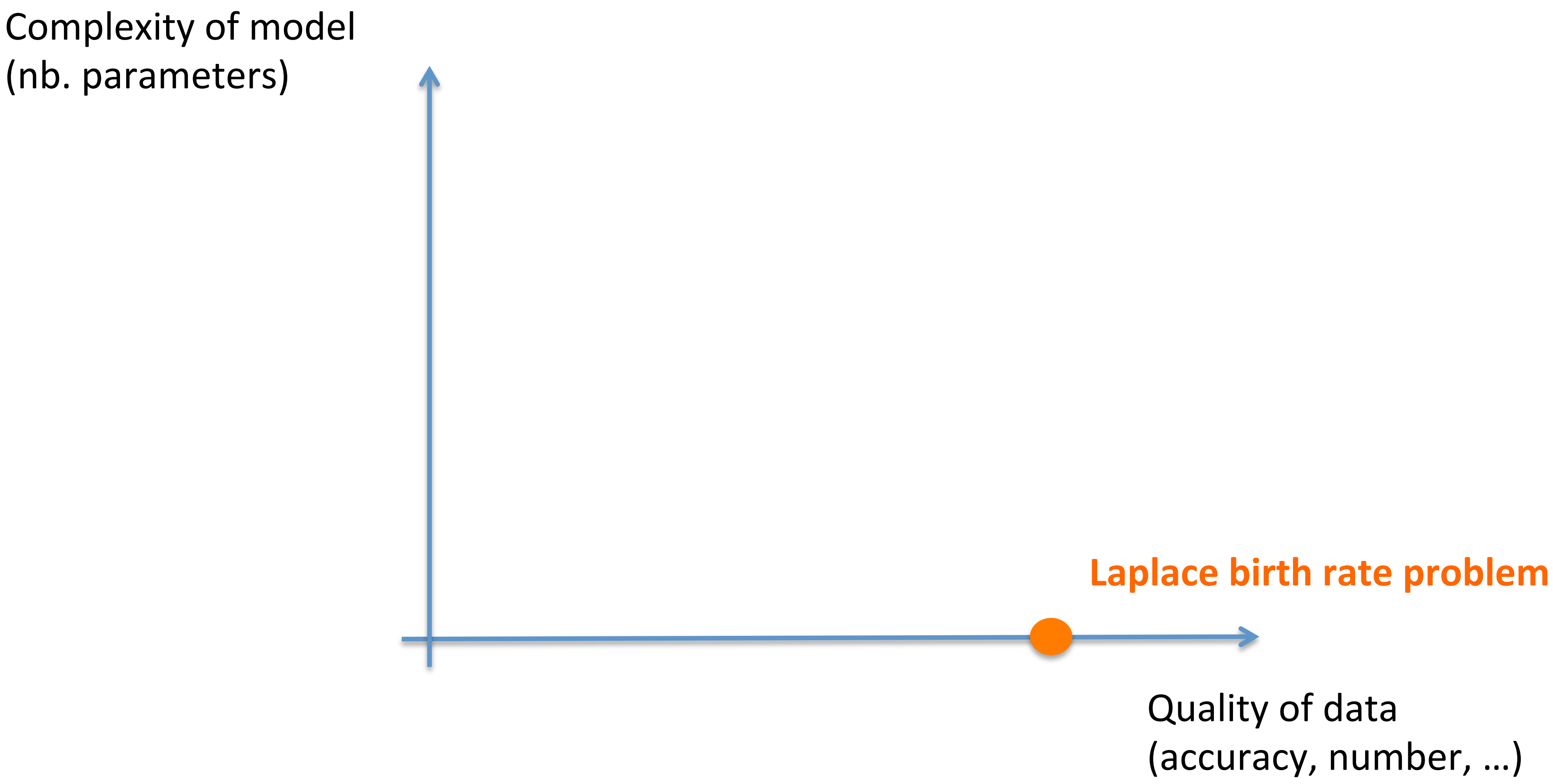
Probability that τ exceeds 0.5 = $\int_{0.5}^{1} d\tau \; p(\tau|\sigma) \approx 10^{-42}$

Very rare event!

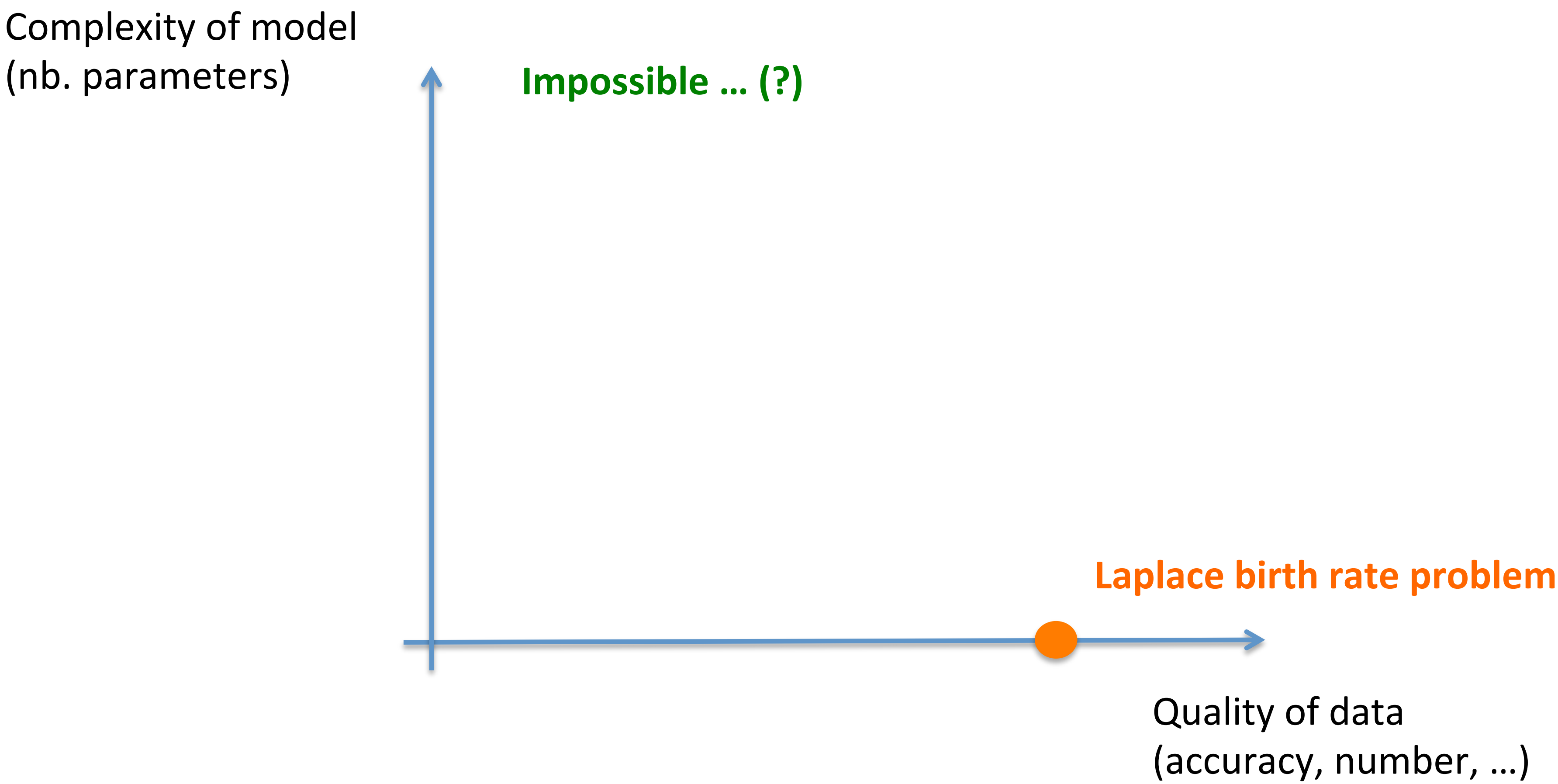
Bayesian inference: High-dimensional setting



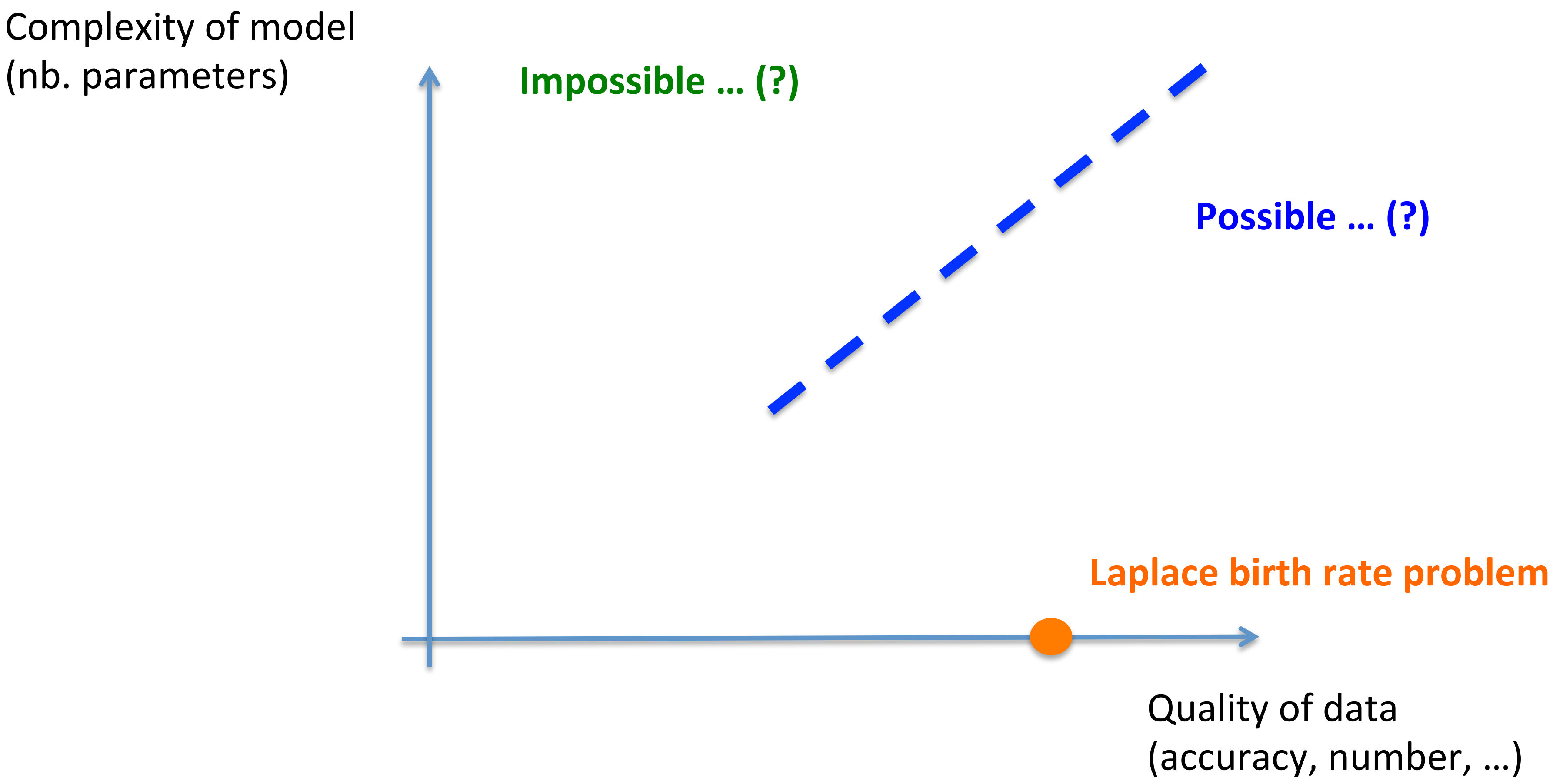
Bayesian inference: High-dimensional setting



Bayesian inference: High-dimensional setting



Bayesian inference: High-dimensional setting



Bayesian inference: High-dimensional setting

Configuration: vectors of p variables: $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_p)$

Model: Gaussian distribution $\rho(\sigma | \boldsymbol{\tau}) = \frac{\sqrt{\det \boldsymbol{\tau}}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \sigma^T \cdot \boldsymbol{\tau} \cdot \sigma \right)$

precision matrix

Moments:

$$\langle \sigma_i \rangle = 0$$

$$\langle \sigma_i \sigma_j \rangle \equiv C_{ij} = \left(\boldsymbol{\tau}^{-1} \right)_{ij}$$

correlation matrix

A trivial case: independent variables (null model)

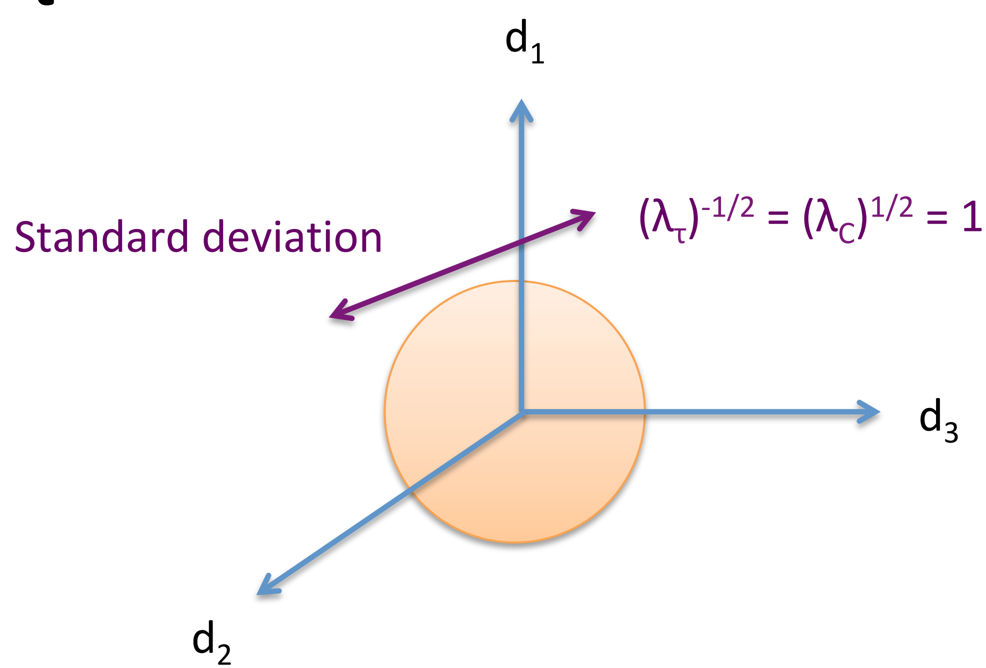
No interaction:

$$\tau = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}$$

[p x p Identity matrix]

Correlation matrix:
(infinite sampling)

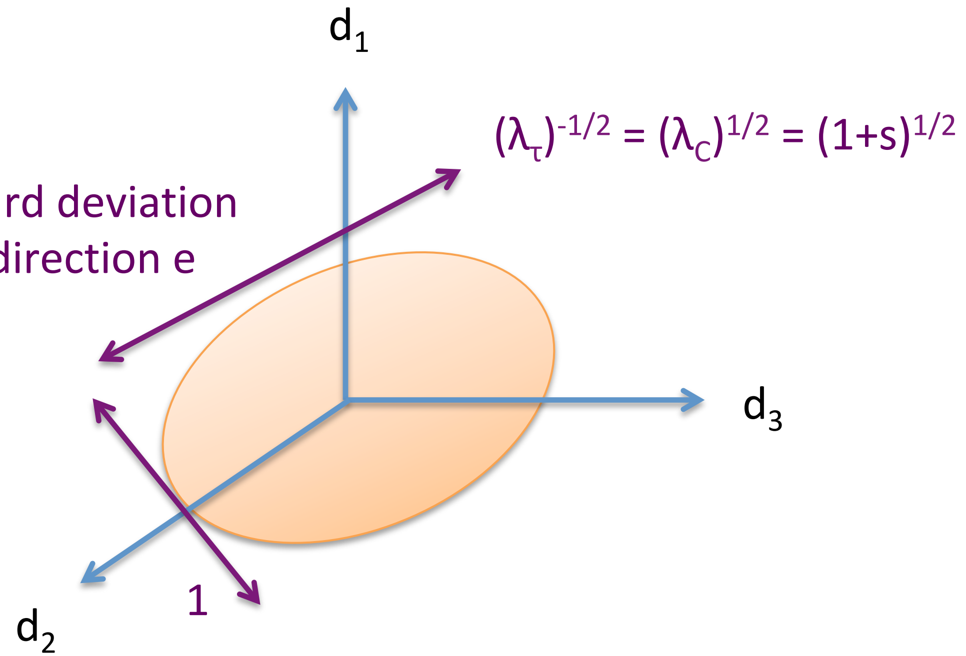
$$C = \tau^{-1}$$



Minimal non trivial case

One special direction: $$\tau = Id - \frac{s}{1+s} |e\rangle\langle e| \qquad (s>0)$$

Correlation matrix: $$C = \tau^{-1} = Id + s\, |e\rangle\langle e|$$
(infinite sampling)



Principal Component Analysis

Bayesian inference: High-dimensional setting

Data: n samples of p multivariate Gaussian variables, (assumed to be independent)

$$\sigma^{(s)} = (\sigma_1^{(s)}, \sigma_2^{(s)}, \ldots, \sigma_p^{(s)})$$

Likelihood:

$$\prod_s \rho(\sigma^{(s)} | \tau) = \left(\frac{\sqrt{\det \tau}}{(2\pi)^{p/2}} \right)^n \exp\left(-\frac{1}{2} \sum_s \sum_{i,j} \sigma_i^{(s)} \tau_{ij} \sigma_j^{(s)} \right)$$

Log-likelihood:

$$L(\tau) = \frac{n}{2} \log \det \tau - \frac{n}{2} \sum_{i,j} \hat{C}_{ij} \tau_{ij} + cst$$

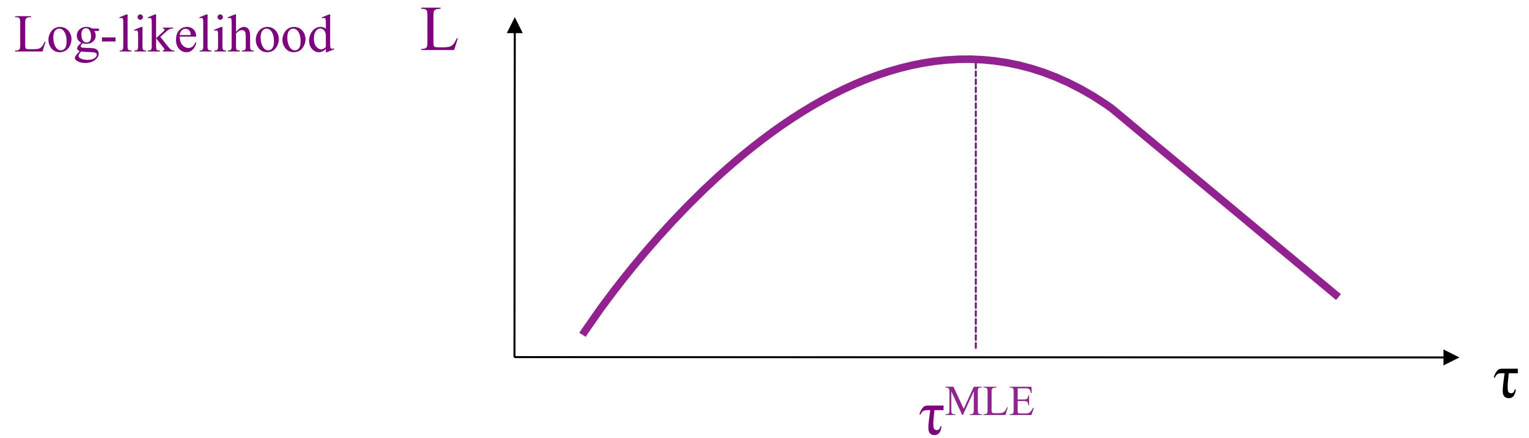
correlations from data

correlations from model

Maximization:

$$\frac{\partial}{\partial \tau_{ij}} L(\tau) \Big|_{\tau_{MLE}} = 0 = \frac{n}{2} (\tau_{MLE}^{-1})_{ji} - \frac{n}{2} \hat{C}_{ij}$$

Bayesian inference: High-dimensional setting



- Hessian of L is negative semi-definite, hence L is concave (easy to show)
- L_2-regularization removes zero modes if necessary: $L(\tau) \to L(\tau) - \frac{\gamma}{2}\sum_{i,j} \tau_{ij}^2$
- But empirical correlation matrix corrupted by sampling noise: inversion is unreliable

Empirical correlations = $\left(\hat{C}_{ij} \pm n^{-1/2}\right)$ $\updownarrow$ p $\Rightarrow$ Errors on inverse matrix of the order of $(p/n)^{1/2}$ …

nb. of data

Minimal non trivial case

How to infer |e> from data?

Log-likelihood: $$L(\tau) = \frac{n}{2}\log\det\tau - \frac{n}{2}\sum_{i,j}\hat{C}_{ij}\tau_{ij} + cst$$

correlations from data

Expression of precision matrix: $$\tau = Id - \frac{s}{1+s}|e\rangle\langle e|$$

Log-likelihood: $$\frac{n\ s}{2(1+s)}\underbrace{\sum_{ij} e_i \hat{C}_{ij} e_j}_{} + \dots$$

Maximum Likelihood Estimator: find top component of empirical C

Example of PCA application

MNIST data set: 60,000 handwritten digits (not Gaussian!!)

Example of PCA application

Top components

of correlation matrix:

Negative entries

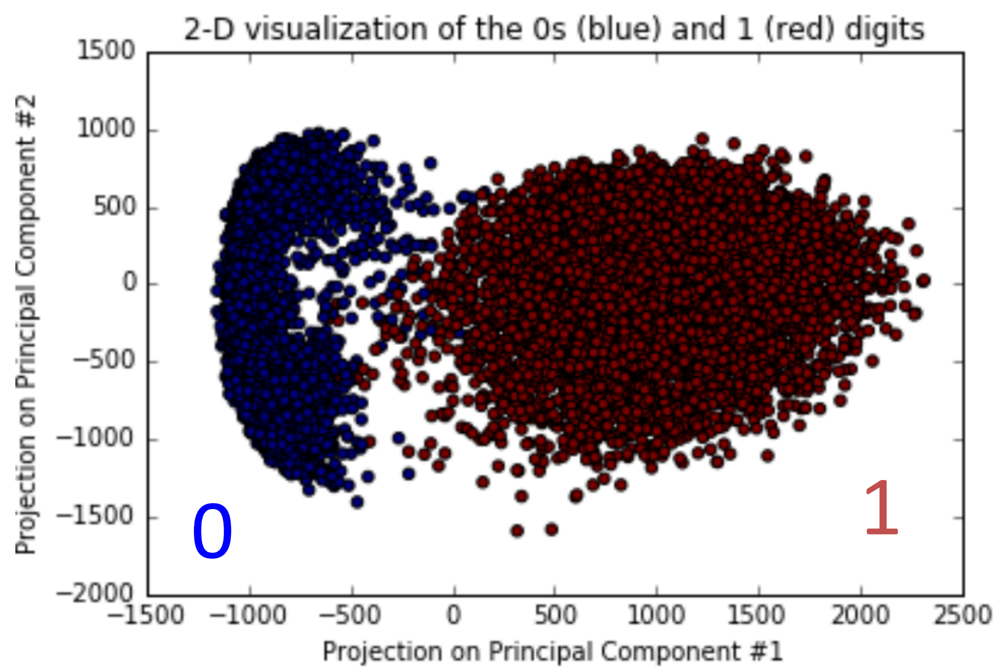
Positive entries

Example of PCA application

Top components
of correlation matrix:

Negative entries
Positive entries

Visualization of
0 and 1 digits:



Many applications, in particular in biology, chemistry, engineering, …

Minimal non trivial case

How to infer |e> from data?

Log-likelihood: $$L(\tau) = \frac{n}{2}\log \det \tau - \frac{n}{2}\sum_{i,j} \hat{C}_{ij}\tau_{ij} + cst$$

correlations from data

Expression of precision matrix: $$\tau = Id - \frac{s}{1+s}|e\rangle\langle e|$$

Log-likelihood: $$\frac{n\ \ s}{2(1+s)}\underbrace{\sum_{ij} e_i \hat{C}_{ij} e_j}_{} + \dots$$

Maximum Likelihood Estimator: find top component of empirical $\hat{C}$

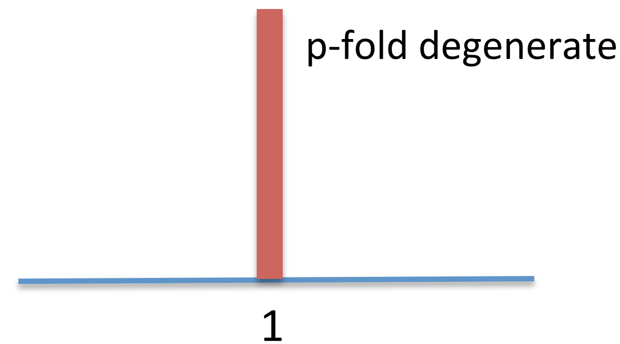
How close are the top components of empirical and « true » correlation matrices??

A trivial case: independent variables (null model)

No interaction:

$$\tau = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & 0 & & \ldots & \\ & & & & 1 \end{pmatrix}$$

[p x p Identity matrix]



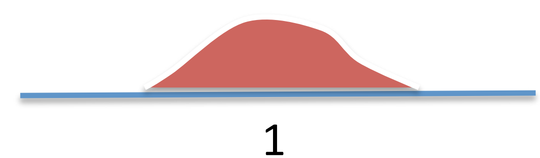
Correlation matrix:
(infinite sampling)

$$C = \tau^{-1}$$

Empirical correlation matrix:
(n samples)

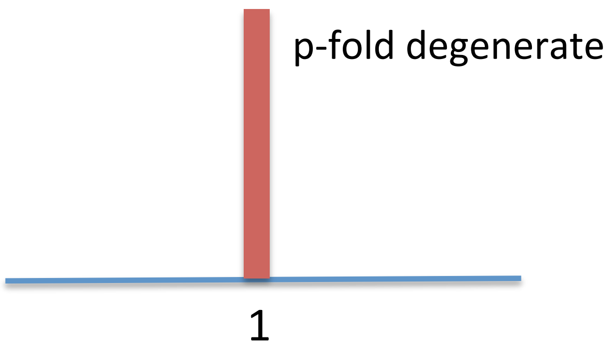
$$\hat{C} = \frac{1}{n} \sum_{s} \sigma^{(s)} \cdot (\sigma^{(s)})^{T}$$

spectrum ???



A trivial case: independent variables (null model)

No interaction: $\tau = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & 0 & & \cdots & \\ & & & & 1 \end{pmatrix}$ [p x p Identity matrix]



Correlation matrix: $C = \tau^{-1}$

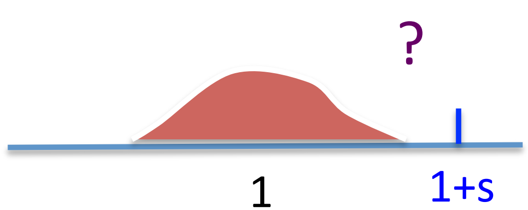
(infinite sampling)

Empirical correlation matrix:

(n samples)

$$\hat{C} = \frac{1}{n}\sum_{s} \sigma^{(s)} \cdot (\sigma^{(s)})^{T}$$

spectrum ???



A trivial case: independent variables (null model)

Random matrix problem: can be solved in many different ways …
(asked me for handwritten notes 1)

A trivial case: independent variables (null model)

Random matrix problem: can be solved in many different ways …
(asked me for handwritten notes 1)

Results: *n* = nb. samples, *p* = nb. Variables

Double limit $n, p \to \infty$ at fixed noise level $\boxed{r = \frac{p}{n}}$

Density of eigenvalues is self-averaging, and equal to

$$\boxed{\rho(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\ r\ \lambda}}$$

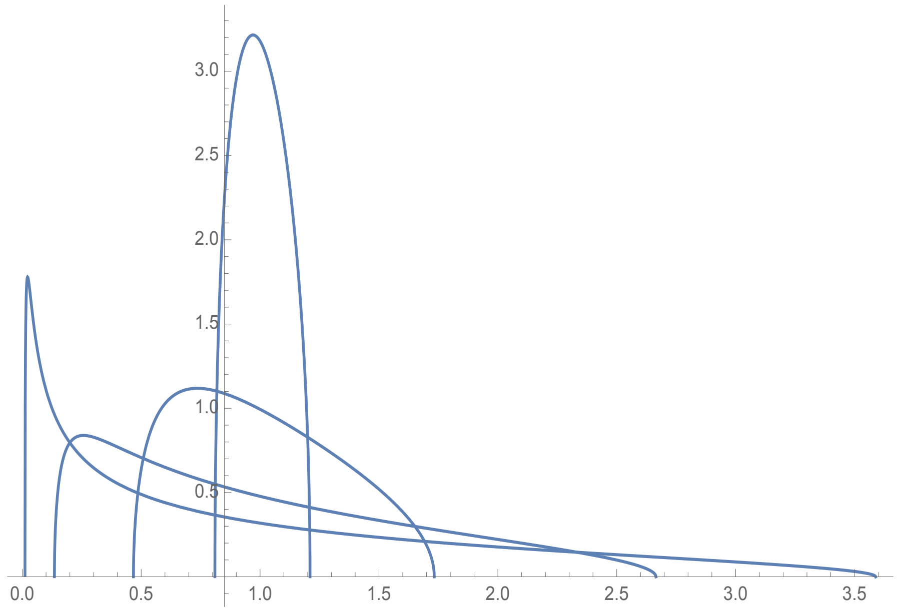
with

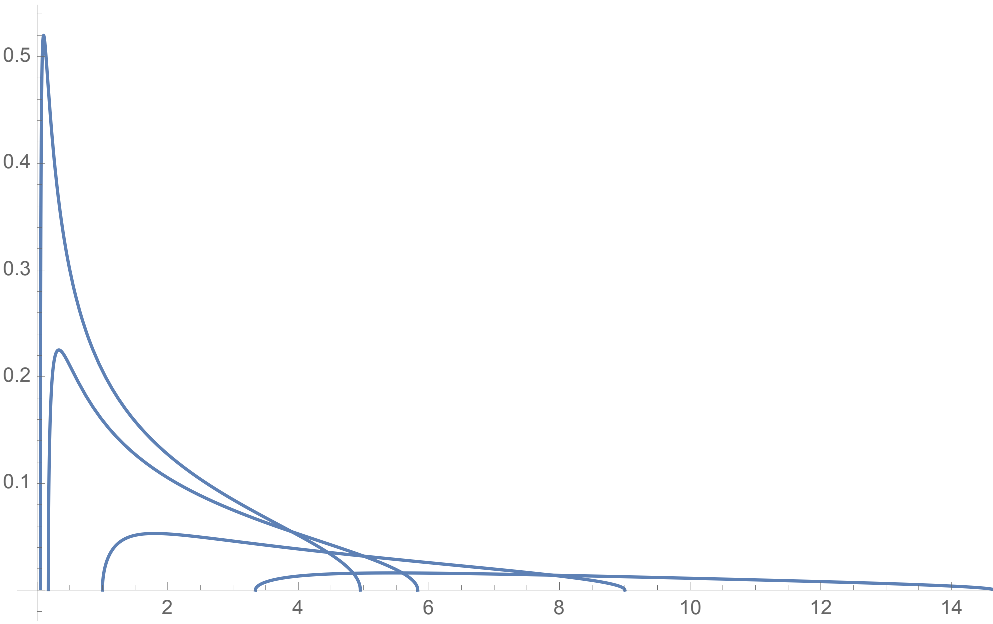
$$\boxed{\lambda_\pm = \left(1 \pm \sqrt{r}\right)^2}$$

- correct for r < 1, otherwise Dirac peak in 0 of height 1-1/r
- graphical representation

A trivial case: independent variables (null model)

```
In[1]:= lamedge[r_, s_] := (1 + s Sqrt[r]) ^ 2
        rho[lam_, r_] := Sqrt[(lamedge[r, +1] - lam) (lam - lamedge[r, -1])] / (2 Pi r lam)
        graph[r_] := Plot[rho[lam, r], {lam, lamedge[r, -1], lamedge[r, +1]}, PlotRange → All]
        Show[graph[.01], graph[.1], graph[.4], graph[.8]]
        Show[graph[1.5], graph[2.], graph[4.], graph[8.]]
```

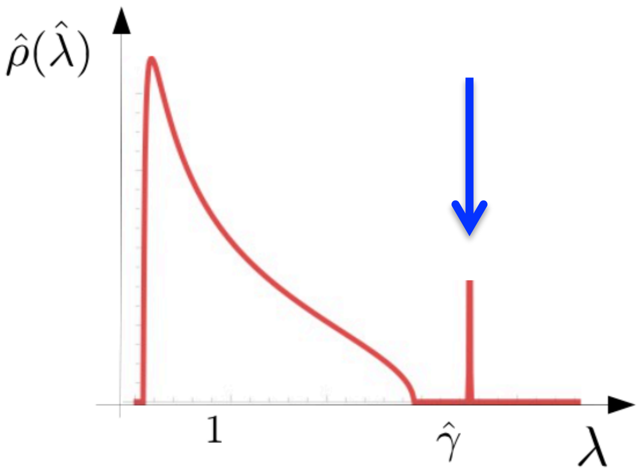




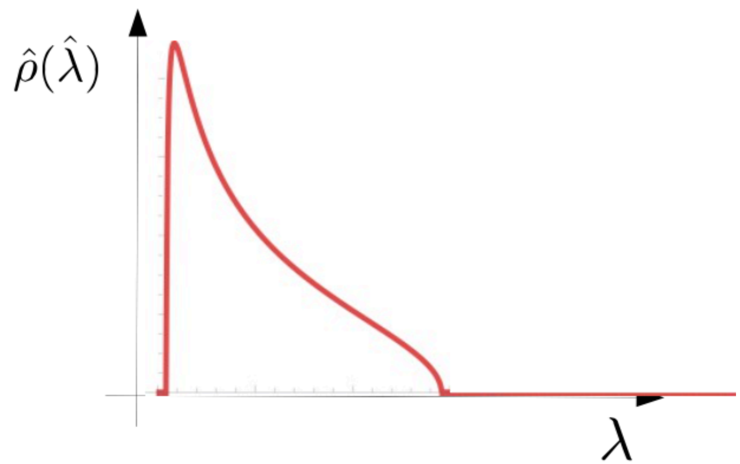
Back to minimal non-trivial model

Correlation matrix: $C = \tau^{-1} = \mathrm{Id} + s\,|e\rangle\langle e|$
(infinite sampling)

Weak noise $r < s^2$



Strong noise $r > s^2$

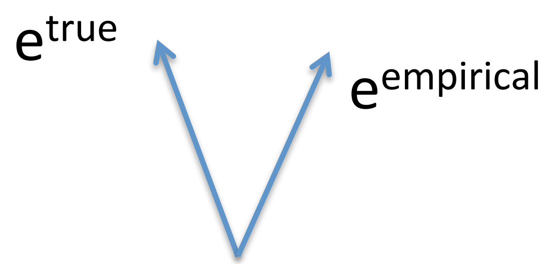


- Phase transition! (Baik, Ben Arous, Peche 2005; Reimann, Van den Broeck, Bex 1996; coined as Retarded Learning by Watkin, Nadal 1994)
- Same phenomenon for any finite nb. K of eigenvalues > 1

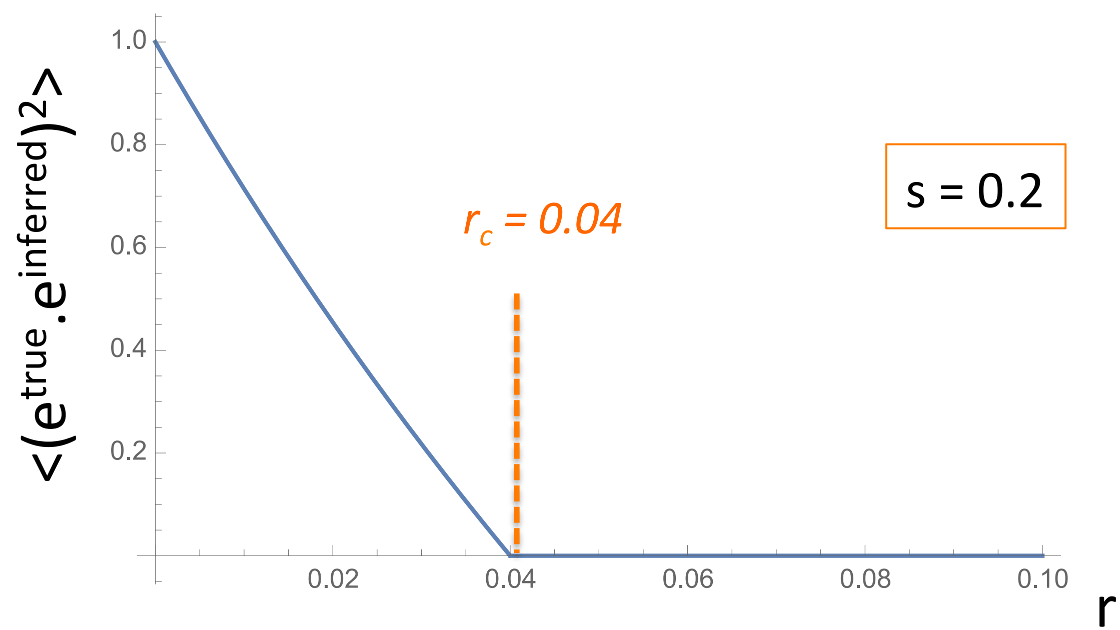
Back to minimal non-trivial model

Correlation matrix:
(infinite sampling)

$$C = \tau^{-1} = \mathrm{Id} + s\,|e\rangle\langle e|$$



?



Unsupervised learning of symmetry-breaking direction from examples

Reimann, Van den Broeck, Bex 1996

- Direction in N-dimension space: $\vec{B},\ |B|^2 = N$
- P examples (i.i.d.): $P_0(\vec{\xi}^{\mu}) \propto \exp\left(-\frac{1}{2}\sum_i \left(\xi_i^{\mu}\right)^2\right) \quad \rightarrow \quad \left|\vec{\xi}^{\mu}\right|^2 \approx N$

$$\boxed{P(\vec{\xi}^{\mu}) \propto P_0(\vec{\xi}^{\mu}) \exp\left(-V\left(\frac{1}{\sqrt{N}}\sum_i \xi_i^{\mu} B_i\right)\right)}$$

$$V\left(\lambda\right) = -\frac{s}{2(1+s)}\lambda^2 \qquad \text{(PCA)}$$

Potential?

$$V\left(\lambda\right) = a\lambda + b\lambda^2 + c\,|\,\lambda\,| + \ldots$$

Unsupervised learning of symmetry-breaking direction from examples

Reimann, Van den Broeck, Bex 1996

- Inference in N-dimension space: $\vec{J},\ |J|^2 = N$
- Bayes: $$\boxed{P(\vec{J}) \propto \exp\left(-\sum_{\mu} V\left(\frac{1}{\sqrt{N}} \sum_i \xi_i^{\mu} J_i \right) \right) \quad \delta\left(\vec{J}^2 - N \right)}$$

Unsupervised learning of symmetry-breaking direction from examples

Reimann, Van den Broeck, Bex 1996

- Inference in N-dimension space: $\vec{J},\ |J|^2 = N$

- Bayes: $$\boxed{P(\vec{J}) \propto \exp\left(-\beta \sum_\mu V\left(\frac{1}{\sqrt{N}} \sum_i \xi_i^\mu J_i\right)\right) \quad \delta\left(\vec{J}^2 - N\right)}$$

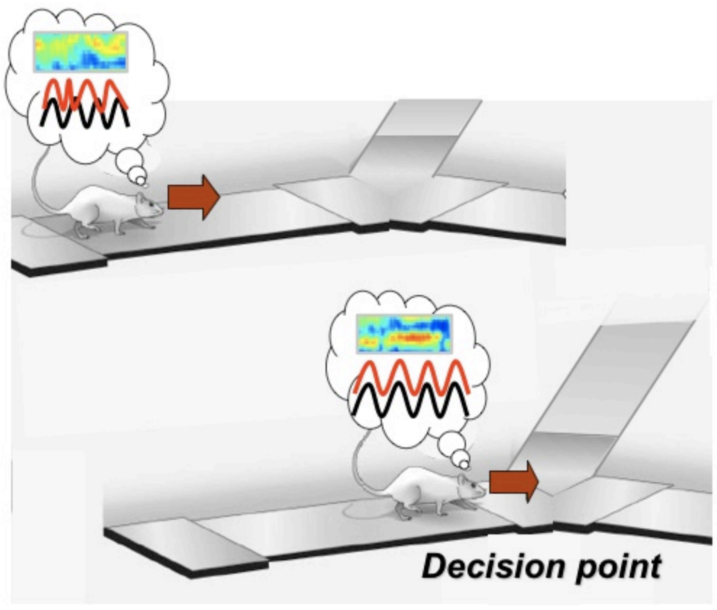
β = 1 Bayes decoding

= infinite ML, MAP

- Crucial quantity: $$\boxed{R = \left[\left\langle \vec{J} \cdot \vec{B} \right\rangle_{P(\vec{J})}\right]_{\left\{\vec{\xi}^\mu\right\}}}$$

(ask me for handwritten notes 2)

Application of PCA & of the Marcenko-Pastur null model to memory consolidation



Peyrache, Battaglia et al.: 37 neural cells recorded in prefrontal cortex

Task: learn correct arm in Y-shaped maze, changed if success rate high enough

Rat perform a rule shift task, with four possible rules

Replay of rule-learning related neural patterns in the prefrontal cortex during sleep A. Peyrache.. F. Battaglia Nature Neuroscience 2009

Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution A.Peyrache … F. Battaglia J. Comput Neurosci. 2009

PCA and analysis of cortical recordings

Activity of prefrontal cortex is recorded during:

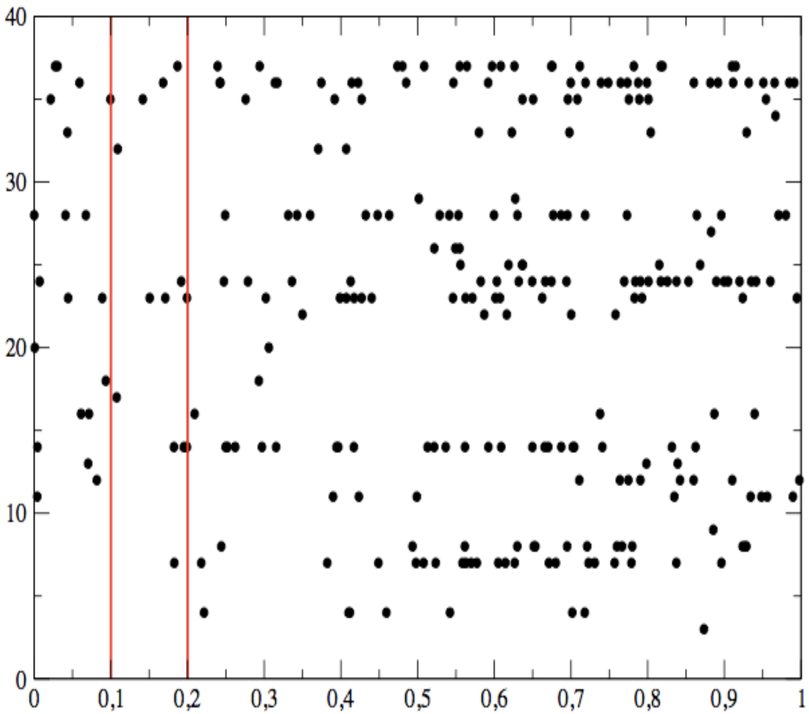
- sleep period **before** the task (PRE)
- task performance
- sleep period **after** the task (POST)

Replay and memory consolidation:

replay of the pattern of activity during the SWS (slow wawe sleep) in period corresponding to coordinated bursts of activity of the hippocampus (sharp waves), Allowing memories to be transferred to prefrontal cortex

1. Spike trains from the awake epoch are binned

neuron



Time(s)

Time is discretized in time windows of size Δt=100 ms

$$s_i^\tau = \begin{cases} 1 \text{ if at least one spike in time window k} \\ 0 \text{ if no spike in time windows k} \end{cases}$$

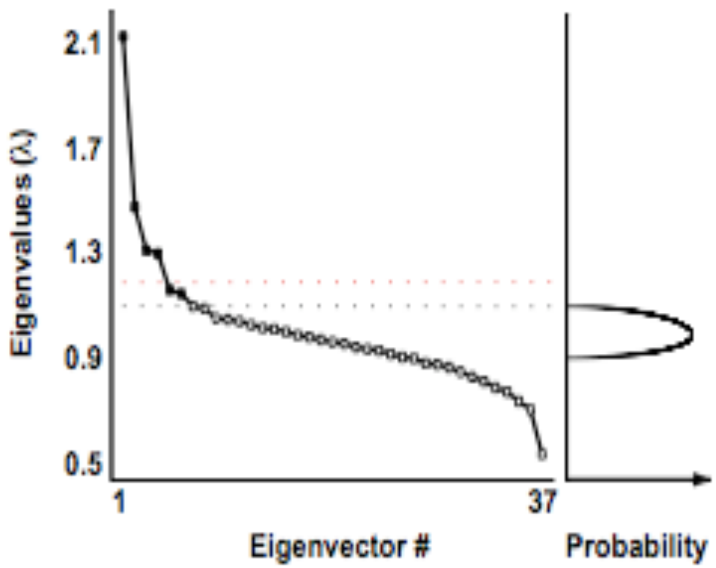
$$p_{kl} = \frac{1}{B} \sum_{\tau=1}^{B} s_k^\tau \, s_l^\tau \; ,$$

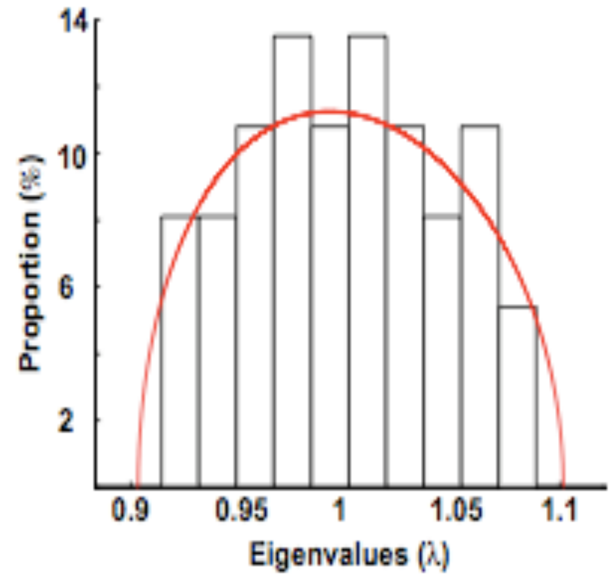
$$p_i = \frac{1}{B} \sum_{\tau=1}^{B} s_i^\tau$$

2. Correlation matrix computed and diagonalized

$$\Gamma_{ij} = \frac{p_{ij} - p_i \, p_j}{\sqrt{p_i (1- p_i) p_j (1- p_j)}}$$

3. Only eigenvectors associated to the largest eigenvalues are retained, threshold value from the upper bound of eigenvalues of correlation matrix of independent, normally distributed spike trains





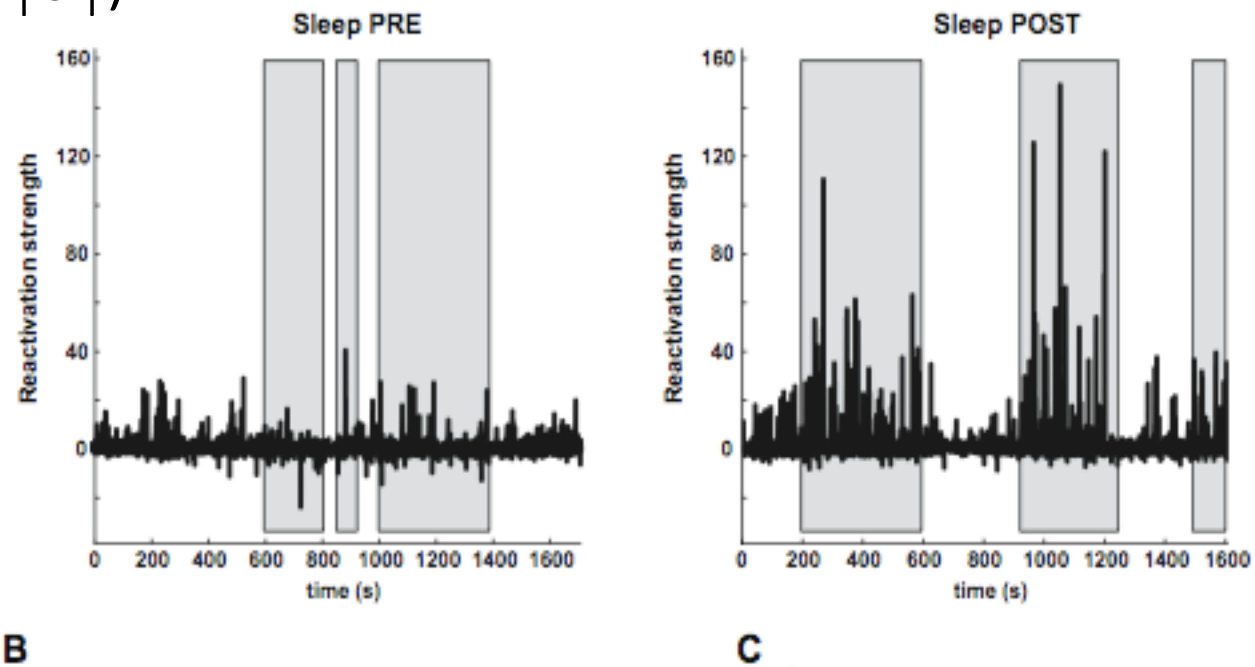
Marcenko-Pastur distribution

4. Spike trains from the sleep epochs are binned

5. The instantaneous similarity of the sleep multi-unit activity with the awake activity is computed through the reactivation strength

$$R^{q}(\tau)=\left(\Sigma_i v^{q}_i s^{\tau}_i\right)^2$$

R^1



- *Instantaneous similarity high in Slow Wave Sleep (SWS) (shaded areas) after learning of the task related to hippocampal sharp waves*

How to cope with too few data/many parameters?

Data
(observations σ)

Model
(parameters τ)

$$p(\tau|\sigma) = \frac{p(\sigma|\tau) \times p(\tau)}{p(\sigma)}$$

Prior knowledge
over model
parameters

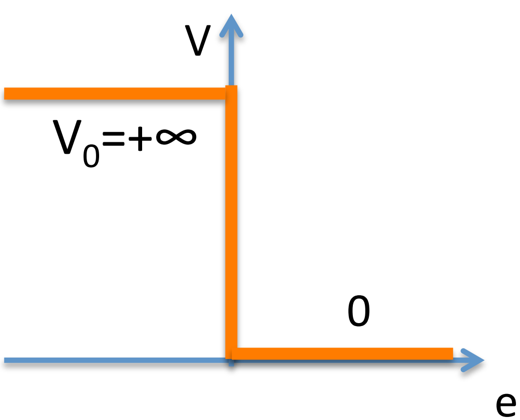
How to beat the phase transition threshold

i.e. to infer |e> when $r > s^2$?

- Log-likelihood: $$\frac{n\ s}{2(1+s)}\sum_{i,j} e_i\ \hat{C}_{ij}\ e_j$$ *(for a normalized vector **e**)*

- Log-Prior over vector e: $$-\sum_i V(e_i)$$ *Prior potential over components*

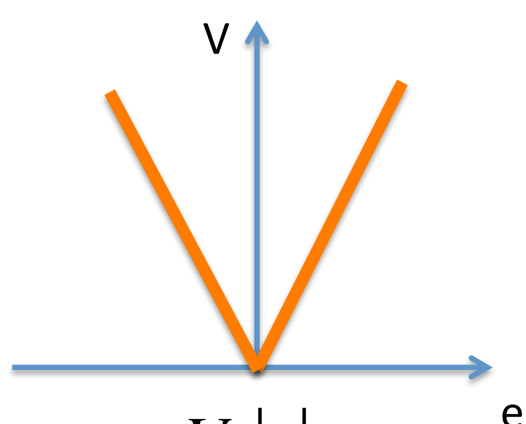
Nonnegative



$r < \underline{2}\ s^2$

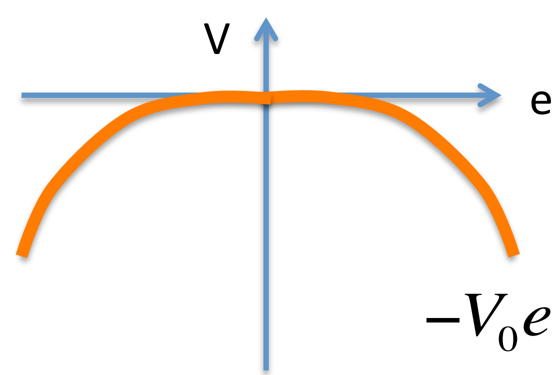
Montanari, Richard (2014)

Sparse (L1)



$V_0|e|$

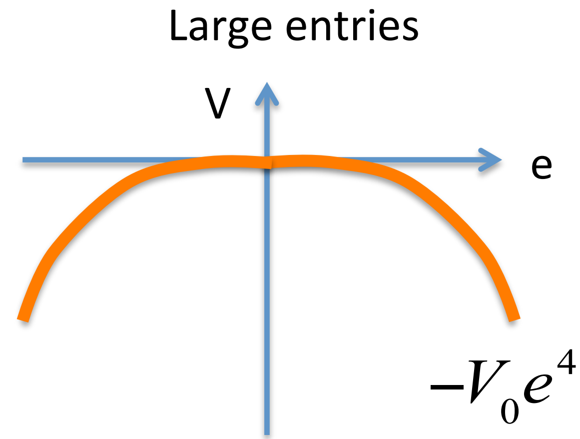
Large entries

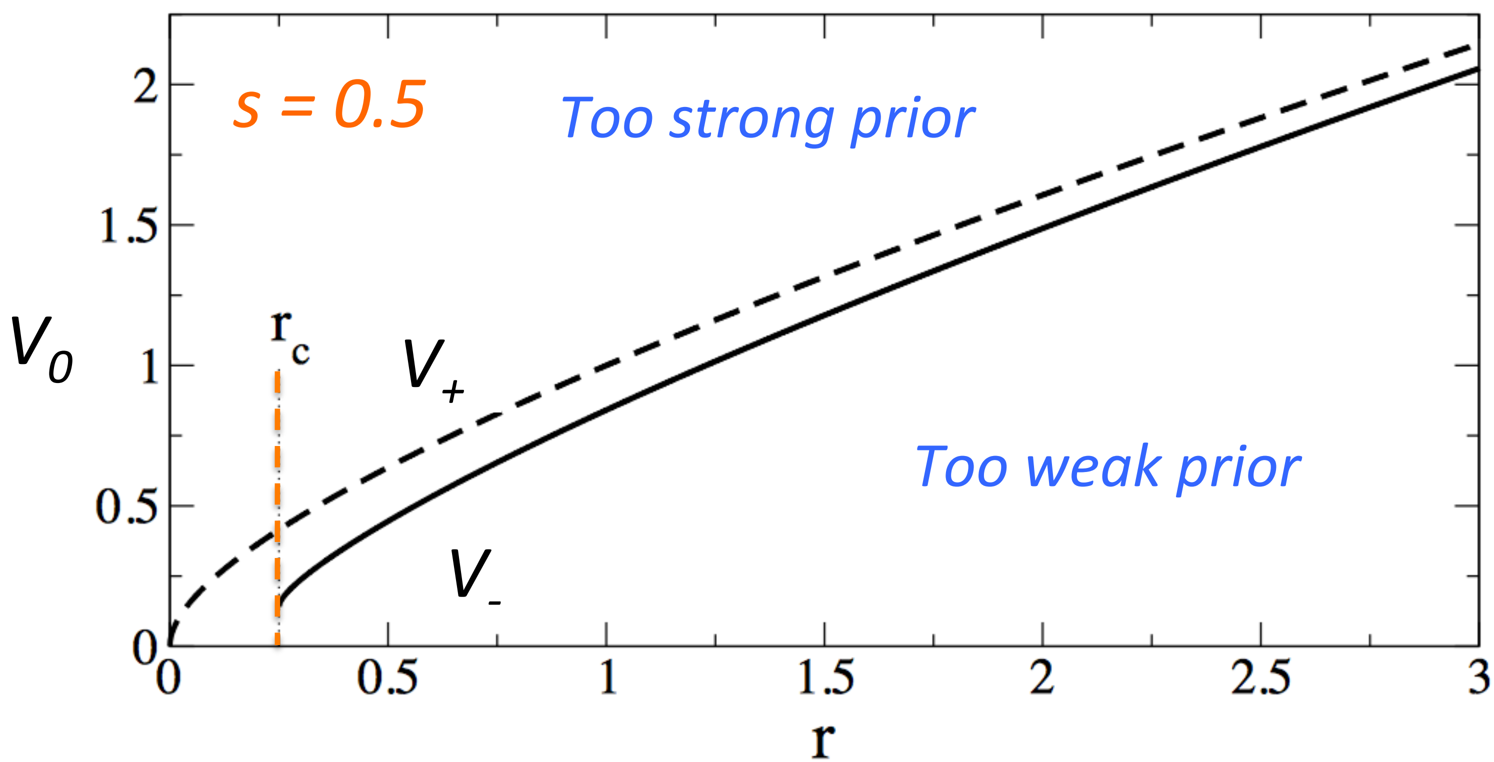


$-V_0 e^4$

Villaimana, R.M. (2015-16)

Can we beat the phase transition threshold?





- **Strength of prior to be chosen carefully …**
- **Beyond PCA: non quadratic potentials in the dot product between data and direction**