Méthode de séparation de sources
Modèles et algorithmes

Applications en Astrophysique

Statistical approaches - take II

from linear to non-linear models

J.Bobin/C. Kervazo

jerome.bobin@cea.fr - christophe.kervazo@telecom-paris.fr

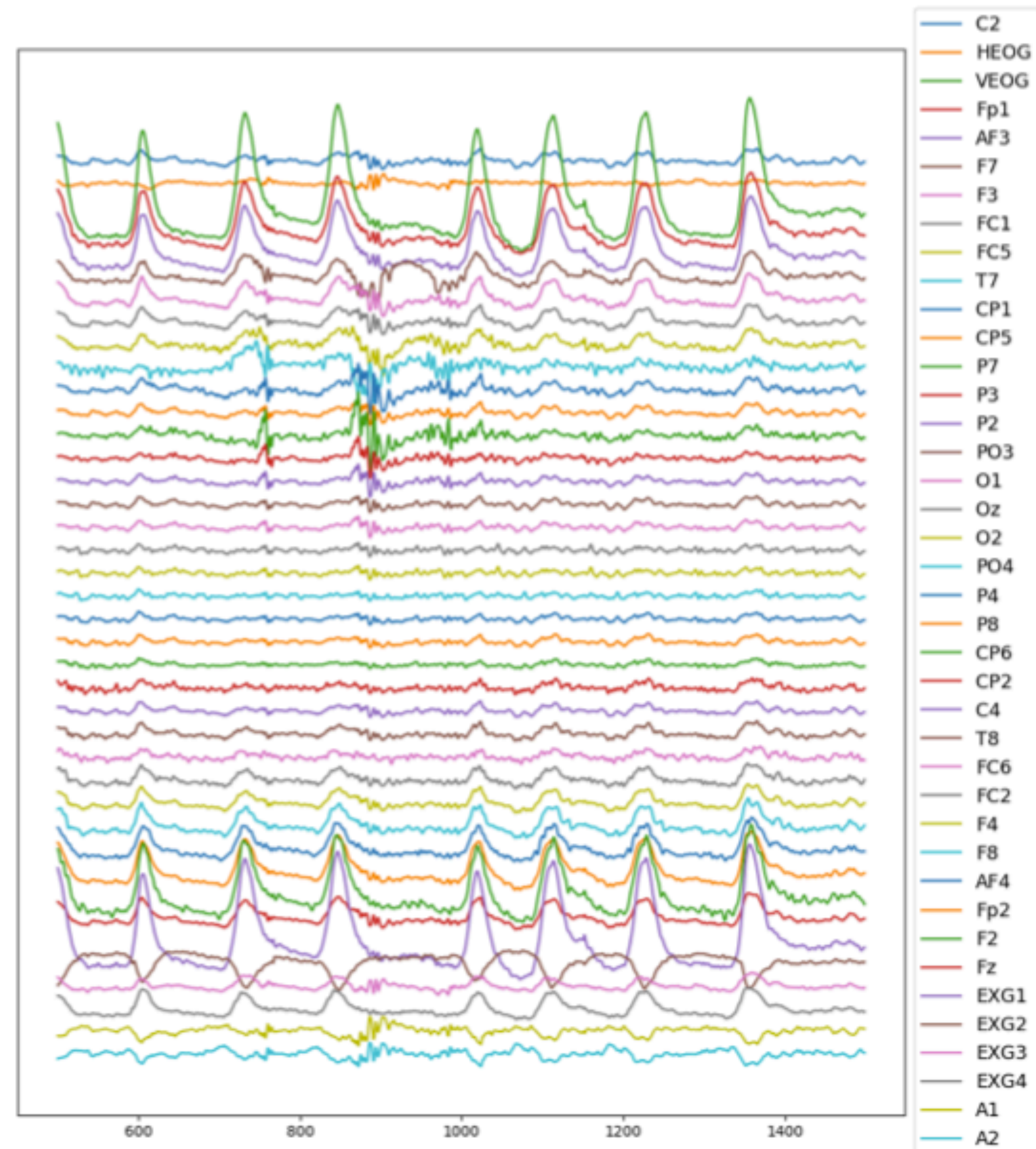
From linear to non-linear mixture models

General principles

Context

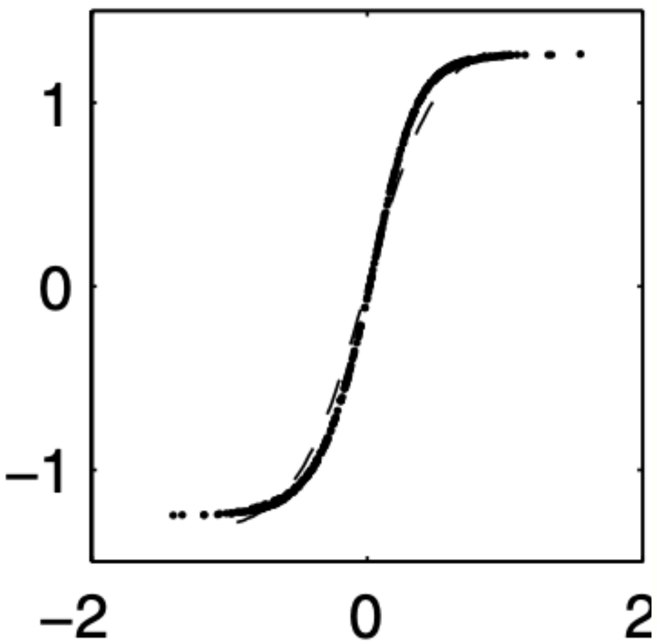
Partiture #1

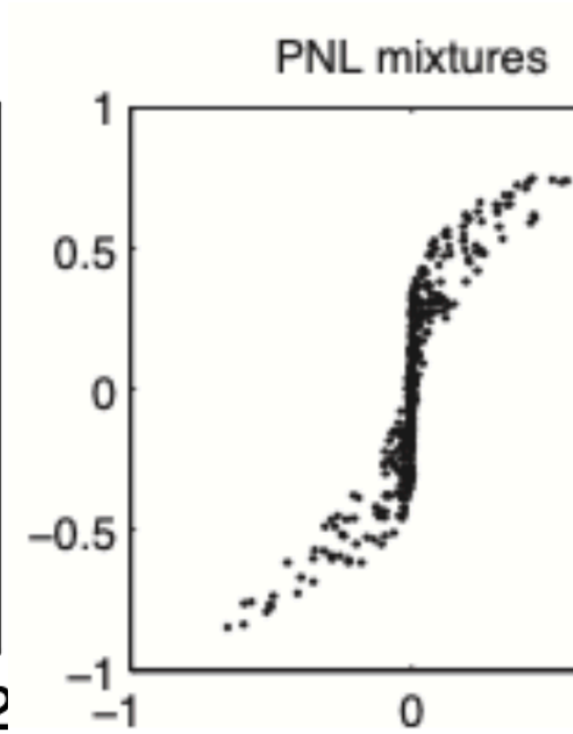
Show-through images



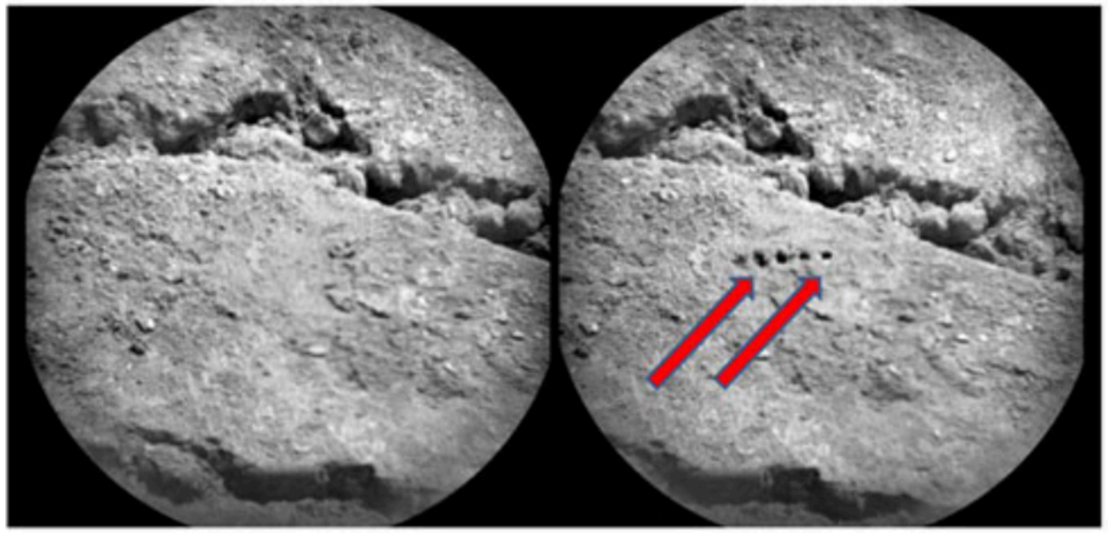
EEG

Context

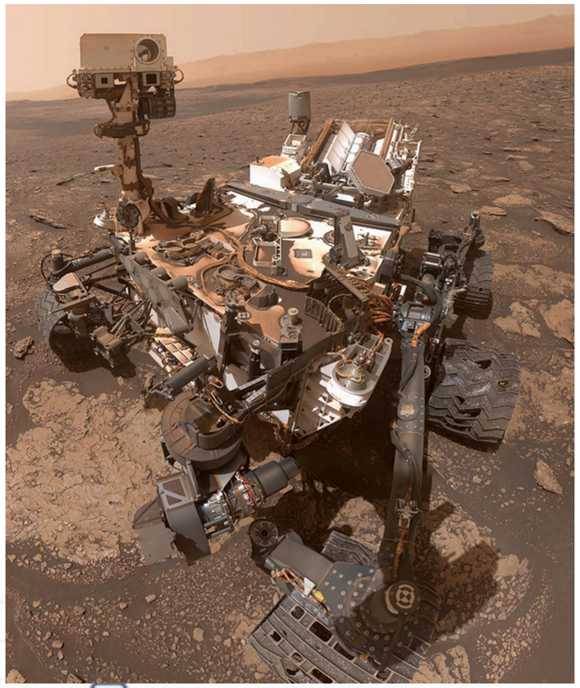


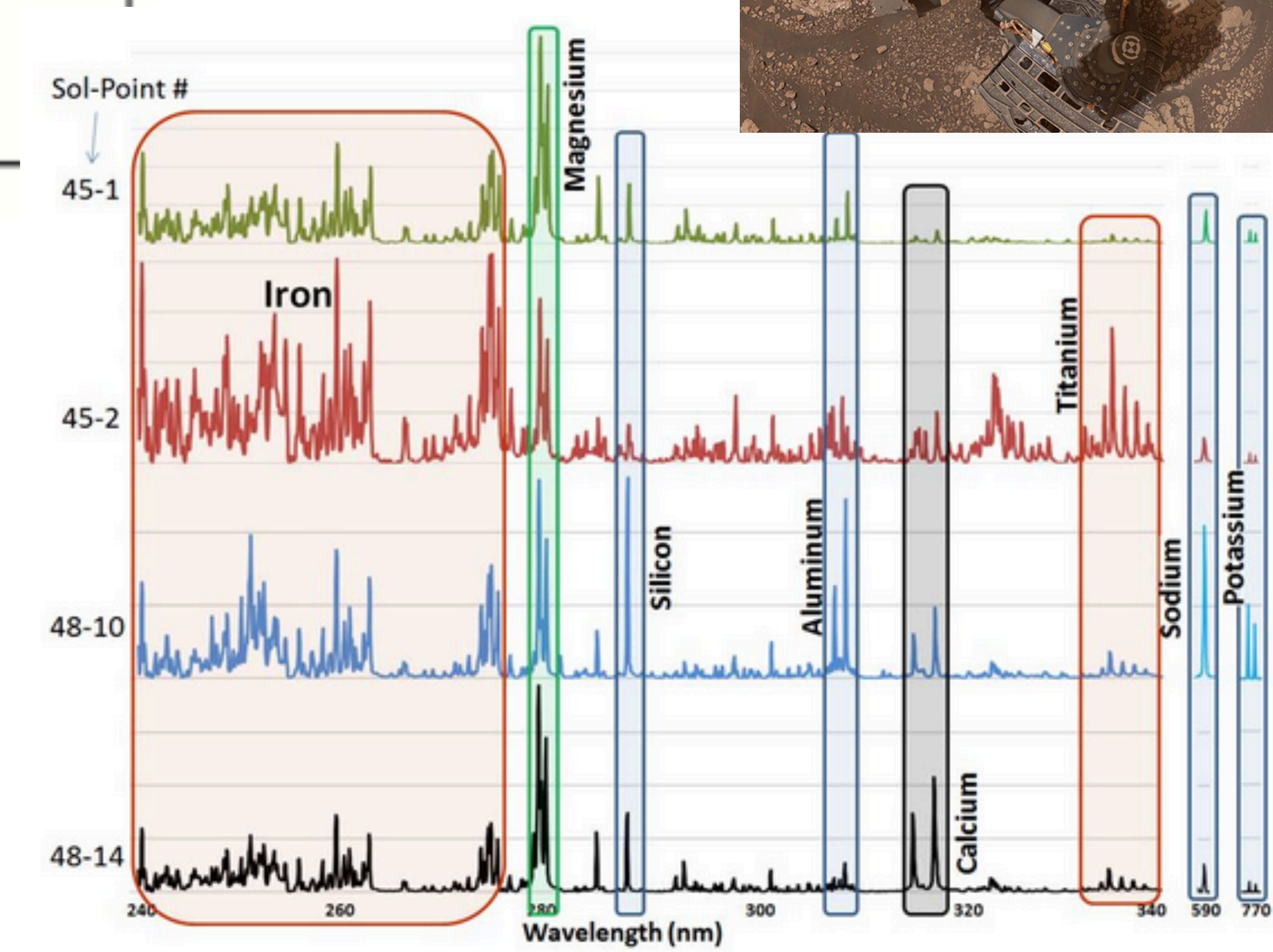


Non-linear sensor response

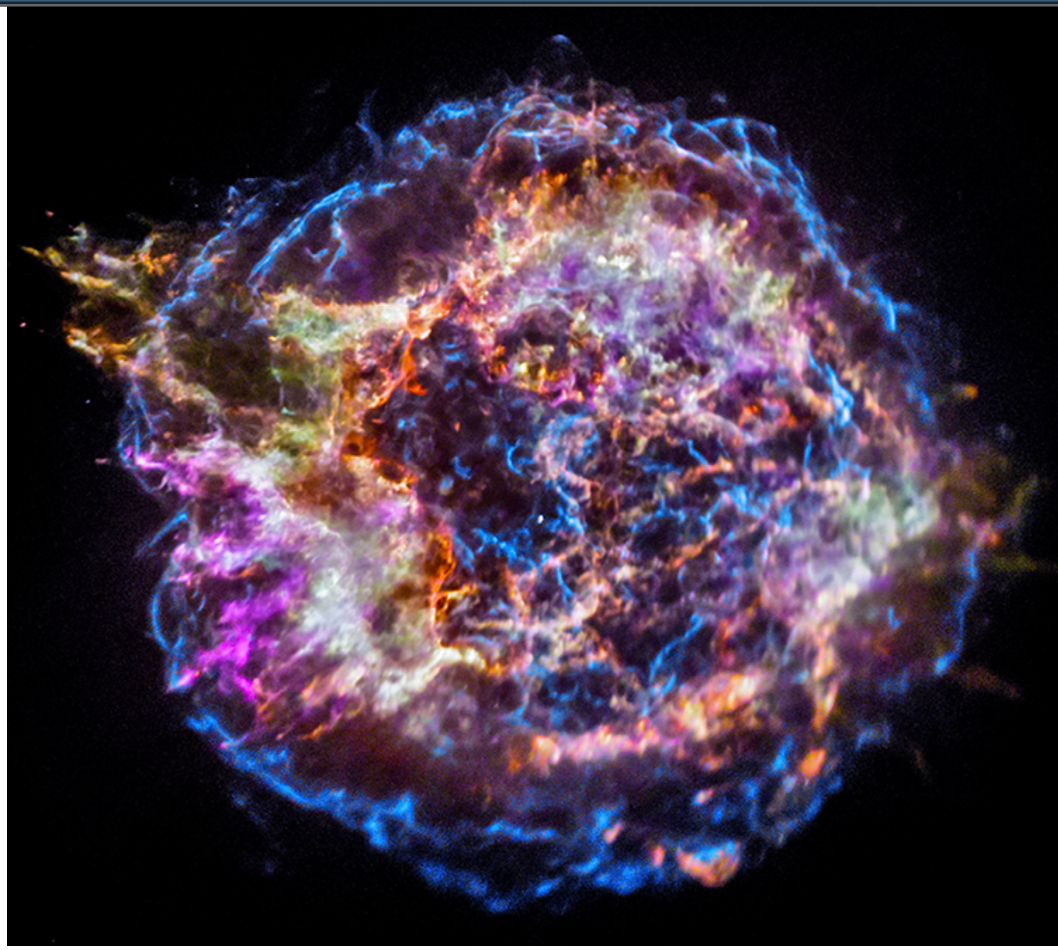


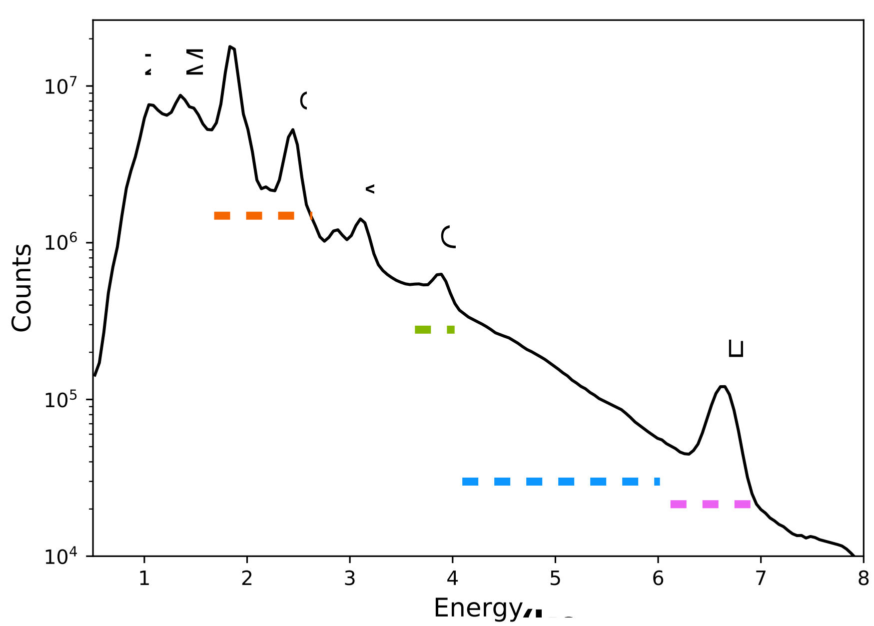
Soil target named "Beechey"

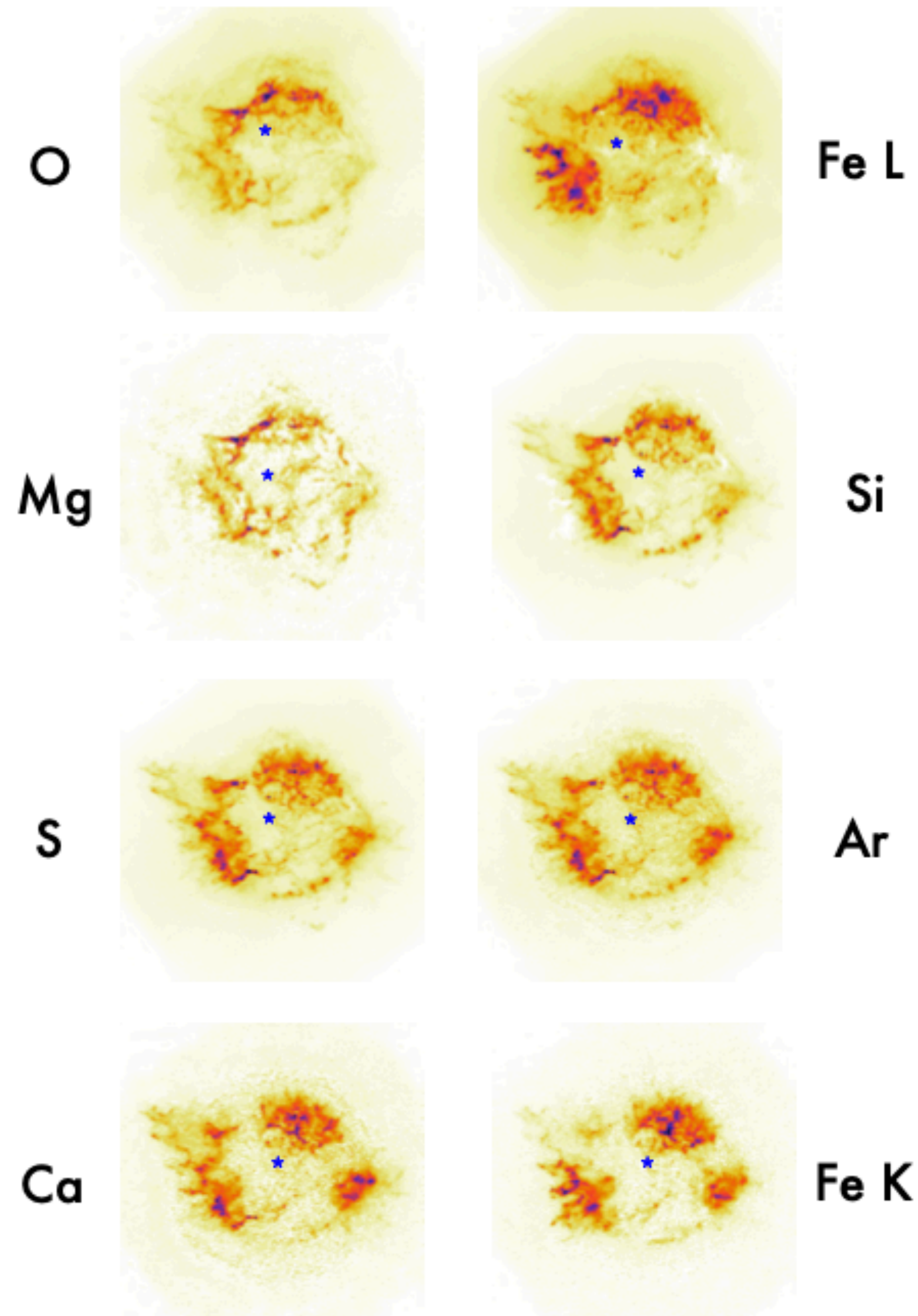




LIBS data from the mars rover curiosity







Special types of non-stationary mixture models

Models' taxonomy

Post non-linear mixture models

$$\mathbf{X} = f(\mathbf{AS}) + \mathbf{N}$$

Composition of a non-linearity with a linear mixture model
e.g. non-linear sensor response

General mixture models

$$\mathbf{X} = f(s_1, \cdots, s_n) + \mathbf{N}$$

General non-linear mixture model
… but f is assumed to be invertible

Post-non linear mixture models

Post non-linear mixture model

Post non-linear mixture models

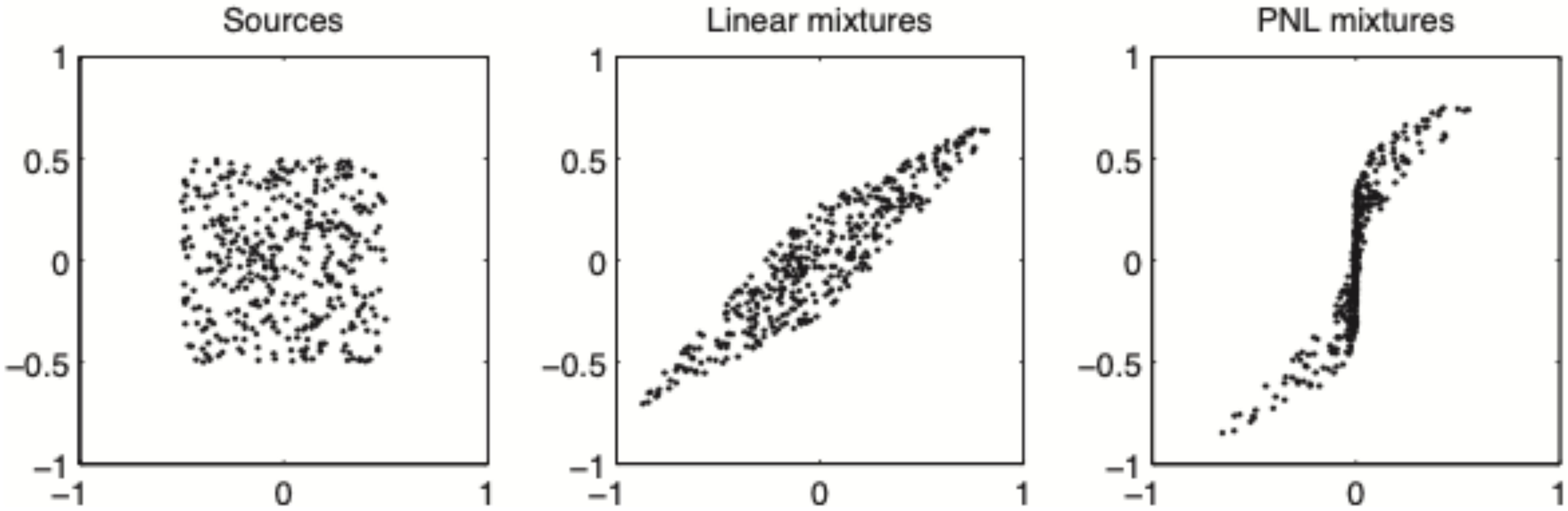
$$x_i^t = f_i\left(\sum_j a_{ij} s_j^t\right) + n_i^t$$

f is invertible and applies to each entry independently

$$\mathbf{X} = (f \circ \mathbf{A})(\mathbf{S}) + \mathbf{N}$$

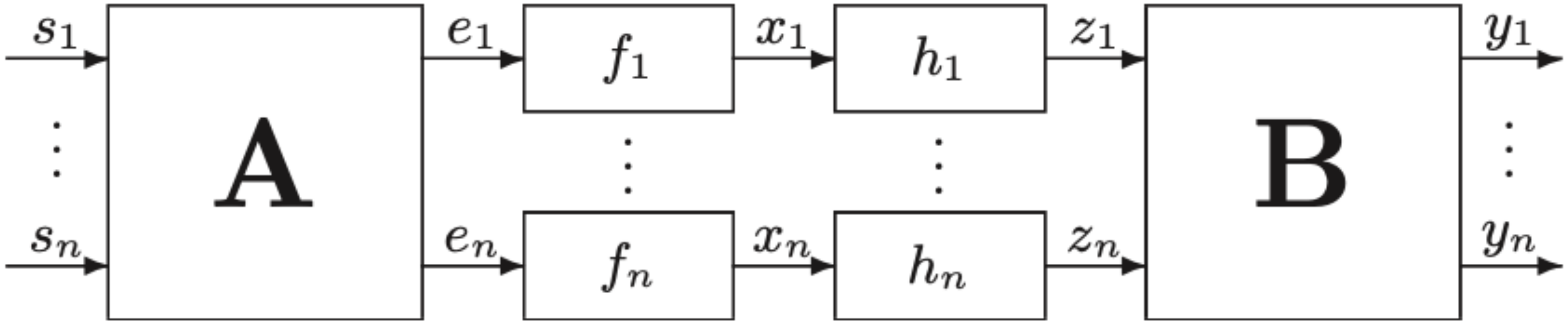
obtained as the composition of a linear mixture with a non-linearity

Example



Courtesy of Jutten et al. 2010

Theoretical results



Mixing process

Unmixing procedure

- h corrects for the PNL
- B is an unmixing matrix as in standard ICA

Can we identify an unmixing structure (h,B) based on independence ?

Theoretical results

An equivalent of the Darmois theorem for PNL *(Achard and Jutten 2005)*:

i. A is invertible and "sufficiently mixing"

i.e. $\forall i, j$ s.t. $a_{ij} \neq 0,$

$\exists k \neq j$ such that $a_{ik} \neq 0$

Or

$\exists l \neq i$ such that $a_{lj} \neq 0$

ii. h is composed of differentiable and invertible functions

iii. S has i.i.d. entries, and at most one is Gaussian

iv. The pdf of the sources are differentiable and their derivatives are continuous

Algorithms for PNL mixture models

$\mathbf{Y} = (B \circ h)(\mathbf{X})$ has independent sources iif

$f \circ h$ are linear and $\mathbf{BA} = \mathbf{\Pi\Delta}$

Sources from PNL are identifiable using standard ICA !

General non-linear mixture models

General mixture models

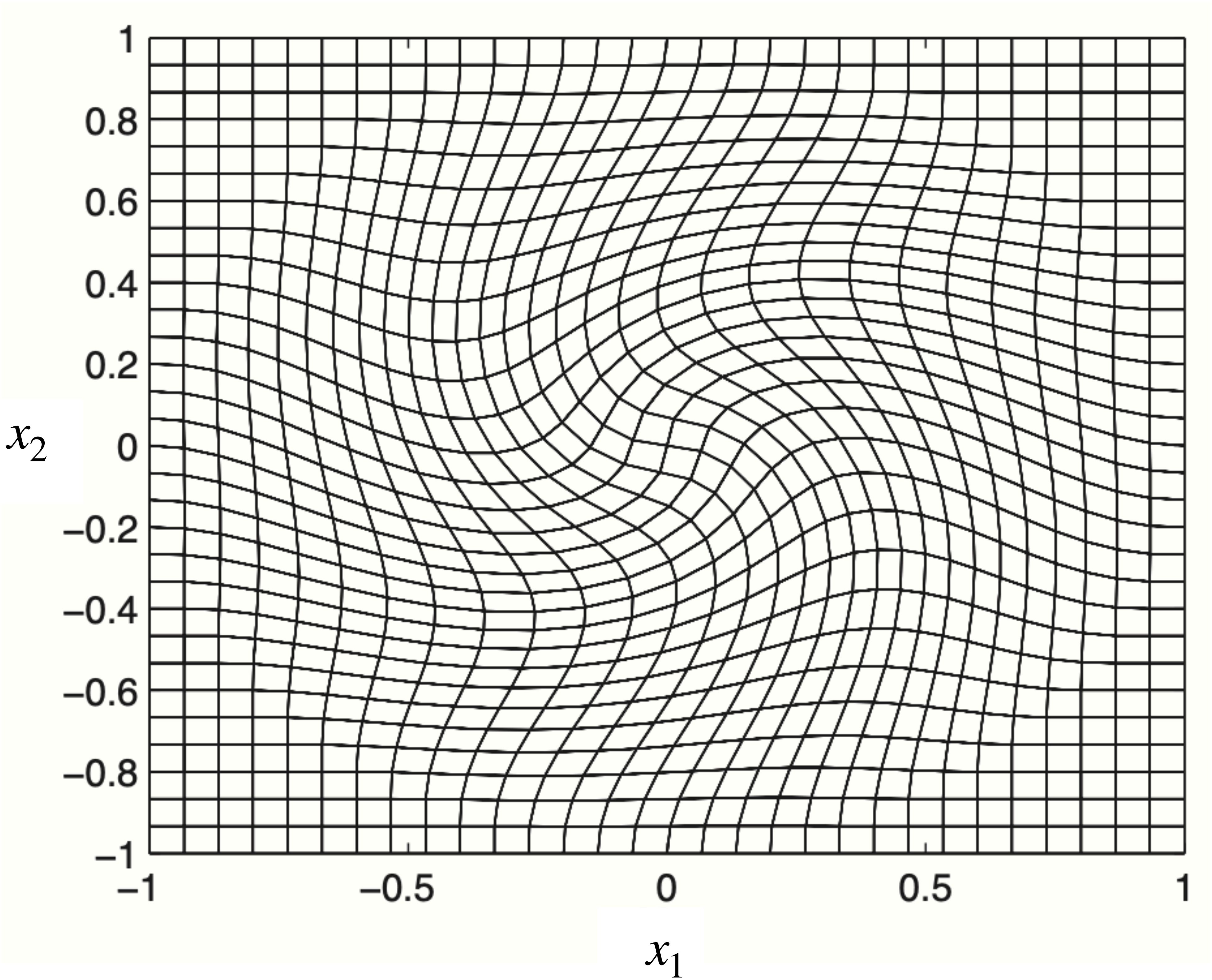
$$\mathbf{X} = f(s_1, \cdots, s_n) + \mathbf{N}$$

General non-linear mixture model

$$\boxed{\mathbf{X}^t = f(s_1^t, \cdots, s_n^t) + \mathbf{N}^t}$$

General instantaneous non-linear mixture model

Context



Sketch of a smooth non-linear mixture

Context

Are there extensions of the Darmois-Linnik theorem for general NL mixtures?

Take I - Kagan, Linnik, Rao (1973!) *and later by koivunen et al. (2002)*

Identifiability limited to non-linearities admitting some mapping $\mathscr{F}$ so that:

$$f(s_1 + s_2) = \mathscr{F}[f(s_1), f(s_2)]$$

$$s_1 + s_2 = f^{-1}\left(\mathscr{F}[f(s_1), f(s_2)]\right)$$

This allows to go back to the standard case

Take II - Hyvarinen et al. (2020)

More general result … see later in this course

ICA for NL mixtures

A quick recall on InfoMAX - minimises the mutual information (MI) of the sources:

$$\min_{\mathbf{B}} \mathscr{D}\left(f_{\mathbf{S}}, \prod_i f_{s_i}\right) \qquad \mathbf{S} = \mathbf{BX}$$

Extension of the InfoMax algorithm to NL mixtures *(Almeida 2004)*

i. The unmixing process is a non-linear mapping h

ii. Estimate h so that h(X) minimises the MI

$$\min_{h} \mathscr{D}\left(f_{\mathbf{S}}, \prod_i f_{s_i}\right) \qquad \mathbf{S} = h\left(\mathbf{X}\right)$$

Extension of the InfoMax algorithm to NL mixtures *(MISEP - Almeida 2004)*

Modelling of the unmixing mapping with a multilayer perceptron (MLP)

$\mathbf{X}$

$$\mathbf{S} = h_\Theta(\mathbf{X})$$

$$u_l = \mathscr{A}(W_l u_l + b_l)$$

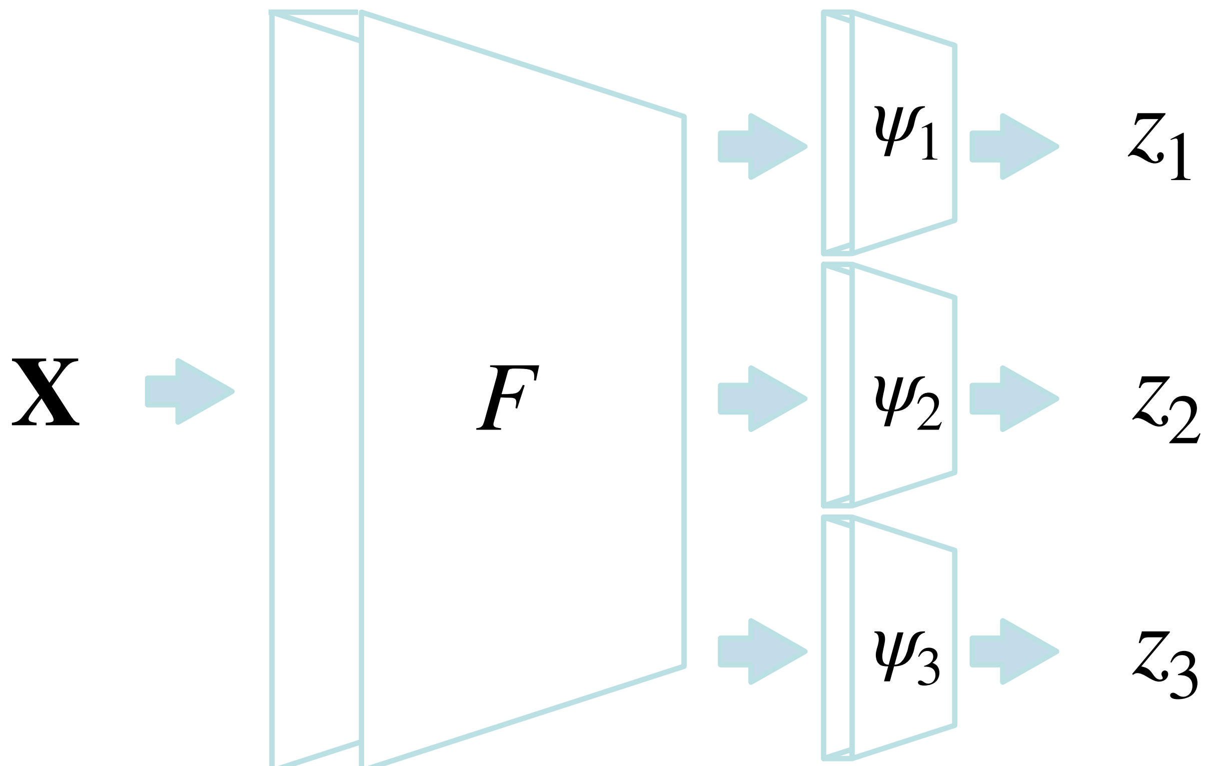
$$\Theta = \{W_1, \cdots, W_L, b_1, \cdots, b_L\}$$

$$\min_\Theta \mathscr{D}\left(f_\mathbf{S}, \prod_i f_{s_i}\right) \qquad \mathbf{S} = h_\Theta(\mathbf{X})$$

MISEP algorithm

- Pre-DL methods, does make use of standard ML coding frameworks

- A simple 2-layers network

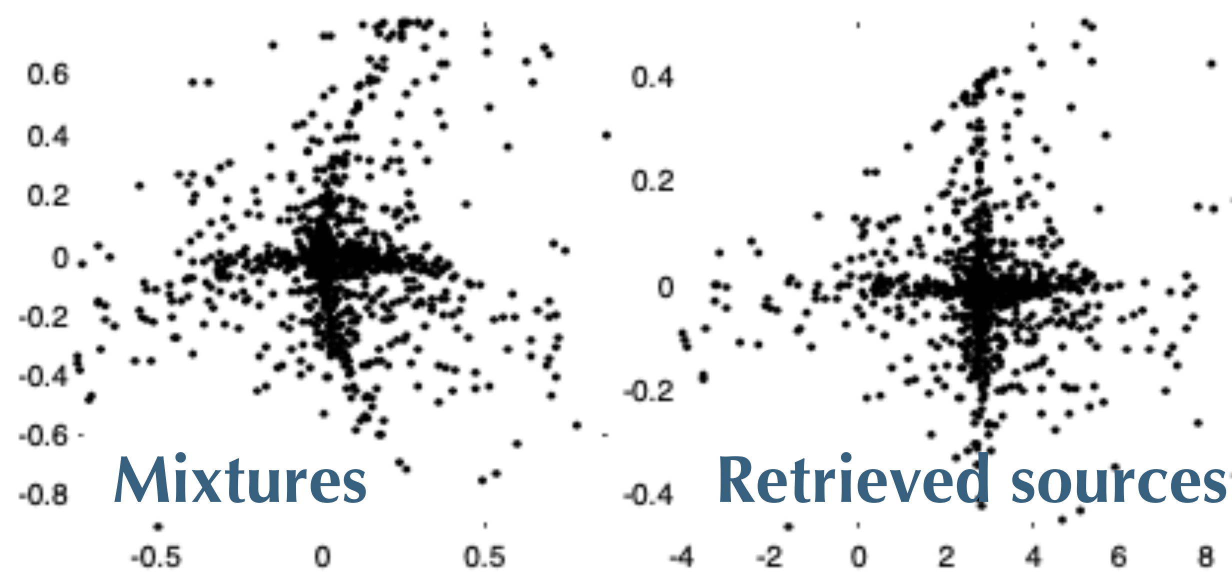


- F is a linear mapping or MLP *(trained)*

- ψ is a non-linear mapping *(trained)*

  e.g. *arctan+linear units*

- Loss: MI of the outputs



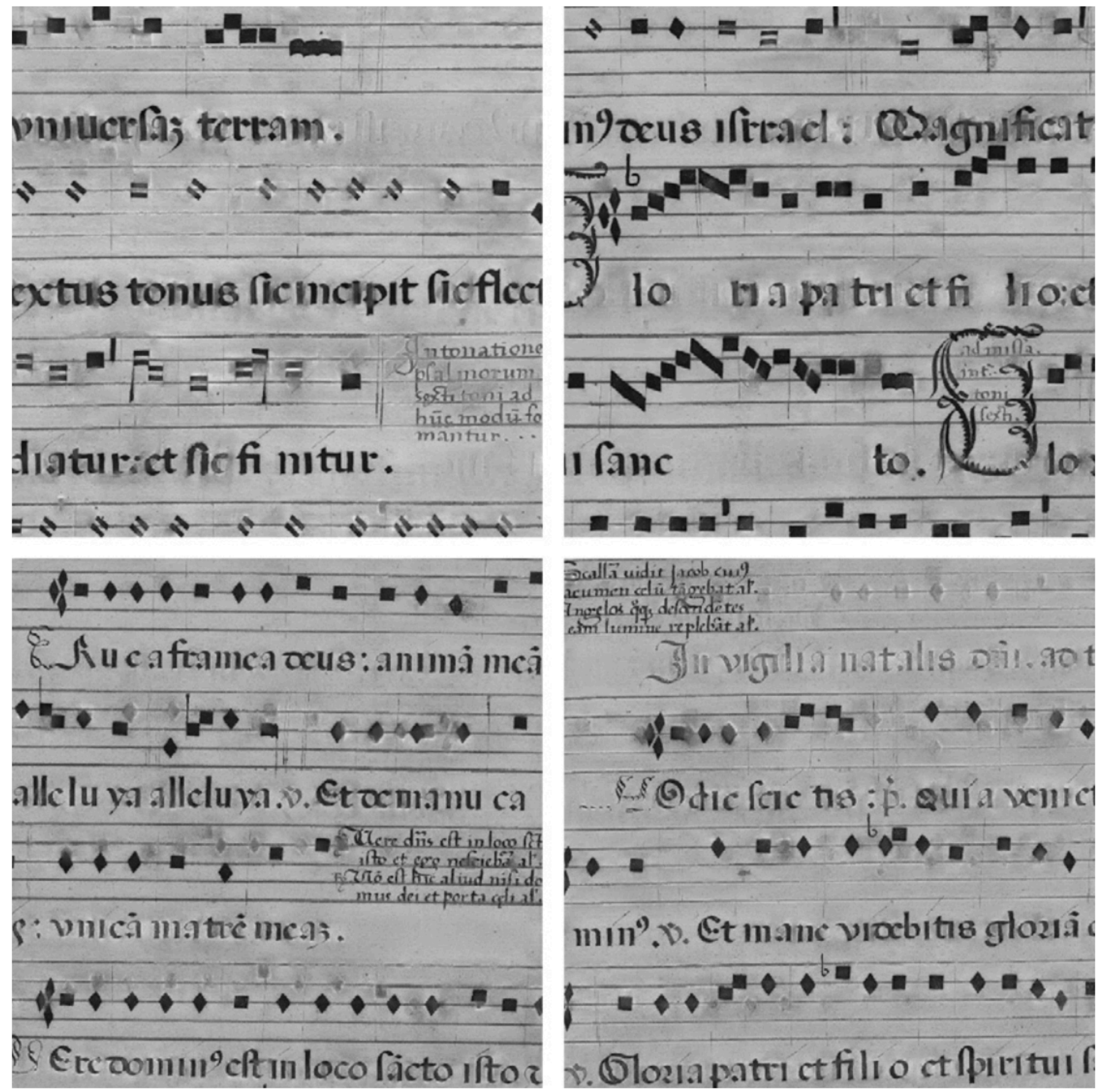
Models for NLMM

Partiture

#1

Measurements

Retrieved sources



Results on show-though data

Encoder/decoder-based approaches

Beyond NL-ICA, from 2004 to 2020:

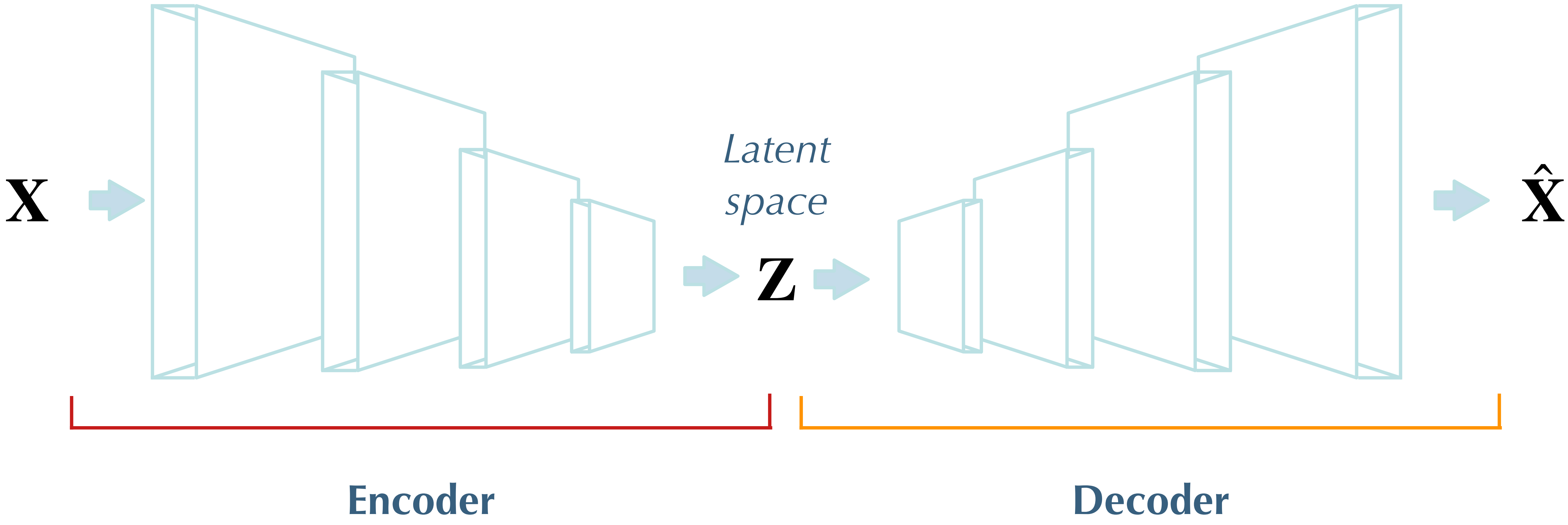
i) **Applicability**: improving the modelling of the sources/non-linearity

ii) **Reconstructing more than an unmixing process ?**

iii) **Identifiability ?** make use of modern-day statistical models

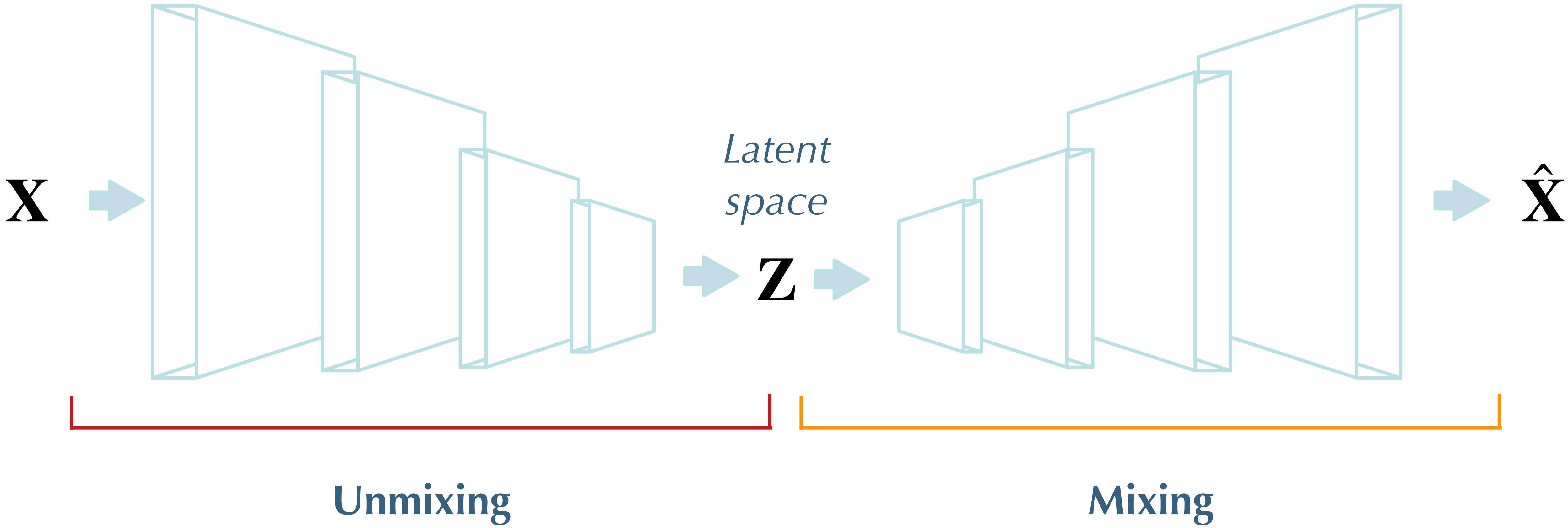
It's time to switch to encoder-decoder (autoencoder) models

Encoder/decoder-based approaches



The Autoencoder (AE) is a very popular model for:

- *(unsupervised)* **representation learning**
- *unsupervised/self-supervised* **learning**
- **manifold learning**
- **Building a non-linear low-dimensional approximation**
- **Generative models**
- **…etc**

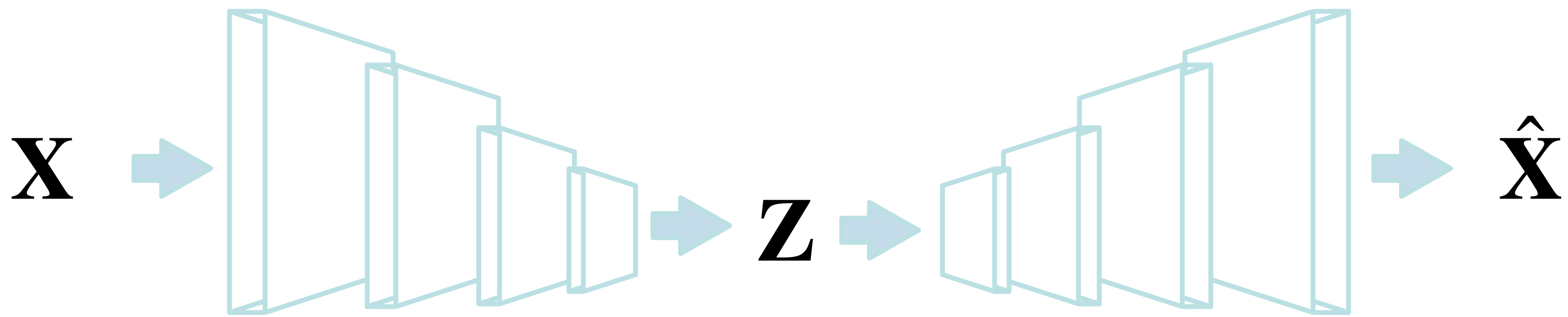


Making use of AE for non-linear ICA is virtually simple:

Learning an unmixing/mixing model that enforces independence in the latent space

But how to do it efficiently ?

ANICA *(Brakel/Bengio 2017)*



Back to the mutual information:

For simplicity, let's pick two sources:

$$I(z_1, z_2) = \int\int p(z_1, z_2) \log \frac{p(z_1, z_2)}{p(z_1)p(z_2)} dz_1 dz_2$$

When p is unknown, how to compute/minimise the mutual information ?

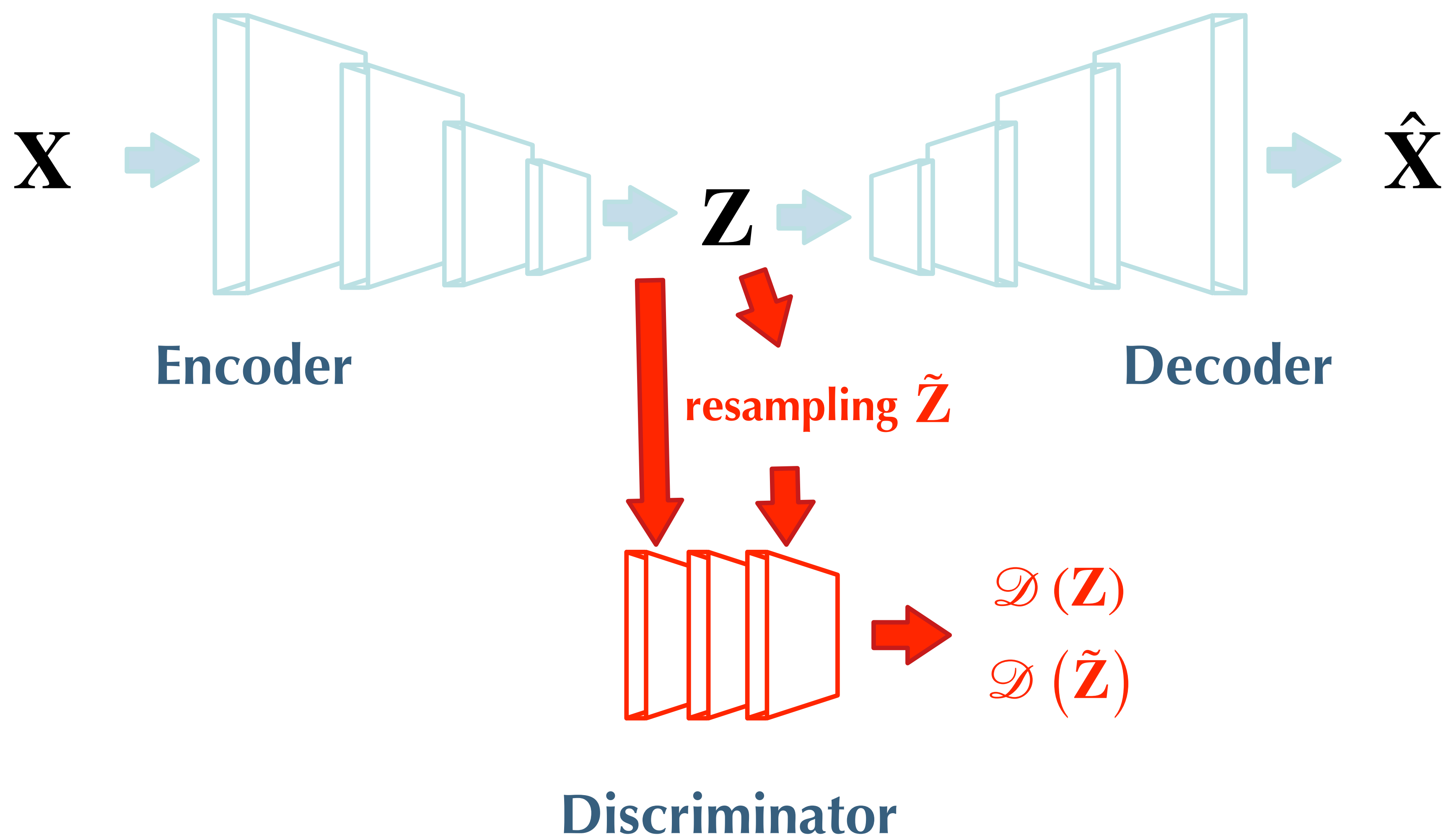
Solution: use a proxy to favour independent latent space variables thanks to
adversarial training

Adversarial regularisation for NL ICA

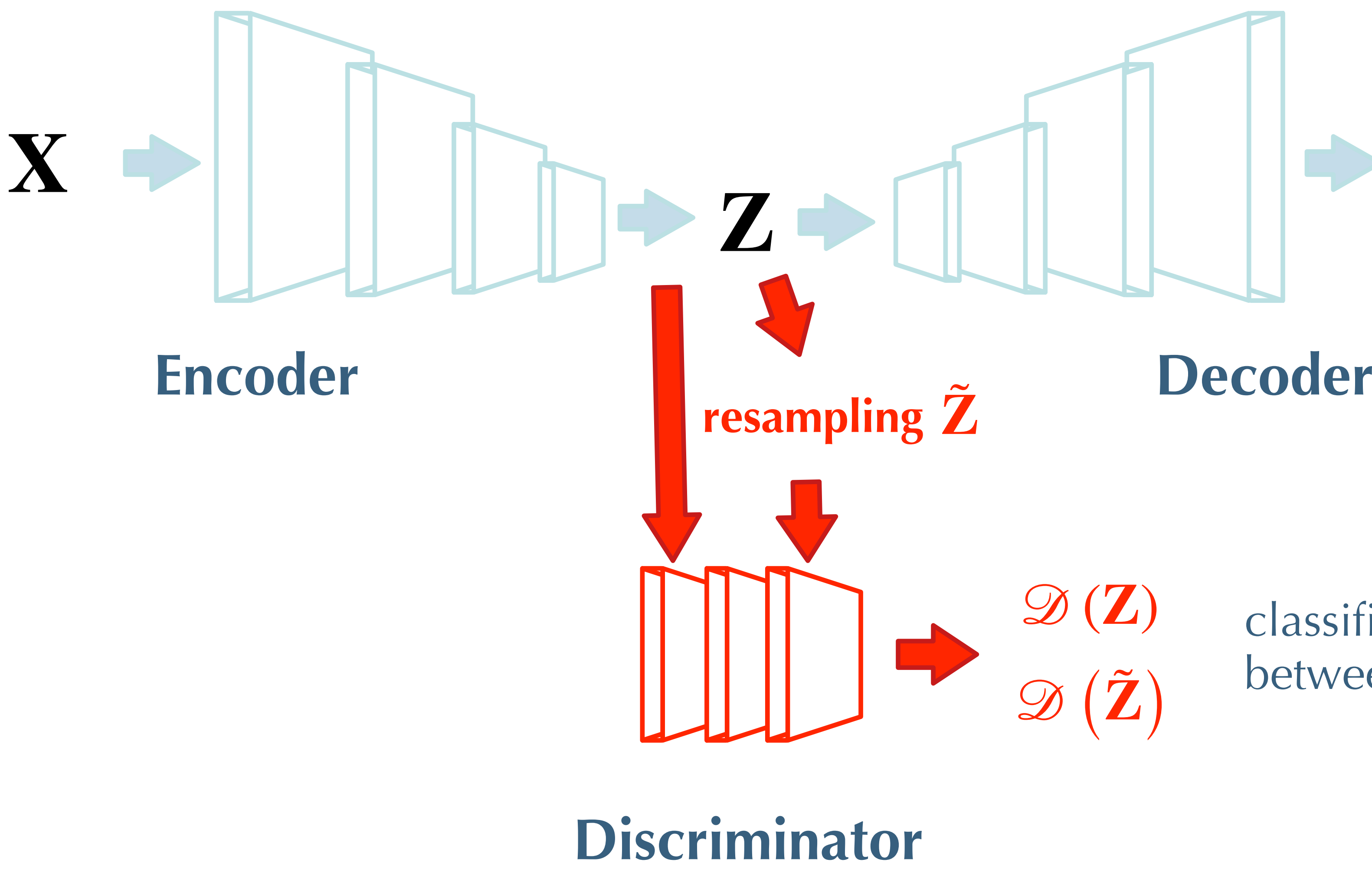
Gist of the approach:

i) Add a regularisation to the latent space variable based on a generator/discriminator model

ii) Generative Adversarial Network *(GAN - Goodfellow 2014)* **allows to model and discriminate dependent/independent sources**



Adversarial regularisation for NL ICA



$\mathscr{D}(\mathbf{Z})$

$\mathscr{D}(\tilde{\mathbf{Z}})$

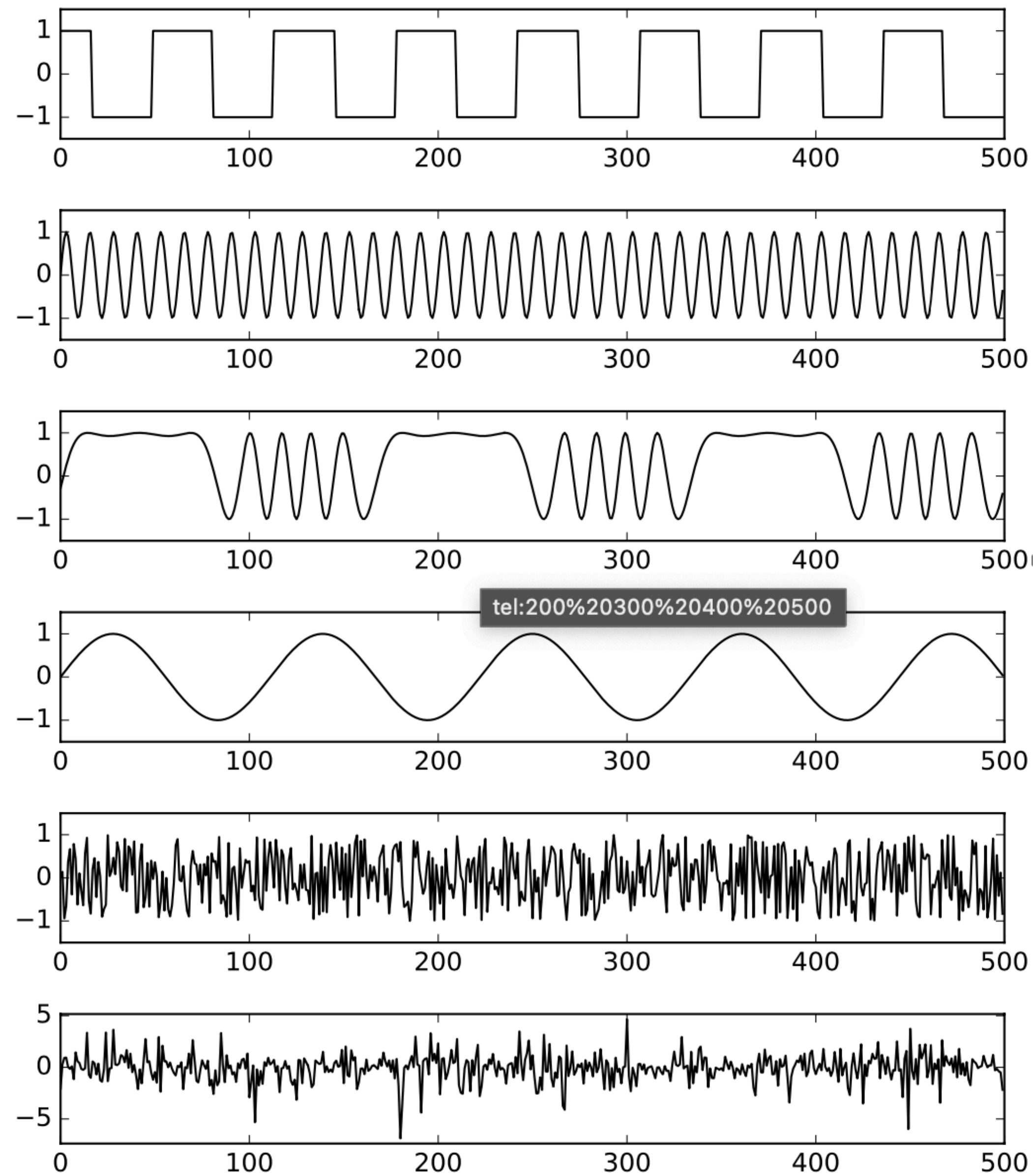
classifier that discriminates between trues/resampled data

Updated loss:

$$\mathscr{L} = \|\mathbf{X} - \hat{\mathbf{X}}\|_2 + \lambda \left[\log \mathscr{D}(\mathbf{Z}) + \log(1 - \mathscr{D}(\mathbf{Z}))\right]$$

Reconstruction error **Discrimination loss**

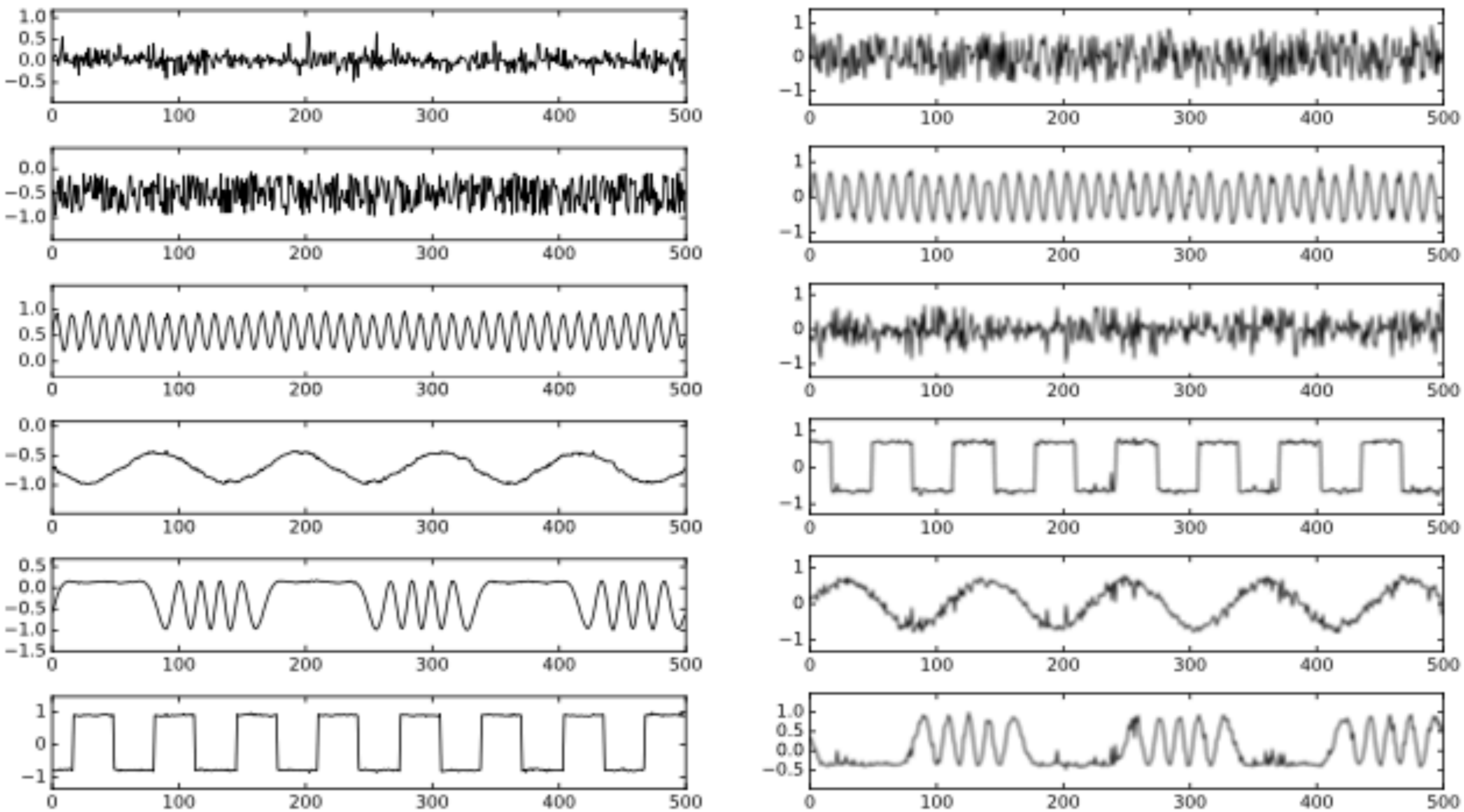
Examples of reconstructions (audio signals) in case of a PNL



Test with NL and PNL mixtures

Adversarial regularisation for NL ICA

Examples of reconstructions (audio signals) in case of a PNL



(a) Anica PNL reconstructions $\rho_{\max} = .997$. (b) Anica MLP reconstructions $\rho_{\max} = .968$.

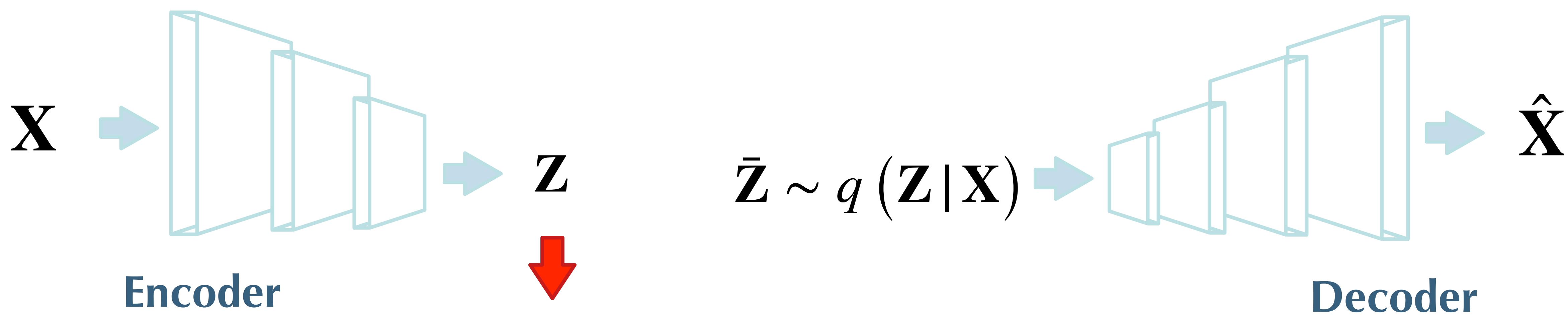
$$X = \tanh(B\ \tanh(AS)) \qquad x_1 = \tanh(x), x_2 = (x + x^3)/2, x = \exp(x)$$

Gist of the approach:

i) better constrain the latent space

ii) statistical model of the latent variables

Solution: make use of a variational autoencoder model



$$p\left(\mathbf{Z}\,|\,\mathbf{X}\right) \simeq q\left(\mathbf{Z}\,|\,\mathbf{X}\right)$$

$$q\left(\mathbf{Z}\,|\,\mathbf{X}\right) \propto \exp((\mathbf{Z}-\mu_{\mathbf{Z}})\Sigma^{-1}(\mathbf{Z}-\mu_{\mathbf{Z}})^{T})$$

Learn the marginal likelihood of the data *(as a generative model)*:

$$\min_{\phi,\psi} \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}\left[-\log p_{\psi}(\mathbf{X}\,|\,\mathbf{Z})\right]$$

ϕ : encoder, ψ : decoder

Reformulation:

$$\log p_{\psi}(\mathbf{X}\,|\,\mathbf{Z}) = \mathscr{D}_{KL}\left(q\left(\mathbf{Z}\,|\,\mathbf{X}\right), p(\mathbf{X})\right) + \mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right)$$

approximation error
of the posterior

β-VAE loss *(Higgins 2017 - Burgess 2018)*:

$$\mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right) = \beta\mathscr{D}_{KL}\left(q\left(\mathbf{Z}\,|\,\mathbf{X}\right), p(\mathbf{X})\right) - \log p_{\psi}(\mathbf{X}\,|\,\mathbf{Z})$$

Factor VAE (Kim, Mnih 2019)

$$\mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right) = \beta \mathscr{D}_{KL}\left(q\left(\mathbf{Z} \,|\, \mathbf{X}\right), p(\mathbf{Z})\right) - \log p_\psi(\mathbf{X} \,|\, \mathbf{Z})$$

$$\mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right) = \beta \mathscr{D}_{KL}\left(q\left(\mathbf{Z} \,|\, \mathbf{X}\right), p(\mathbf{Z})\right) + \boxed{\gamma \mathscr{D}_{KL}\left(q\left(\mathbf{Z}\right), \prod_i q(z_i)\right)} - \log p_\psi(\mathbf{X} \,|\, \mathbf{Z})$$

"Total correlation" :

$$\mathscr{D}_{KL}\left(q\left(\mathbf{Z}\right), \prod_i q(z_i)\right) = \mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{q\left(\mathbf{Z}\right)}{\prod_i q(z_i)}\right]$$

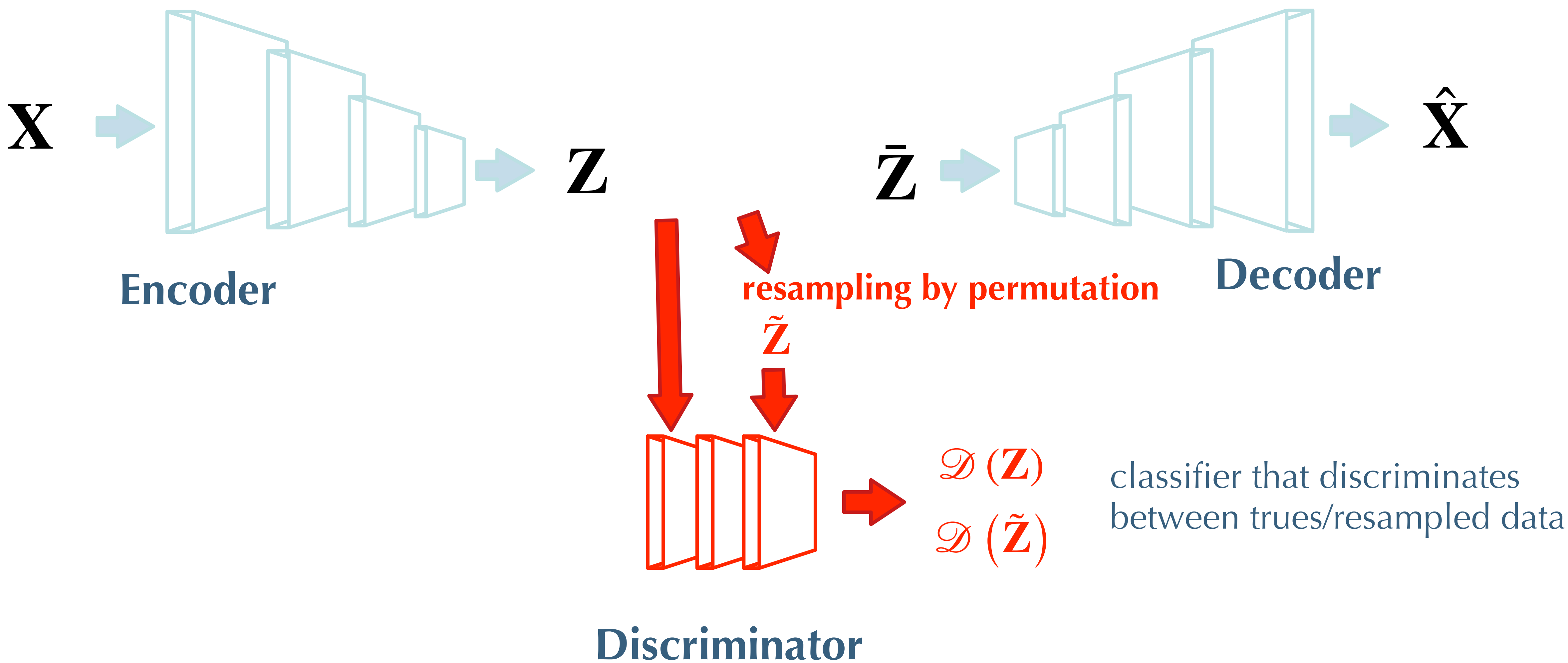
$$\simeq \mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{\mathscr{D}(\mathbf{Z})}{1 - \mathscr{D}(\mathbf{Z})}\right] \qquad \textit{(Sugyama 2012)}$$

$$\mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right) = \beta \mathscr{D}_{KL}\left(q\left(\mathbf{Z} \,|\, \mathbf{X}\right), p(\mathbf{Z})\right) + \boxed{\gamma \frac{\log(\mathscr{D}(\mathbf{Z}))}{\log(1 - \mathscr{D}(\tilde{\mathbf{Z}}))}} - \log p_\psi(\mathbf{X} \,|\, \mathbf{Z})$$

Factor VAE (Kim, Mnih 2019)

$$\mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right) = \beta \mathscr{D}_{KL}\left(q\left(\mathbf{Z} \,|\, \mathbf{X}\right), p(\mathbf{Z})\right) - \log p_{\psi}(\mathbf{X} \,|\, \mathbf{Z})$$

$$\mathscr{L}\left(\phi, \psi; \mathbf{X}, \mathbf{Z}\right) = \beta \mathscr{D}_{KL}\left(q\left(\mathbf{Z} \,|\, \mathbf{X}\right), p(\mathbf{Z})\right) + \gamma \boxed{\frac{\log(\mathscr{D}(\mathbf{Z}))}{\log(1 - \mathscr{D}(\tilde{\mathbf{Z}}))}} - \log p_{\psi}(\mathbf{X} \,|\, \mathbf{Z})$$



Towards identifiability: a unified framework

Some good and bad news *(Khemakhem 2020):*

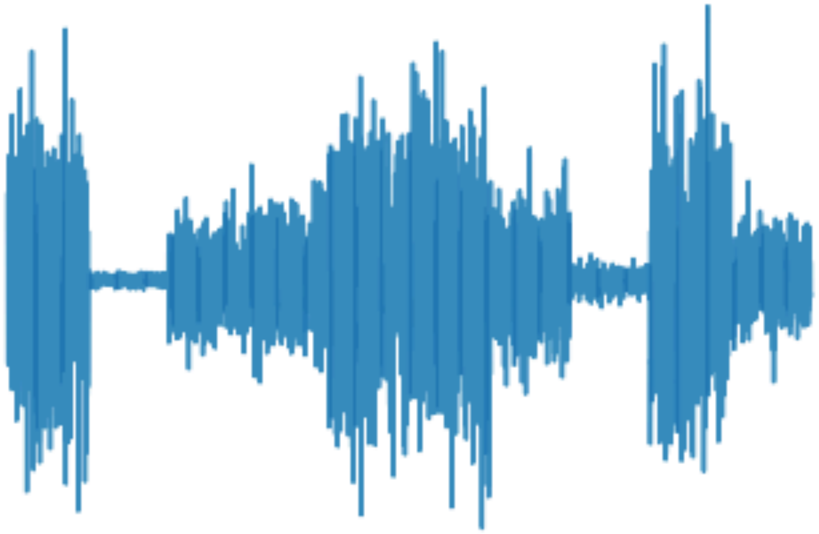
i) we generally don't have identifiability .. unless

ii) prior information is known about the sought-after sources

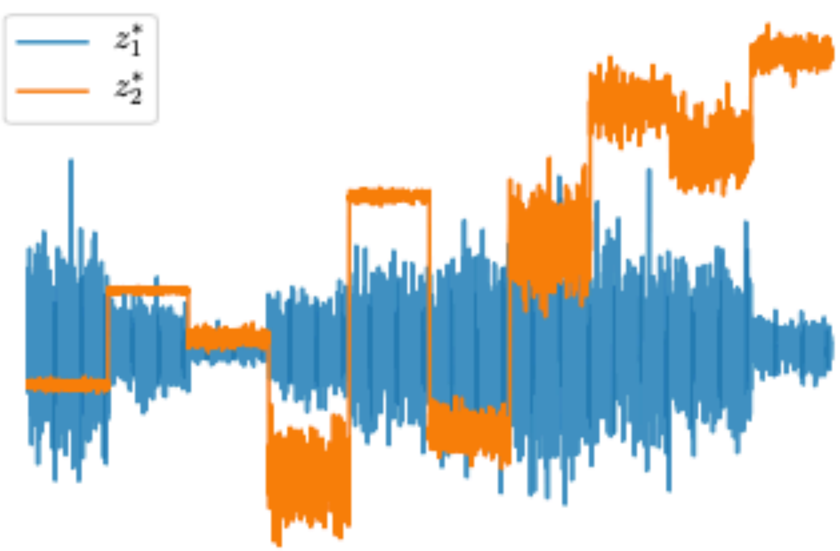
$$q\left(\mathbf{Z}\,|\,\mathbf{X}\right), p(\mathbf{Z}) \rightarrow q\left(\mathbf{Z}\,|\,\mathbf{X}, \mathbf{U}\right), p(\mathbf{Z}\,|\,\mathbf{U})$$

*Where **U** is an auxiliary variable describing the sources*

e.g. time-dependent sources, class labels, etc



(a)



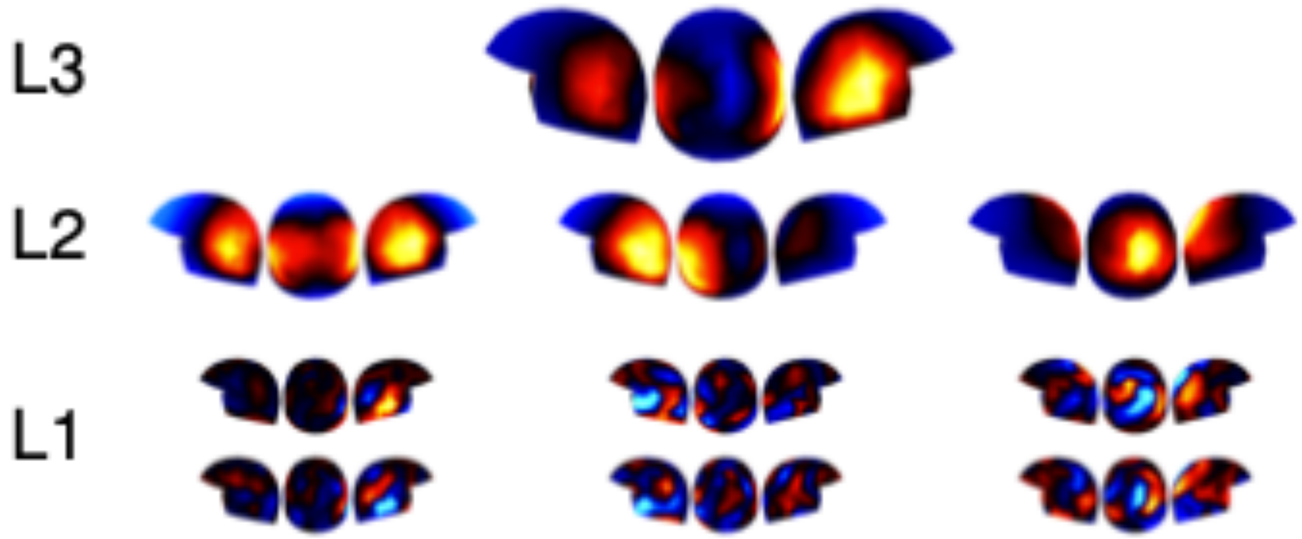
(b)

(Temporary conclusion)

- VAE-based architectures provide a powerful and general framework to tackle NL-ICA

- Identifiability results have quickly emerged during the past 3 years

- Applicability and extensions to tackle real problems are fully open !



MEG data
(Hyvarinen & Morioka, 2016)