Multi-Channel Stochastic Variational Inference for the Joint Analysis of Heterogeneous Biomedical Data in Alzheimer's Disease

Luigi Antelmi, Nicholas Ayache, Philippe Robert, Marco Lorenzi

Multi-Channel Stochastic Variational Inference for the Joint Analysis of Heterogeneous Biomedical Data in Alzheimer's Disease

Luigi Antelmi[1], Nicholas Ayache[1], Philippe Robert[2,3], and Marco Lorenzi[1] for the Alzheimer's Disease Neuroimaging Initiative *

[1] University of Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France
[2] University of Côte d'Azur, CoBTeK, France
[3] Centre Memoire, CHU of Nice, France

Abstract. The joint analysis of biomedical data in Alzheimer's Disease (AD) is important for better clinical diagnosis and to understand the relationship between biomarkers. However, jointly accounting for heterogeneous measures poses important challenges related to the modeling of heterogeneity and to the interpretability of the results. These issues are here addressed by proposing a novel multi-channel stochastic generative model. We assume that a latent variable generates the data observed through different channels (*e.g.,* clinical scores, imaging) and we describe an efficient way to estimate jointly the distribution of the latent variable and the data generative process. Experiments on synthetic data show that the multi-channel formulation allows superior data reconstruction as opposed to the single channel one. Moreover, the derived lower bound of the model evidence represents a promising model selection criterion. Experiments on AD data show that the model parameters can be used for unsupervised patient stratification and for the joint interpretation of the heterogeneous observations. Because of its general and flexible formulation, we believe that the proposed method can find various applications as a general data fusion technique.

1 Introduction

Physicians investigate their patients' status through various sources of information that in this work we call *channels*. For Alzheimer's Disease (AD), for example, the anamnestic questionnaire, genetic tests and various imaging modalities are channels providing specific, complementary, and sometimes overlapping views on the patient's state [3, 7].

Tackling a complex disease like AD requires to establish a link between heterogeneous data channels. However, simple univariate correlation analyses are limited in modeling power, and are prone to false positives when the data

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. A detailed list of funding actors can be found in the *Acknowledgments* section of the *Supplementary Material* .

dimension is high. To overcome the limitations of mass-univariate analysis, more advanced methods, such as Partial Least Squares (PLS), Reduced Rank Regression (RRR), or Canonical correlation analysis (CCA) [5] have successfully been applied in biomedical research [12], along with multi-channel [8, 13] and non-linear [1, 6] variants.

A common drawback of standard multivariate methods is that they are not generative. Indeed, their formulation consists in projecting the observations in a latent lower dimensional space in which they exhibit certain desired characteristics like maximum correlation (CCA), maximum covariance (PLS), minimum regression error (RRR); however these methods are limited in providing information on how this latent representation is expressed in the observations [4]. Moreover, techniques for model comparison should be applied to select the best number of dimensions for the latent representation and avoid overfitting. While cross-validation is the standard model validation procedure, this requires holding-out data from the original dataset, and thus leading to data loss at the training stage.

We need generative models that can actually describe the direct influence of the latent space on the observations, and model selection techniques leveraging solely on training data. *Bayesian-CCA* [11] actually goes in this direction: it is a generative formulation of the CCA defined on a latent variable that captures the shared variation between data channels. Moreover, the Bayesian formulation allows the use of probabilistic model comparison, without recurring to cross-validation. However, *Bayesian-CCA* may not scale well to large dimensions and several channels.

In this work we aim at addressing the current methodological limitations in multi-channel analysis. By leveraging on the recent developments on performing efficient approximate *Variational Inference* in Bayesian modeling in an efficient way, we propose a novel multi-channel stochastic generative model for the joint analysis of multi-channel heterogeneous data. Our hypothesis is that a latent variable $\mathbf{z}$ generates the heterogeneous data $\mathbf{x}_1, \ldots, \mathbf{x}_C$ observed through different channels C. In this work we propose an efficient way to estimate jointly the latent variable distribution and the data likelihood $p(\mathbf{x}_1, \ldots, \mathbf{x}_C|\mathbf{z})$, and we also investigate a mean for Bayesian model selection. Our work generalizes the *Variational Autoencoder* [10] and the *Bayesian-CCA*, making possible to jointly model multiple channels simultaneously and efficiently.

The next sections of this paper are organized as follows. In Section 2 we present the derivation of the multi-channel variational model and we describe a possible implementation with Gaussian distributions parametrized by linear functions. In Section 3 we apply our method on a synthetic dataset, as well as on a real multi-channel Alzheimer's disease dataset, to test the descriptive and predictive properties of the model. In the last section we provide our discussions and conclusions. Further experimental tests are provided in the *Supplementary Material* [1].

[1] https://hal.inria.fr/hal-01844733

2 Method

2.1 Multi-Channel Variational Inference

Let $\mathbf{x} = \{\mathbf{x}_c\}_{c=1}^{C}$ be a single observation of a set of C channels, where each $\mathbf{x}_c \in \mathbb{R}^{d_c}$ is a d_c-dimensional vector. Also, let $\mathbf{z} \in \mathbb{R}^l$ denote the l-dimensional latent variable commonly shared by each $\mathbf{x}_c$. We propose the following generative process:

$$\begin{aligned} &\mathbf{z} \sim p(\mathbf{z}) \\ &\mathbf{x}_c \sim p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c) \qquad \text{for} \quad c \quad \text{in} \quad 1 \dots C \end{aligned} \tag{1}$$

where $p(\mathbf{z})$ is a prior distribution for the latent variable and $p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c)$ is a likelihood distribution for the observations conditioned on the latent variable. We assume that the likelihood functions belong to a distribution family $\mathcal{P}$ parametrized by $\boldsymbol{\theta}_c$. When the distributions are Gaussians parametrized by linear transformations, the model is equivalent to the *Bayesian-CCA* (*cf.* [11], Eq. 3). In the scenario depicted so far, solving the inference problem allows the discovery of the common latent space from which the observed data in each channel is generated. The solution to the inference problem is given by deriving the posterior $p(\mathbf{z}|\mathbf{x}_1, \dots, \mathbf{x}_C, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_C)$, that is not always computable analytically. In this case, *Variational Inference* [2] can be applied to compute an approximate posterior. In our setting, variational inference is carried out by introducing probability density functions $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ that are on average as close as possible to the true posterior in terms of Kullback-Leibler divergence:

$$\underset{q \in \mathcal{Q}}{arg\,min}\ \mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||\, p(\mathbf{z}|\mathbf{x}_1, \dots, \mathbf{x}_C, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_C)\big) \right] \tag{2}$$

where the approximate posteriors $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ belong to a distribution family $\mathcal{Q}$ parametrized by $\boldsymbol{\phi}_c$, and represent the view on the latent space that can be inferred from each channel $\mathbf{x}_c$. Practically, solving the objective in Eq. (2) allows to use on average every $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ to approximate the true posterior distribution. It can be shown that the maximization of the model evidence $p(\mathbf{x}_1, \dots, \mathbf{x}_C)$ is equivalent to the optimization of the evidence lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) &= \frac{1}{C} \sum_{c=1}^{C} \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)} \left[\sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z}, \boldsymbol{\theta}_c) \right]}_{\text{cross-reconstruction of all } \mathbf{x}_i \text{ from } \mathbf{x}_c} - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||\, p(\mathbf{z})\big) \\ &= \ln \underbrace{p(\mathbf{x}_1, \dots, \mathbf{x}_C)}_{\text{Evidence}} - \underbrace{\mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) \,||\, p(\mathbf{z}|\mathbf{x}_1, \dots, \mathbf{x}_C, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_C)\big) \right]}_{\geq 0} \end{aligned} \tag{3}$$

It can be shown that maximizing $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$ is equivalent to solving the objective in Eq. (2) (*cf. Sup. Mat.*). Moreover, being the lower bound linked to the data evidence up to a positive constant, Eq. (3) allows to test $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$ as a surrogate measure of $p(\mathbf{x}_1, \dots, \mathbf{x}_C)$ for Bayesian model selection. This formulation is valid for any distribution family $\mathcal{P}$ and $\mathcal{Q}$, and the complete derivation of Eq. (3) is in

the *Sup. Mat.*

Comparison with variational autoencoder (VAE). Our model extends the VAE [10]: the novelty is in the cross-reconstruction term labeled in Eq. (3). In case $C = 1$ the model collapses to a VAE. In the case $C > 1$ the cross-term forces each channel to the joint decoding of the other channels. For this reason, our model is different from a stack of VAEs. The dependence between encoding and decoding across channels stems from the joint approximation of the posteriors (Formula (2)).

Optimization of the lower bound. The optimization starts with a random initialization of the generative parameters $\boldsymbol{\theta}$ and the variational parameters $\boldsymbol{\phi}$. The expectation in the first row of Eq. (3) can be computed by sampling from the variational distributions $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c)$ and, when the prior and the variational distributions are Gaussians, the Kullback-Leibler term can be computed analytically (*cf.* [10], appendix 2.A). The maximization of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ is efficiently carried out through minibatch stochastic gradient descent implemented with the backpropagation algorithm. For each parameter, adaptive learning rates are computed with *Adam* [9].

2.2 Gaussian linear case

Model (1) is completely general and can account for complex non-linear relationships modeled, for example, through deep neural networks. However, for simplicity of interpretation, and validation purposes, in the next experimental section we will restrict our multi-channel variational framework to the *Gaussian Linear Model*. This is a special case, analogous to Bayesian-CCA, where the members of the generative family $\mathcal{P}$ and variational family $\mathcal{Q}$ are Gaussians parametrized by linear transformations. The parameters of these transformations are thus optimized by maximizing lower bound. We define the members of the generative family $\mathcal{P}$ as Gaussians whose first moments are linear transformations of the latent variable $\mathbf{z}$, and the second moments are parametrized by a diagonal covariance matrix, such that $p(\mathbf{x}_c|\mathbf{z}, \boldsymbol{\theta}_c) = \mathcal{N}\left(\mathbf{x}_c|\mathbf{G}_c^{(\mu)}\mathbf{z}, diag(\mathbf{g}_c^{(\sigma)})\right)$, where $\mathbf{G}_c^{(\mu)} \in \mathbb{R}^{d_c \times l}$ and $\mathbf{g}_c^{(\sigma)} \in \mathbb{R}^{d_c}$. The elements of $\boldsymbol{\theta}_c = \{\mathbf{G}_c^{(\mu)}, \mathbf{g}_c^{(\sigma)}\}$ are the generative parameters to be optimized for every channel. We also define the members of variational family $\mathcal{Q}$ to be Gaussians whose moments are linear transformation of the observations, such that $q(\mathbf{z}|\mathbf{x}_c, \boldsymbol{\phi}_c) = \mathcal{N}\left(\mathbf{x}_c|\mathbf{V}_c^{(\mu)}\mathbf{x}_c, diag(\mathbf{V}_c^{(\sigma)}\mathbf{x}_c)\right)$ where $\mathbf{V}_c^{(\mu)} \in \mathbb{R}^{l \times d_c}$ and $\mathbf{V}_c^{(\sigma)} \in \mathbb{R}^{l \times d_c}$. The elements of $\boldsymbol{\phi}_c = \{\mathbf{V}_c^{(\mu)}, \mathbf{V}_c^{(\sigma)}\}$ are the variational parameters to be optimized for every channel.

3 Experiments

In this section we illustrate the performance of the method extensively tested on a large scale synthetic dataset, and we provide a real case example by jointly analyzing multimodal brain imaging and clinical scores in AD data. Further experimental tests are provided in *Sup. Mat.*

3.1 Experiments on Linearly generated synthetic datasets

Data generation procedure. Datasets $\mathbf{x} = \{\mathbf{x}_c\}$ with $c = 1 \dots C$ channels where created as $\mathbf{x}_c = \mathbf{G}_c\mathbf{z} + snr^{-1/2}\boldsymbol{\epsilon}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}; \mathbf{I}_l)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}; \mathbf{I}_{d_c})$. snr is the signal-to-noise ratio and $\mathbf{G}_c$ is the linear generative law initialized as $\mathbf{G}_c = diag\left(\mathbf{R}_c\mathbf{R}_c^T\right)^{-1/2}\mathbf{R}_c$ where, for every channel c, $\mathbf{R}_c \in \mathbb{R}^{d_c \times l}$ is a random matrix with l orthonormal columns (*i.e.,* $\mathbf{R}_c^T\mathbf{R}_c = \mathbf{I}_l$). It's easy to demonstrate that the diagonal elements of the covariance matrix of $\mathbf{x}_c$ are inversely proportional to snr, *i.e., diag* $\left(\mathbb{E}\left[\mathbf{x}_c\mathbf{x}_c^T\right]\right) = (1 + snr^{-1})\mathbf{I}_{d_c}$. Scenarios where generated by varying one-at-a-time the dataset attributes (*e.g.,* noise level, number of observations, ...), as listed in *Sup. Mat.* We fitted several instances of the model specified in Section 2.2, changing each time the number of fitted latent dimensions, for a total of 40 000 experiments.

Results. At convergence, the loss function (negative lower bound) has a minimum when the number of fitted latent dimension corresponds to the number of the latent dimensions used to generate the data, as depicted in Fig. 1a. When increasing the number of fitted latent dimensions, a sudden decrease of the loss (elbow effect) is indicative that the true number of latent dimensions has been found. In the *Sup. Mat.* we show also that the elbow effect becomes more pronounced with increasing the number of data channels. Ambiguity in identifying the elbow, instead, may rise for high-dimensional data channels. In these cases, increasing the sample size or the data quality in terms of snr can make the elbow point more noticeable.

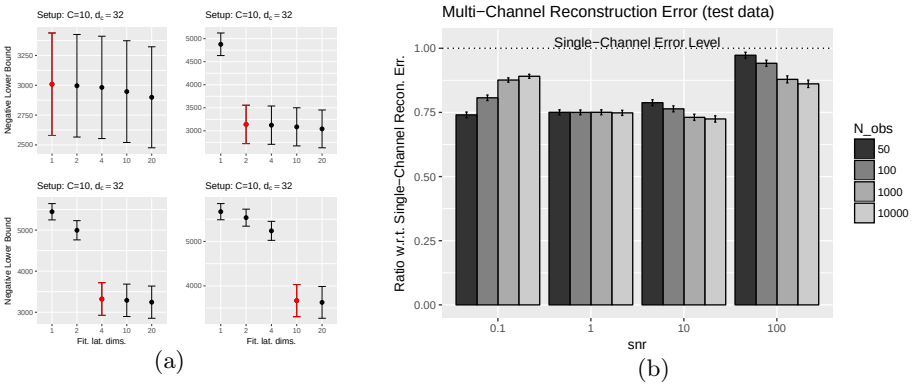


Fig. 1: (a) Negative lower bound (NLB) on the synthetic training set computed at convergence for all the scenarios. Each bar shows mean ± s.e. of $N = 80$ total experiments as a function of the number of fitted latent dimensions. Red bars represents experiments where the number of true and fitted latent dimensions coincide. (b) Ratio between Multi- *vs* Single-Channel reconstruction error computed as mean squared error from the ground truth test data.

Concerning the estimation of the ground truth parameters and data reconstruction, we observed that the performance of the model increases with higher snr, sample size, and number of channels (*Sup. Mat.*); moreover we notice that the error made in ground truth data recovery with multi-channel information is systematically lower than the one obtained with a single-channel decoder (Fig. 1b).

3.2 Application to clinical and medical imaging data in AD

Data preparation. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.

We fit our model with linear parameters to clinical imaging channels acquired on 504 subjects. The clinical channel is composed of six continuous variables generally recorded in memory clinics (age, mini-mental state examination, adas-cog, cdr, faq, scholarity); the three imaging channels are structural MRI (gray matter only), functional FDG-PET, and Amyloid-PET, each of them composed by continuous measures averaged over 90 brain regions mapped in the AAL atlas [14]. Raw data from the imaging channels where coregistered in a common geometric space, and visual quality check was performed to exclude registration errors. Data was centered and standardized across dimensions. Model selection was carried out by comparing the lower bound for several fitted latent dimensions.
Results. As depicted in Fig. 2a, we found that model selection through the lower bound identifies in a range around 16 the number of latent dimensions that optimally describe the observations. When fixing 16 latent dimensions, in one of them (Fig. 2c) subjects appear stratified by disease status, an information that was not directly introduced ahead. For each model, the classification accuracy in predicting the disease status was assessed through split-half cross-validation linear discriminant analysis on the latent variables (Fig. 2b). Maximum accuracy for disease classification occurs at 16 and 32 latent dimensions, an optimum location also identified through the lower bound. Fig. 3 shows the generative parameters $\boldsymbol{\phi}_c$ of the four channels associated to the latent dimension shown in Fig. 2c. The generative parameters describe a plausible relationship between this latent dimension and the heterogeneous observations in the data channels.

4 Discussion and Conclusion

We presented a multi-channel stochastic framework based on a probabilistic generative formulation. The performance of our multi-channel model was shown in the case of Gaussian distributions with moments parametrized by linear functions. In the real case scenario of AD modeling, the model allowed the unsupervised stratification of the latent variable by disease status, providing evidence for a physiological interpretation of the latent space. The generative

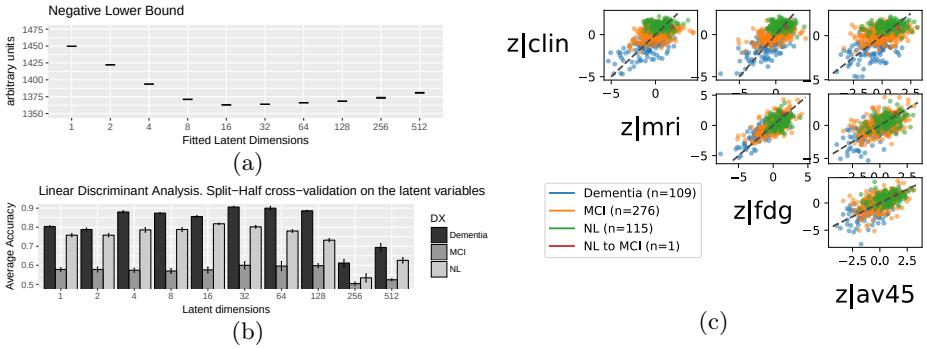


Fig. 2: Modeling results on ADNI data. (a) The negative lower bound has a minimum when fitting 16 latent dimensions. (b) Classification performance of the models: maximum accuracy for classes identification occurs with 16 and 32 lat. dims., in agreement with (a). (c) Pairwise representations of one latent dimension (out of 16) inferred from each of the four data channel. Although the optimization is not supervised to enforce clustering, subjects appear stratified by disease classes.

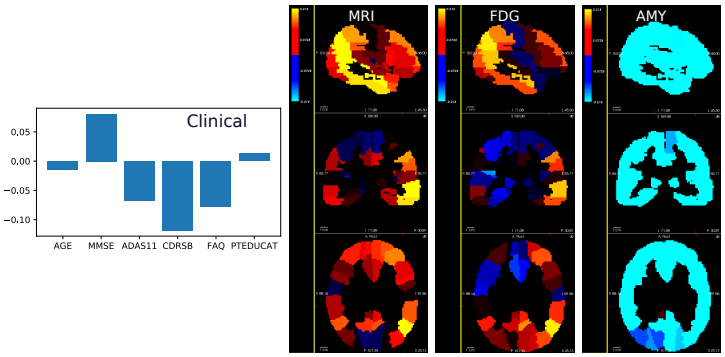


Fig. 3: Generative parameters $\phi_c^{(\mu)}$ of the four channels associated to the latent dimension in Fig. 2c. The clinical parameters are age, mini-mental state examination (mmse), adas-cog (adas11), cdr-sb, faq, scholarity (pteducat); The generative parameters describe a plausible relationship between the latent variable and the heterogeneous observations in the ADNI dataset, coherently with the research literature on Alzheimer's Disease (e.g. low amyloid deposition, high mmse, high scholarity, low cdr, etc.).

parameters can therefore encode clinically meaningful relationships across multi-channel observations. Although the use of the lower bound for model selection presents theoretical limitations [2], we found that it leads to good approximation of the marginal likelihood, thus providing a basis for model selection.

Future extension of this work will concern model with non-linear parameterization of the distributions, easily implementable through deep neural networks. The use of non-Gaussian distributions can also be tested. Given the scalability of our variational model, application to high resolution images may be also easily implemented. To increase the model classification performance, supervised clustering of the latent space will be introduced, for example, by adding an appropriate cost function to the lower bound. Also, introducing sparsity to remove redundancies may ease the identification and interpretation of the most informative parameters. Lastly, due to the general formulation, the proposed method can find various applications as a general data fusion technique, not limited to the biomedical research area.

Acknowledgments. This work has been supported by: the French government, through the UCA^{JEDI} Investments in the Future project managed by the National Research Agency (ref. n. ANR-15-IDEX-01); the grant AAP Santé 06 2017-260 DGA-DSH, and by the Inria Sophia Antipolis - Méditerranée, "NEF" computation cluster.

Bibliography

[1] Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep Canonical Correlation Analysis. Proc. Mach. Learn. Res. **28**(3), 1247–1255 (2013)
[2] Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational Inference: A Review for Statisticians (2016), http://arxiv.org/abs/1601.00670
[3] Dubois, B., Feldman, H.H., Jacova, C., Hampel, H., Molinuevo, J.L., et al.: Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet. Neurol. **13**(6), 614–29 (jun 2014)
[4] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., et al.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage **87**, 96–110 (2014)
[5] Hotelling, H.: Relations Between Two Sets of Variates. Biometrika **28**(3/4), 321 (dec 1936)
[6] Huang, S.Y., Lee, M.H., Hsiao, C.K.: Nonlinear measures of association with kernel canonical correlation analysis and applications. J. Stat. Plan. Inference **139**(7), 2162–2174 (2009)
[7] Jack, C.R., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., et al.: NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's Dement. **14**(4), 535–562 (2018)
[8] Kettenring, J.R.: Canonical analysis of several sets of variables. Biometrika **58**(3), 433–451 (1971)
[9] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. CoRR **abs/1412.6** (2014), http://arxiv.org/abs/1412.6980
[10] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: Proc. 2nd Int. Conf. Learn. Represent. (ICLR2014). (dec 2014), http://arxiv.org/abs/1312.6114
[11] Klami, A., Seppo, V., Kaski, S.: Bayesian Canonical Correlation Analysis. J. Mach. Learn. Res. **14**, 965–1003 (2013)
[12] Liu, J., Calhoun, V.D.: A review of multivariate analyses in imaging genetics. Front. Neuroinform. **8**, 29 (mar 2014)
[13] Luo, Y., Tao, D., Ramamohanarao, K., Xu, C., Wen, Y.: Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction. IEEE Trans. Knowl. Data Eng. **27**(11), 3111–3124 (2015)
[14] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., et al.: Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. Neuroimage **15**(1), 273–289 (jan 2002)

Supplementary Material

Derivation of the Lower Bound

In the following derivation we will interchangeably use $\mathbf{x}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_C$ to leave the notation uncluttered. For the same reason we will omit the variational and generative parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

Variational inference is carried out by introducing a set of probability density functions $q(\mathbf{z}|\mathbf{x}_c)$, belonging to a distribution family $\mathcal{Q}$, that are on average as close as possible to the true posterior over the latent variable $p(\mathbf{z}|\mathbf{x})$. In other words we aim to solve the following minimization problem:

$$\underset{q \in \mathcal{Q}}{arg\,min}\ \mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_C)\big) \right] \tag{4}$$

Given the intractability of $p(\mathbf{z}|\mathbf{x})$ for most complex models, we cannot solve directly this optimization problem. We look then for an equivalent problem, by rearranging the objective:

$$\begin{aligned}
&\mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z}|\mathbf{x})\big) \right] = \mathbb{E}_c \left[\int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}_c) \left(\ln q(\mathbf{z}|\mathbf{x}_c) - \ln p(\mathbf{z}|\mathbf{x}) \right) d\mathbf{z} \right] \\
&= \mathbb{E}_c \left[\int_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}_c) \left(\ln q(\mathbf{z}|x_c) - \ln p(\mathbf{x}|\mathbf{z}) - \ln p(\mathbf{z}) + \ln p(\mathbf{x}) \right) d\mathbf{z} \right] \\
&= \ln p(\mathbf{x}) + \mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z})\big) - \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)} \left[\ln p(\mathbf{x}|\mathbf{z}) \right] \right]
\end{aligned} \tag{5}$$

where in the middle line we use the Bayes' theorem to factorize the true posterior $p(\mathbf{z}|\mathbf{x})$. Now, we can reorganize the terms, such that:

$$\begin{aligned}
&\ln p(\mathbf{x}) - \underbrace{\mathbb{E}_c \left[\mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z}|\mathbf{x})\big) \right]}_{\geq 0} = \\
&= \underbrace{\mathbb{E}_c \left[\mathbb{E}_{q(\mathbf{z})} \left[\ln p(\mathbf{x}_1, \ldots, \mathbf{x}_C|\mathbf{z}) \right] - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z})\big) \right]}_{\text{lower bound } \mathcal{L}}
\end{aligned} \tag{6}$$

Since the KL term in the left hand side is always non-negative, the right hand side lower bounds the log evidence. By maximizing the lower bound we obtain the result of maximizing the data log evidence while solving the minimization problem in (4).

The hypothesis that every channel is conditionally independent from all the others given $\mathbf{z}$, allows to factorize the data likelihood as $p(\mathbf{x}_1, \ldots, \mathbf{x}_C|\mathbf{z}) = \prod_{i=1}^{C} p(\mathbf{x}_i|\mathbf{z})$, so that the lower bound becomes:

$$\mathcal{L} = \mathbb{E}_c \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)} \left[\sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z}) \right] - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z})\big) \right] \tag{7}$$

Finally, assuming every channel is equally likely to be observed with probability $1/C$, we can rewrite equation (6) as:

$$\ln p(\mathbf{x}_1, \ldots, \mathbf{x}_C) \geq \frac{1}{C} \sum_{c=1}^{C} \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_c)} \left[\sum_{i=1}^{C} \ln p(\mathbf{x}_i|\mathbf{z}) \right] - \mathcal{D}_{KL}\big(q(\mathbf{z}|\mathbf{x}_c) \,||\, p(\mathbf{z})\big) \tag{8}$$

Data generation procedure

Datasets $\mathbf{x} = \{\mathbf{x}_c\}$ with $c = 1..C$ channels where created according to the following model:

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}_l) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}_{d_c}) \\ \mathbf{G}_c &= diag\left(\mathbf{R}_c \mathbf{R}_c^T\right)^{-1/2} \mathbf{R}_c \\ \mathbf{x}_c &= \mathbf{G}_c \mathbf{z} + snr^{-1/2} \cdot \boldsymbol{\epsilon} \end{aligned} \tag{9}$$

where for every channel c, $\mathbf{R}_c \in \mathbb{R}^{d_c \times l}$ is a random matrix with l orthonormal columns (*i.e.,* $\mathbf{R}_c^T \mathbf{R}_c = \mathbf{I}_l$), $\mathbf{G}_c$ is the linear generative law, and snr is the signal-to-noise ratio. It's easy to demonstrate that the diagonal elements of the covariance matrix of $\mathbf{x}_c$ are inversely proportional to snr, *i.e.,*$diag\left(\mathbb{E}\left[\mathbf{x}_c \mathbf{x}_c^T\right]\right) = (1 + snr^{-1})\mathbf{I}_{d_c}$. Scenarios where generated by varying one-at-a-time the dataset attributes, as listed in Tab. 1.

Table 1: Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to Eq. (9).

Attribute description	range / iteration list	symbol
Total channels	2, 3, 5, 10	C
Channel dimension	4, 8, 16, 32, 500	d_c
Latent space dimension	1, 2, 4, 10, 20	l
Number of samples/observations	50, 100, 1000, 10000	S
Signal-to-noise ratio	100, 10, 1, 0.1	snr
Replication number (re-initialize $\mathbf{R}_c$)	1, 2, 3, 4, 5	

Acknowledgments

This work has been supported by:

- the French government, through the UCA$^{\text{JEDI}}$ 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.
- Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through

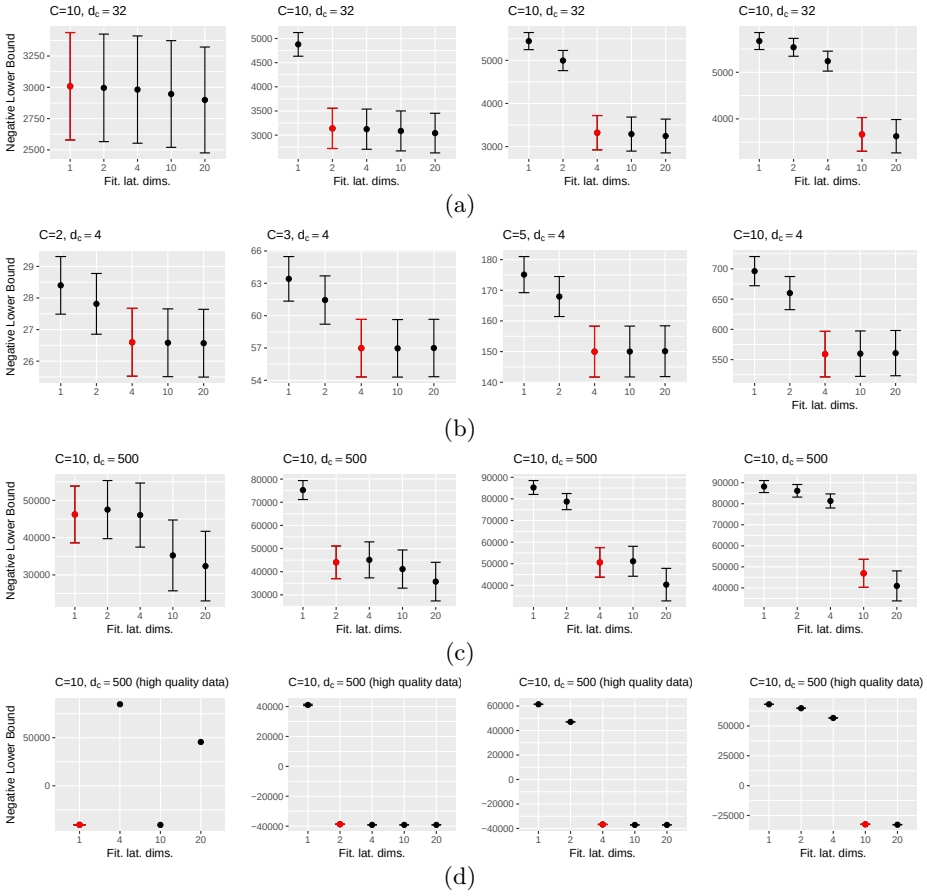
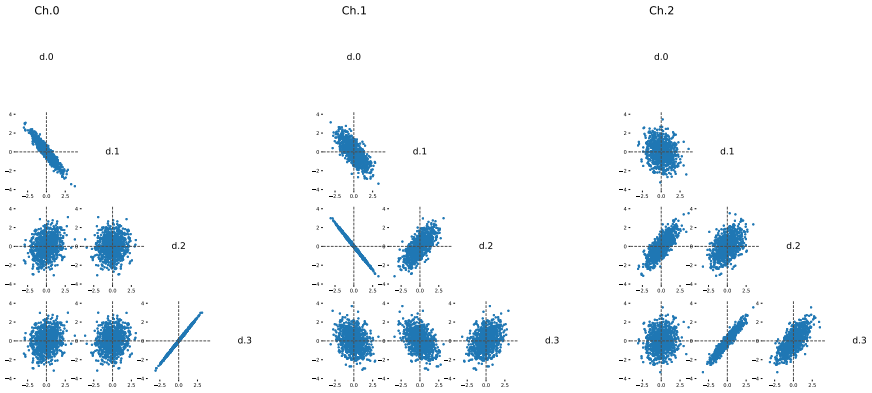
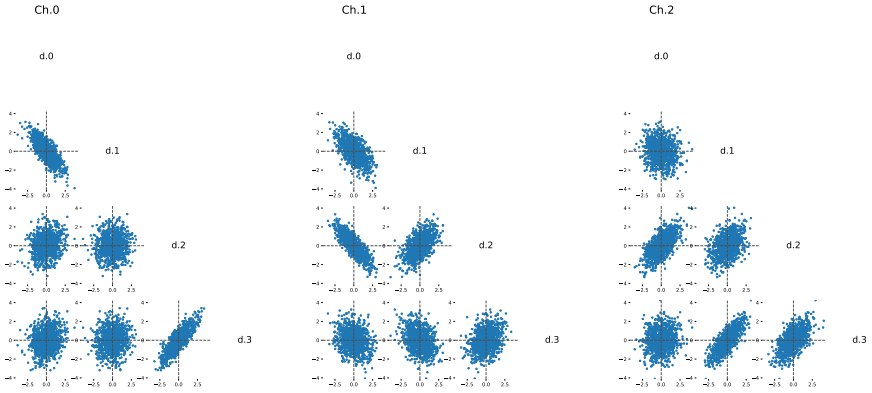


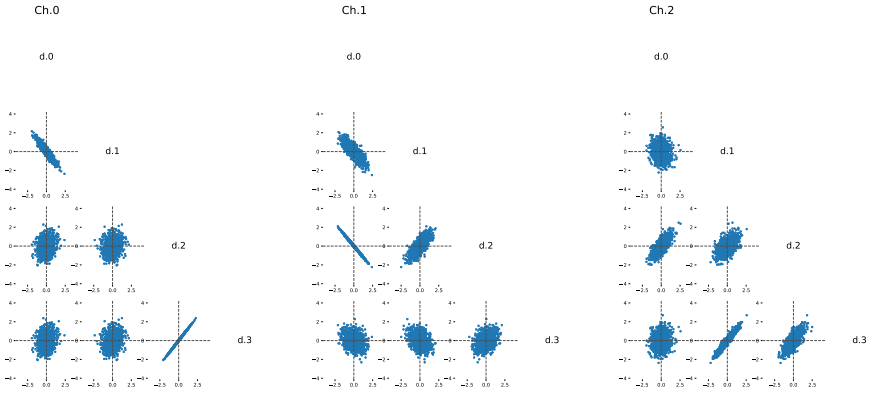
Fig. 4: Negative lower bound (NLB) on the synthetic training set computed at convergence for all the scenarios. Each bar shows mean $\pm$ std.err. of $N = 80$ total experiments as a function of the number of fitted latent dimensions. Red bars represents experiments where the number of true and fitted latent dimensions coincide. (top) Experimental setup $C = 10$, $d_c = 32$: NLB stops decreasing when the number of fitted latent dimension coincide with the generated ones; notable gap between the under-fitted and over-fitted experiments (elbow effect). (2^{nd} row) Experimental setup $d_c = 4$, $l = 4$: increasing the number of channels C makes the elbow effect more pronounced. (3^{rd} row) Experimental setup $C = 10$, $d_c = 500$: with high dimensional data ($d_c = 500$) using the lower bound as a model selection criteria to assess the true number of latent dimensions may end up in overestimation. (bottom) Restricted ($N = 5$ total experiments) high quality experimental setup $C = 10$, $d_c = 500$, $S = 10000$, $snr = 100$: the risk to overestimate the true number of latent dimensions can be mitigated by increasing the snr and S of the observations in the dataset.



(a) Ground truth



(b) Noisy observations ($snr = 5$)



(c) Reconstruction

Fig. 5: Pairwise representation of the four dimensions d of three data channels Ch, generated from a two-dimensional latent dimension $\mathbf{z} \sim \mathcal{N}(\mathbf{0}; \mathbf{I})$, according to Eq. (9). Noisy data was fitted with our model with a linear reparameterization. (a) Ground truth ($snr = 0$). (b) Observations used to fit the multi-channel model ($snr = 5$). (c) Data generated from the latent variable inferred from the noisy data.

generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.