ICML | 2019

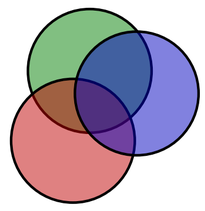
Sparse Multi-Channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data

Luigi Antelmi[1], Nicholas Ayache[1], Philippe Robert[2,3], Marco Lorenzi[1]

1) Université Côte d'Azur, INRIA, Epione Project-Team, France. 2) Université Côte d'Azur, CoBTeK, France. 3) Centre Mémoire, CHU of Nice, France.

The Context

High dimensional data sources are pervasive, especially in complex healthcare applications where data come from multiple channels, such as clinical scores, medical imaging, ...
We propose a method leading to an interpretable joint representation of complex multi-channel data.



Challenges
1) Joint description of multi-channel data.
2) Interpretable relationship between channels.
3) Ability to cope with missing data in training and testing.

Idea: 1) generalize the Variational AutoEncoder [Kingma et al., 2014; Rezende et al., 2014] to joint modeling independently encoded multi-channel data, by constraining the encoded latent space distributions to match a common target distribution. 2) Enhance interpretability and parsimony with sparse latent representation.

The Generative Model

In the case of C **channels** we hypothesize the following parametrized sampling process:

$$\text{latent variable: } \mathbf{z} \sim p(\mathbf{z})$$
$$\text{observed channels: } \mathbf{x}_c \sim p(\mathbf{x}_c|\mathbf{z}, \theta_c) \text{ for } c = 1..C.$$

As depicted in Fig.1, to **infer** the latent variable from the observed channels, we variationally approximate the posterior distribution $p(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_C, \theta_1 \ldots, \theta_C)$ with C independently encoded latent representations $q(\mathbf{z}|\mathbf{x}_c, \phi_c)$.

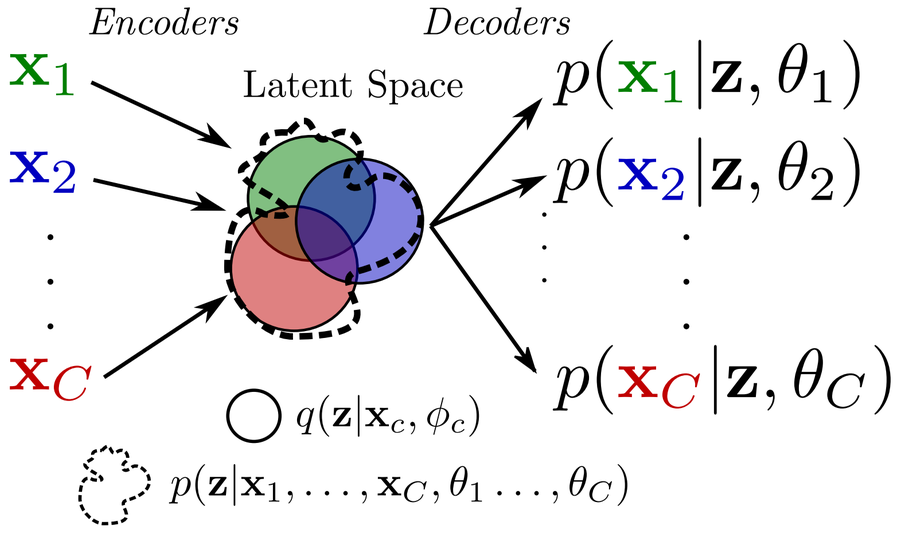


Fig.1: A set of approximate density functions ◯ ,one for each channel, are optimized to be, on average, as close as possible to the exact uncomputable posterior .

The Evidence Lower Bound

With the variational approach is possible to construct C approximate distributions, one for each channel, as close as possible, on average, to the exact posterior one, by solving the following optimization problem:

$$\underset{q \in \mathcal{Q}}{arg\,min}\ \mathbb{E}_c \left[\mathcal{D}_{\mathrm{KL}} \left(q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) || \, p\left(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_C, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C\right) \right) \right]$$

that is equivalent to maximize the data **evidence** through the **lower bound**:

$$\underbrace{\ln p\left(\mathbf{x}_1, \ldots, \mathbf{x}_C\right)}_{\text{Evidence}} \geq$$
$$\underbrace{\frac{1}{C} \sum_{c=1}^{C} \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c, \phi_c)} \left[\sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\theta}_i) \right] - \mathcal{D}_{\mathrm{KL}} \left(q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right) || \, p\left(\mathbf{z}\right) \right)}_{\text{Lower Bound}}$$

Parameters $\boldsymbol{\phi}_c, \boldsymbol{\theta}_i \ \ (c, i = 1 \ldots C)$ are optimized to maximize the Lower Bound. The latent variable $\mathbf{z}$ is inferred from each channel $\mathbf{x}_c$ through $q\left(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c\right)$ and contributes to the joint decoding of all the channels $\mathbf{x}_i$. The dimension of the latent space, dim($\mathbf{z}$), is not generally known *a priori*, but it can be selected via *Variational Dropout* [Molchanov et al.; 2017] (see next panel).

Inducing sparsity *via* Variational Dropout

Latent Parameterization

Standard	Sparse
$q(z_i\|\mathbf{x}) = \mathcal{N}(\mu_i; \sigma_i^2\|\mathbf{x})$	$q(z_i\|\mathbf{x}) = \mathcal{N}(\mu_i; \alpha\mu_i^2\|\mathbf{x})$
$p(z_i) = \mathcal{N}(0; 1)$	$p(z_i) \propto 1/\|z_i\|$

Why it works? $\lim_{\mu_i \to 0} \mathcal{N}(\mu_i; \alpha\mu_i^2|\mathbf{x}) = \delta(0)$

Relationship between α and the probability of pruning the *i*-th dimension:

$$\alpha_i = \frac{p_i}{1 - p_i}$$

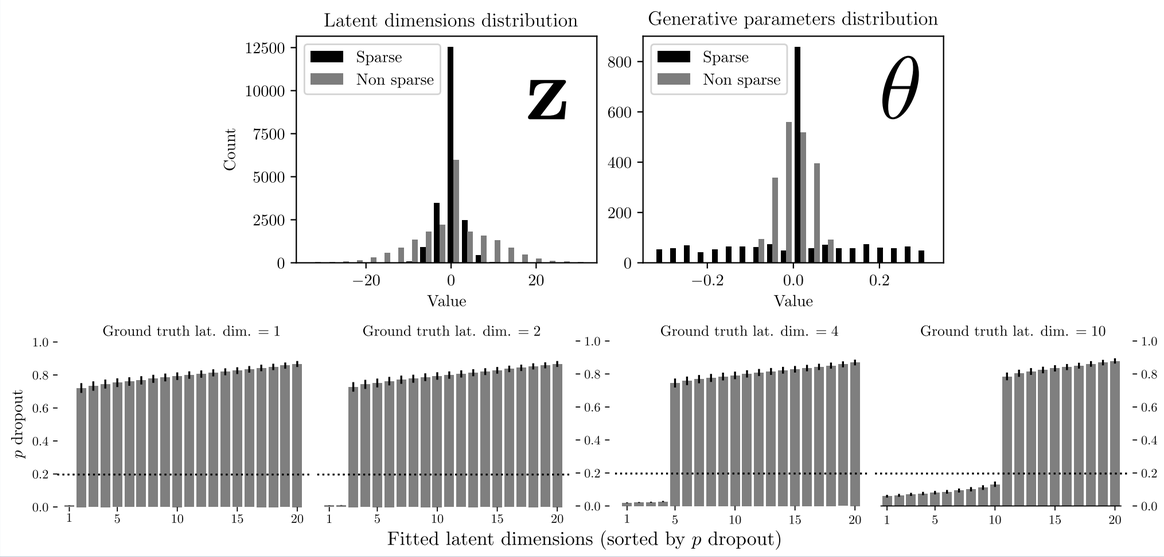


Fig.2: Sparsity in action in synthetic experiments. (Top) Latent dimensions $\mathbf{z}$ and generative parameters θ going to zero within the sparse framework. (Bottom) Selection of the latent dimensions by sorting them for their dropout probability.

Inference and Prediction on Medical Data

Experiment on real data: 504 subjects, four channels:

1) Clinical (age, mmse, adas11, cdr-sb, faq, pteducat)
2) MRI (gray matter - AAL atlas parcellation)
3) FDG-PET (Glucose uptake - AAL atlas parcellation)
4) AV45-PET (Amyloid uptake - AAL atlas parcellation)

Sparse model: 5 dims retained through dropout;
linear parameterization.

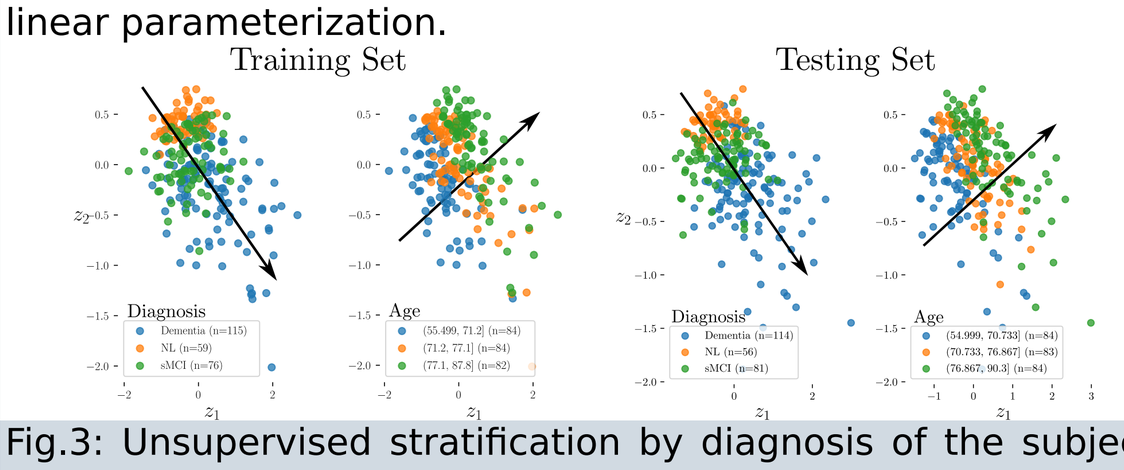


Fig.3: Unsupervised stratification by diagnosis of the subjects in the latent space (first 2 dims shown). Stratification by age occurs on a roughly orthogonal direction with respect to the diagnosis one.

At test time, prediction of unseen channels $\{\hat{\mathbf{x}}_i\}$ can be inferred from the available ones $\{\tilde{\mathbf{x}}_j\}$ through the formula:

$$\hat{\mathbf{x}}_i = \mathbb{E}_j \left[\mathbb{E}_{q(\mathbf{z}|\tilde{\mathbf{x}}_j)} \left[p\left(\mathbf{x}_i|\mathbf{z}\right) \right] \right]$$

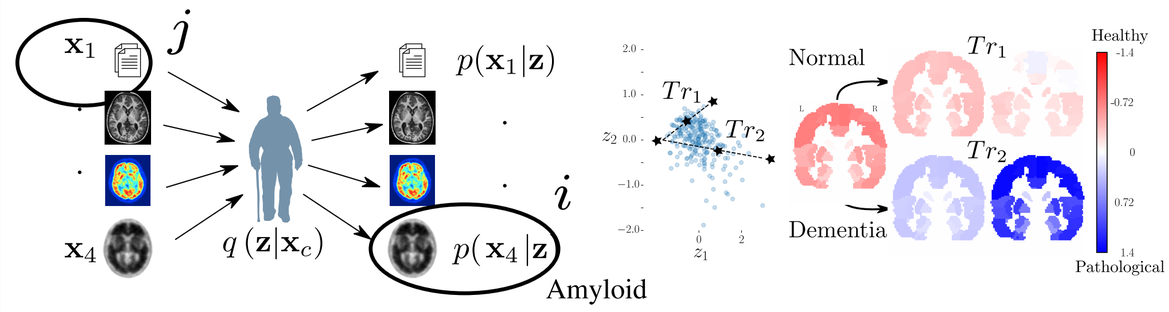


Fig.4: Generation of imaging data from trajectories in the latent space. Trajectory 1 (Tr1) follows an ageing path centered on the healthy subject group. Trajectory 2 (Tr2) follows a path where ageing is entangled with pathological variability.

This work has been supported by:
- the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01;
- the grant AAP Santé 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.