Machine Intelligence 2

2.2 ICA: The Infomax method

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

Statistical Independence & Infomax

$$I(X,Y) = KL[P_j(X,Y); P_i(X,Y)] = H(Y) - H(Y|X)$$

Statistical independence

Scenario

observations: $\underline{\mathbf{x}} \in \mathbb{R}^N$ with distribution $P_{\underline{\mathbf{x}}}(\underline{\mathbf{x}})$
estimated sources: $\widehat{\underline{\mathbf{s}}} = \underline{\mathbf{W}} \cdot \underline{\mathbf{x}}$

$\rightsquigarrow$ $P_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}})$: family of true (unknown) densities, parametrized by $\underline{\mathbf{W}}$

$\rightsquigarrow$ $P_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}}) = P_{\underline{\mathbf{s}}}(\underline{\mathbf{W}} \cdot \underline{\mathbf{x}})$

Model selection

$$\widehat{P}_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}}) = \prod_{i=1}^{N} \widehat{P}_{s_i}(\widehat{s}_i) \leftarrow \text{ assumption: statistical independence}$$

$$\mathrm{D}_{\mathrm{KL}} = \mathrm{D}_{\mathrm{KL}}[P_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}}), \widehat{P}_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}})] = \int d\widehat{\underline{\mathbf{s}}} P_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}}) \ln \frac{P_{\underline{\mathbf{s}}}(\widehat{\underline{\mathbf{s}}})}{\prod_{i=1}^{N} \widehat{P}_{s_i}(\widehat{s}_i)} \stackrel{!}{=} \min_{\underline{\mathbf{W}}}$$

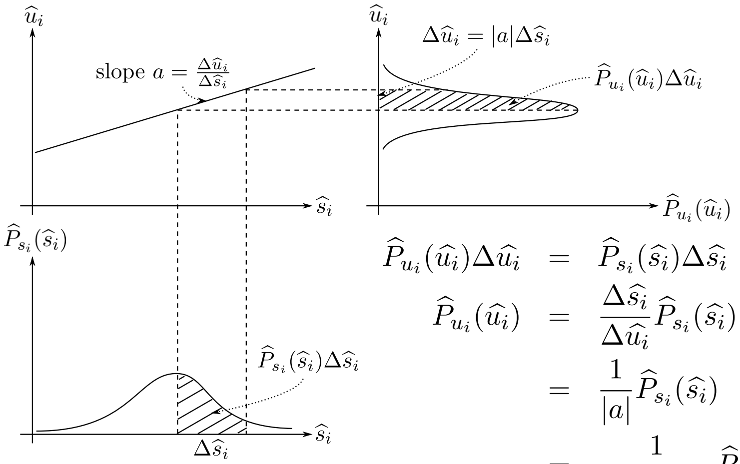
Equally distributed sources

non-linear transformations: $\widehat{u}_i = \widehat{f}_i(\widehat{s}_i)$, such that $\widehat{P}_{u_i}(\widehat{u}_i) = \text{const.}$

conservation of probability

$$\widehat{P}_{u_i}(\widehat{u}_i) d\widehat{u}_i = \widehat{P}_{s_i}(\widehat{s}_i) d\widehat{s}_i$$

Conservation of probability



$$
\begin{aligned}
\widehat{P}_{u_i}(\widehat{u}_i) \Delta \widehat{u}_i &= \widehat{P}_{s_i}(\widehat{s}_i) \Delta \widehat{s}_i \\
\widehat{P}_{u_i}(\widehat{u}_i) &= \frac{\Delta \widehat{s}_i}{\Delta \widehat{u}_i} \widehat{P}_{s_i}(\widehat{s}_i) \\
&= \frac{1}{|a|} \widehat{P}_{s_i}(\widehat{s}_i) \\
&= \frac{1}{\left| d\widehat{u}_i / d\widehat{s}_i \right|} \widehat{P}_{s_i}(\widehat{s}_i) \\
&= \left| \frac{d\widehat{s}_i}{d\widehat{u}_i} \right| \widehat{P}_{s_i}(\widehat{s}_i)
\end{aligned}
$$

Equally distributed sources

non-linear transformations: $\widehat{u}_i = \widehat{f}_i(\widehat{s}_i)$, such that $\widehat{P}_{u_i}(\widehat{u}_i) = \text{const.}$

conservation of probability

$$\widehat{P}_{u_i}(\widehat{u}_i) d\widehat{u}_i = \widehat{P}_{s_i}(\widehat{s}_i) d\widehat{s}_i$$

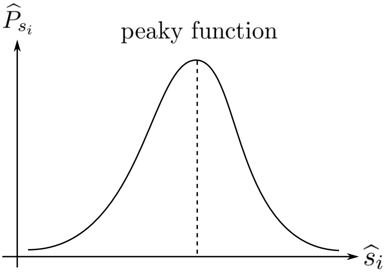
$$\widehat{P}_{u_i}(\widehat{u}_i) \quad = \quad \left| \frac{d\widehat{s}_i}{d\widehat{u}_i} \right| \widehat{P}_{s_i}(\widehat{s}_i) \quad = \quad \frac{1}{\left| \widehat{f}_i'(\widehat{s}_i) \right|} \widehat{P}_{s_i}(\widehat{s}_i) \quad \stackrel{!}{=} 1$$

It follows:

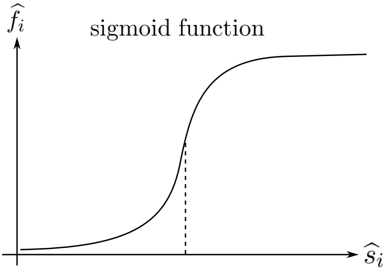
$$\left| \widehat{f}_i'(\widehat{s}_i) \right| = \widehat{P}_{s_i}(\widehat{s}_i) \qquad \Rightarrow \widehat{f}_i(\widehat{s}_i) = \int_{-\infty}^{\widehat{s}_i} dy \widehat{P}_{s_i}(y)$$

$\widehat{f}_i$: cumulative density function (cdf) of $\widehat{P}_{s_i}(\widehat{s}_i)$

Equally distributed sources



pdf



cdf

The Infomax principle

Statistical independence:

$$\mathrm{D_{KL}} = \int d\underline{\widehat{\mathbf{s}}} P_{\underline{\mathbf{s}}}(\underline{\widehat{\mathbf{s}}}) \ln \frac{P_{\underline{\mathbf{s}}}(\underline{\widehat{\mathbf{s}}})}{\prod_{i=1}^{N} \widehat{P}_{s_i}(\widehat{s}_i)} \overset{!}{=} \min$$

Transformation:

$$\widehat{u}_i = \widehat{f}_i\big(\underbrace{\underline{\mathbf{e}}_i^T \underline{\mathbf{W}} \cdot \underline{\mathbf{x}}}_{\widehat{s}_i}\big)$$

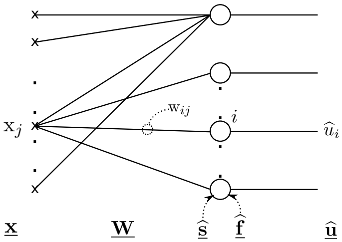
see blackboard

The Infomax principle (Bell, Sejnowski, 1995)

$$H = -\int d\underline{\widehat{\mathbf{u}}} P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \ln P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \overset{!}{=} \max_{\underline{\mathbf{W}}}$$

Perceptron implementation

Architecture



perceptron: $\widehat{u}_i = \widehat{f}_i\Big(\sum_j \mathrm{w}_{ij}\mathrm{x}_j\Big)$ observations: $\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N, \alpha = 1, \ldots, p$

Cost function

$$H = -\int d\underline{\widehat{\mathbf{u}}} P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \ln P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \stackrel{!}{=} \max_{\underline{\mathbf{W}}}$$

Empirical Risk Minimization

$$H = -\int d\underline{\widehat{\mathbf{u}}} P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \ln P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \stackrel{!}{=} \max_{\underline{\mathbf{W}}}$$

see blackboard

Empirical Risk Minimization

$$H = -\int d\underline{\widehat{\mathbf{u}}} P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \ln P_{\underline{\mathbf{u}}}(\underline{\widehat{\mathbf{u}}}) \stackrel{!}{=} \max_{\underline{\mathbf{W}}}$$

Model selection: maximize E^G

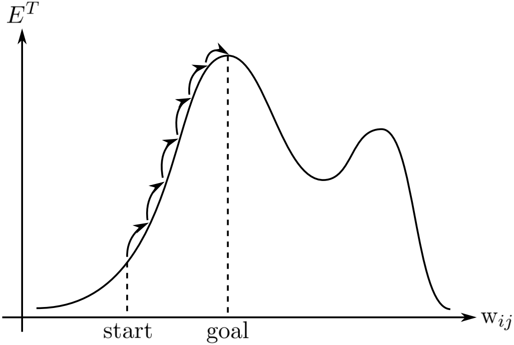
$$E^G = \ln |\det \underline{\mathbf{W}}| + \int d\underline{\mathbf{x}} P_{\underline{\mathbf{x}}}(\underline{\mathbf{x}}) \left\{ \sum_{l=1}^{N} \ln \widehat{f}_l' \left(\sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k \right) \right\}$$

ERM principle: mathematical expectation $E^G \longrightarrow$ empirical average E^T

$$E^T = \ln |\det \underline{\mathbf{W}}| + \frac{1}{p} \sum_{\alpha=1}^{p} \sum_{l=1}^{N} \ln \widehat{f}_l' \left(\sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k^{(\alpha)} \right)$$

Gradient based optimization

Gradient Ascent



$$\Delta w_{ij} = \underbrace{\varepsilon}_{\text{learning step}} \cdot \frac{\partial E^T}{\partial w_{ij}}$$

Gradient ascent on the training cost.

Gradient based optimization

Batch Learning: $\Delta \mathrm{w}_{ij} = \frac{\partial E^T}{\partial \mathrm{w}_{ij}} = \frac{\varepsilon}{p} \sum_\alpha \frac{\partial e^{(\alpha)}}{\partial \mathrm{w}_{ij}}$

On-line learning: $\Delta \mathrm{w}_{ij} = \varepsilon \frac{\partial e^{(\alpha)}}{\partial \mathrm{w}_{ij}}$ and time-dependent ($\searrow$) learning rate ε

Gradient based optimization

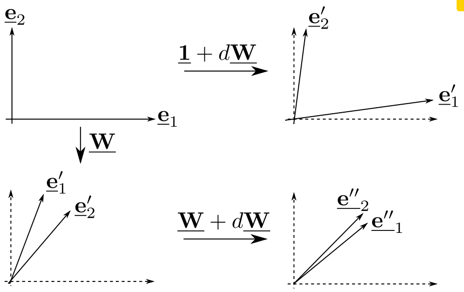
Batch Learning: $\Delta \mathrm{w}_{ij} = \frac{\partial E^T}{\partial \mathrm{w}_{ij}} = \frac{\varepsilon}{p} \sum_\alpha \frac{\partial e^{(\alpha)}}{\partial \mathrm{w}_{ij}}$

On-line learning: $\Delta \mathrm{w}_{ij} = \varepsilon \frac{\partial e^{(\alpha)}}{\partial \mathrm{w}_{ij}}$ and time-dependent ($\searrow$) learning rate ε

$$
\begin{aligned}
e^{(\alpha)} &= \ln |\det \underline{\mathbf{W}}| + \sum_{l=1}^{N} \ln \widehat{f}_l' \left(\sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k^{(\alpha)} \right) \\
\frac{\partial e^{(\alpha)}}{\partial \mathrm{w}_{ij}} &= \underbrace{\left(\underline{\mathbf{W}}^{-1}\right)_{ji}}_{\text{costly computation}} + \frac{\widehat{f}_i'' \left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik} \mathrm{x}_k^{(\alpha)} \right)}{\widehat{f}_i' \left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik} \mathrm{x}_k^{(\alpha)} \right)} \cdot \mathrm{x}_j^{(\alpha)}
\end{aligned}
$$

Natural Gradient

linear transformations: $d\underline{\mathbf{W}}, \underline{\mathbf{W}}$



Normalized step size:

$$d\underline{\mathbf{Z}} = \underbrace{d\underline{\mathbf{W}}}_{\text{then do } d\underline{\mathbf{W}}} \cdot \underbrace{\underline{\mathbf{W}}^{-1}}_{\text{transform back to } \underline{\mathbf{1}}}$$

$\Rightarrow$ make learning steps "comparable"

Natural Gradient

Taylor expansion of e ($e^{(\alpha)}$, but α suppressed in the following):

$$\begin{aligned} e_{(\underline{\mathbf{W}}+d\underline{\mathbf{W}})} &= e_{(\underline{\mathbf{W}})} + \nabla e_{(\underline{\mathbf{W}})} d\underline{\mathbf{W}} \\ &\underbrace{=}_{d\underline{\mathbf{W}}=\varepsilon\underline{\mathbf{D}}_{\mathrm{w}}} e_{(\underline{\mathbf{W}})} + \varepsilon\left[\nabla e_{(\underline{\mathbf{W}})}\right]^T \cdot \underline{\mathbf{D}}_{\mathrm{w}} \end{aligned}$$

learning step:

$$\begin{aligned} d\underline{\mathbf{Z}} &= d\underline{\mathbf{W}} \cdot \underline{\mathbf{W}}^{-1} \\ &= \varepsilon\underline{\mathbf{D}}_{\mathrm{w}} \cdot \underline{\mathbf{W}}^{-1} \end{aligned}$$

direction of steepest ascent under normalized step-size:

$$\begin{aligned} \left[\nabla e_{(\underline{\mathbf{W}})}\right]\underline{\mathbf{D}}_{\mathrm{w}} &\overset{!}{=} \max_{\underline{\mathbf{D}}_{\mathrm{w}}} \\ \left(\underline{\mathbf{D}}_{\mathrm{w}} \cdot \underline{\mathbf{W}}^{-1}\right)^2 &\overset{!}{=} 1 \end{aligned}$$

Natural Gradient

Solution using Lagrange multipliers:

$$\sum_{i,j=1}^{N} \frac{\partial e}{\partial \mathrm{w}_{ij}} \left(\underline{\mathbf{D}}_{\mathrm{w}}\right)_{ij} - \lambda \sum_{i,j,k,l=1}^{N} \left(\underline{\mathbf{D}}_{\mathrm{w}}\right)_{ij} \left(\underline{\mathbf{W}}^{-1}\right)_{jl} \left(\underline{\mathbf{D}}_{\mathrm{w}}\right)_{ik} \left(\underline{\mathbf{W}}^{-1}\right)_{kl} \stackrel{!}{=} \max_{\underline{\mathbf{D}}_{\mathrm{w}}}$$

Taking the derivative w.r.t. $(\underline{\mathbf{D}})_{ps}$ and setting to zero yields:

$$\begin{aligned}
\frac{\partial e}{\partial (\underline{\mathbf{D}}_{\mathrm{w}})_{ps}} = \frac{\partial e}{\partial (\underline{\mathbf{W}})_{ps}} &- \lambda \sum_{k,l=1}^{N} \left(\underline{\mathbf{W}}^{-1}\right)_{sl} \left(\underline{\mathbf{D}}_{\mathrm{w}}\right)_{pk} \left(\underline{\mathbf{W}}^{-1}\right)_{kl} \\
&- \lambda \sum_{i,j=1}^{N} \left(\underline{\mathbf{D}}_{\mathrm{w}}\right)_{pj} \left(\underline{\mathbf{W}}^{-1}\right)_{jl} \left(\underline{\mathbf{W}}^{-1}\right)_{sl} \\
= \frac{\partial e}{\partial (\underline{\mathbf{W}})_{ps}} &- 2\lambda \sum_{k,l=1}^{N} \left(\underline{\mathbf{D}}_{\mathrm{w}}\right)_{pk} \left(\underline{\mathbf{W}}^{-1}\right)_{kl} \left(\underline{\mathbf{W}}^{-1}\right)_{sl} \stackrel{!}{=} 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial e}{\partial \underline{\mathbf{W}}} &= 2\lambda \underline{\mathbf{D}}_{\mathrm{w}} \underline{\mathbf{W}}^{-1} \left(\underline{\mathbf{W}}^{-1}\right)^{T} \\
\underline{\mathbf{D}}_{\mathrm{w}} &= \frac{1}{2\lambda} \frac{\partial e}{\partial \underline{\mathbf{W}}} \underline{\mathbf{W}}^{T} \underline{\mathbf{W}}
\end{aligned}$$

Natural Gradient

$$e_{(\underline{\mathbf{W}}+d\underline{\mathbf{W}})} = e_{(\underline{\mathbf{W}})} + \varepsilon \left[\nabla e_{(\underline{\mathbf{W}})}\right] \cdot \underline{\mathbf{D}}_{\mathrm{w}}$$

Inserting the optimal direction for "natural" gradient ascent yields:

$$\Delta\underline{\mathbf{W}} = \varepsilon \frac{\overbrace{\partial e}^{\text{"original" gradient}}}{\partial \underline{\mathbf{W}}} \; \underbrace{\underline{\mathbf{W}}^T \underline{\mathbf{W}}}_{\text{normalization of step size}}$$

i.e.

$$\Delta \mathrm{w}_{ij} = \varepsilon \sum_{l=1}^{N} \left\{ \delta_{il} + \frac{\widehat{f}_i''\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik}\mathrm{x}_k^{(\alpha)}\right)}{\widehat{f}_i'\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik}\mathrm{x}_k^{(\alpha)}\right)} \sum_{k=1}^{N} \mathrm{w}_{lk}\mathrm{x}_k^{(\alpha)} \right\} \mathrm{w}_{lj}$$

Summary: The Infomax method

data $\quad \underline{\mathbf{x}}^{(\alpha)}, \quad \alpha = 1, \ldots, p$

$\Downarrow$

model $\quad \hat{\mathrm{u}}_i = \hat{f}_i(\overbrace{\underline{\mathbf{e}}_i^T \cdot \underline{\mathbf{W}} \cdot \underline{\mathbf{x}}}^{\hat{s}_i})$

$\Downarrow$

performance measure

$$H = -\int d\widehat{\underline{\mathbf{u}}} P_{\underline{\mathbf{u}}}(\widehat{\underline{\mathbf{u}}}) \ln P_{\underline{\mathbf{u}}}(\widehat{\underline{\mathbf{u}}}) \stackrel{!}{=} \max_{\underline{\mathbf{W}}}$$

$$E^T = \ln|\det \underline{\mathbf{W}}| + \frac{1}{p} \sum_{\alpha=1}^{p} \sum_{l=1}^{N} \ln \widehat{f}_l' \left(\sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k^{(\alpha)} \right) \stackrel{!}{=} \max_{\underline{\mathbf{W}}}$$

$\Downarrow$

optimization $\quad$ Natural Gradient ascent on E^T

Practical aspects: Source amplitudes

Problem: Undetermined source amplitudes $\rightsquigarrow$ convergence problems

1. Bell-Sejnowski solution:

$$\Delta \mathrm{w}_{ii} = 0 \text{ and } \mathrm{w}_{ii} = 1 \text{ for all } i$$

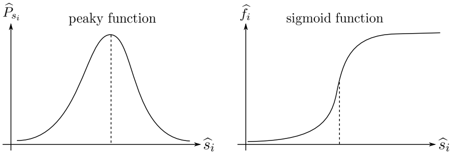
2. Amari solution: Learning steps are always orthogonal to subspace of equivalent unmixing matrices.

$$\Delta \mathrm{w}_{ij} = \varepsilon \frac{\widehat{f}_i''\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik}\mathrm{x}_k^{(\alpha)}\right)}{\widehat{f}_i'\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik}\mathrm{x}_k^{(\alpha)}\right)} \sum_{l \neq i}^{N} \left(\sum_{k=1}^{N} \mathrm{w}_{lk}\mathrm{x}_k^{(\alpha)}\right) \mathrm{w}_{lj}$$

Practical aspects: Choice of $\widehat{f}_i$

Problem: True distribution (of u and s) and its cumulative distribution function is **unknown**.

Solution: Choose peaky PDF $\rightsquigarrow$ sigmoid CDF



pdf cdf

typical choice:

$$\widehat{f}_{(y)} = \frac{1}{1+\exp(-y)} \Rightarrow \frac{\widehat{f}''_{(y)}}{\widehat{f}'_{(y)}} = 1 - 2\widehat{f}_{(y)}$$

Observation: ICA is fairly robust against details of the choice of $\widehat{f}$!

Practical aspects: Choice of $\widehat{f}_i$

Natural Gradient (batch):

$$\Delta \mathrm{w}_{ij} = \varepsilon \sum_{l=1}^{N} \left\{ \delta_{il} + \frac{1}{p} \sum_{\alpha=1}^{p} \frac{\widehat{f}_i''\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik} \mathrm{x}_k^{(\alpha)} \right)}{\widehat{f}_i'\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik} \mathrm{x}_k^{(\alpha)} \right)} \sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k^{(\alpha)} \right\} \mathrm{w}_{lj}$$

Stationary state:

$$\Delta \underline{\mathbf{w}}_{ij} \stackrel{!}{=} 0 \quad \Rightarrow \quad \delta_{il} \stackrel{!}{=} -\frac{1}{p} \sum_{\alpha=1}^{p} \underbrace{\frac{\widehat{f}_i''\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik} \mathrm{x}_k^{(\alpha)} \right)}{\widehat{f}_i'\left(\sum\limits_{k=1}^{N} \mathrm{w}_{ik} \mathrm{x}_k^{(\alpha)} \right)}}_{\varphi_i\left(\widehat{s}_i^{(\alpha)}\right)} \cdot \underbrace{\sum_{k=1}^{N} \mathrm{w}_{lk} \mathrm{x}_k^{(\alpha)}}_{\widehat{s}_l^{(\alpha)}}$$

Practical aspects: Choice of $\widehat{f}_i$

$$-\frac{1}{p}\sum_{\alpha=1}^{p} \varphi_i\big(\widehat{s}_i^{(\alpha)}\big)\,\widehat{s}_l^{(\alpha)} \quad \stackrel{!}{=} \quad \delta_{il}$$

Ansatz: $\widehat{s}_i = \lambda_i s_i \rightarrow$ estimated $\sim$ true source signals

$i = l$: Through proper choice of λ_i we can always fulfill:

$$-\frac{1}{p}\sum_{\alpha=1}^{p} \varphi_i\Big(\widehat{s}_i^{(\alpha)}\Big)\lambda_i s_i^{(\alpha)} \stackrel{!}{=} 1$$

$i \neq l$: Limit of large number of observations:

$$\frac{1}{p}\sum_{\alpha=1}^{p} \varphi_i\Big(\widehat{s}_i^{(\alpha)}\Big)\lambda_l s_l^{(\alpha)} \rightarrow \Big\langle \varphi_i\Big(\widehat{s}_i^{(\alpha)}\Big)\lambda_l s_l^{(\alpha)} \Big\rangle_{P_{\underline{\mathbf{s}}}}$$

$$\Big\langle \varphi_i\big(\lambda_i s_i\big)\lambda_l s_l \Big\rangle \underbrace{=}_{\substack{\text{statistical} \\ \text{independence}}} \Big\langle \varphi_i\big(\lambda_i s_i\big)\Big\rangle\Big\langle \lambda_l s_l \Big\rangle \stackrel{!}{=} 0$$

Can always be fulfilled if data is centered: $< s_l >= 0$

Practical aspects: Choice of $\widehat{f}_i$

True (independent) source signals are always a fixed point of the natural gradient ascent $\rightarrow$ independent of choice of $\widehat{f}_i$

However: if $\widehat{f}_i$ deviates too strongly from its true shape, the fixed point may become unstable:

$\Rightarrow$ if in doubt (and enough data available):

- $\rightsquigarrow$ make a parametrized ansatz for $\widehat{f}_i$
- $\rightsquigarrow$ estimate parameters in addition to $\underline{\mathbf{w}}$