Deep Linear Modeling of Hierarchical Functional Connectivity in the Human Brain

Wei Zhang[1], Eva Palacios[1], Pratik Mukherjee[1,2#]

[1]Department of Radiology and Biomedical Imaging,

[2]Department of Bioengineering and Therapeutic Sciences,

University of California San Francisco,

San Francisco, CA, 94143-0628, USA;

E-mail: pratik.mukherjee@ucsf.edu

[#]Corresponding author

Abstract

The human brain exhibits hierarchical modular organization, which is not depicted by conventional fMRI functional connectivity reconstruction methods such as independent component analysis (ICA). To map hierarchical brain connectivity networks (BCNs), we propose a novel class of deep (multilayer) linear models that are constructed such that each successive layer decomposes the features of the preceding layer. Three of these are multilayer variants of Sparse Dictionary Learning (SDL), Non-Negative Matrix Factorization (NMF) and Fast ICA (FICA). We present a fourth deep linear model, Deep Matrix Fitting (MF), which incorporates both rank reduction for data-driven hyperparameter determination as well as a distributed optimization function. We also introduce a novel framework for theoretical comparison of these deep linear models based on their combination of mathematical operators, the predictions of which are tested using simulated resting state fMRI data with known ground truth BCNs. Consistent with the theoretical predictions, Deep MF and Deep SDL performed best for connectivity estimation of 1st layer networks, whereas Deep FICA and Deep NMF were modestly better for spatial mapping. Deep MF provided the best overall performance, including computational speed. These deep linear models can efficiently map hierarchical BCNs without requiring the manual hyperparameter tuning, extensive fMRI training data or high-performance computing infrastructure needed by deep nonlinear models, such as convolutional neural networks (CNNs) or deep belief networks (DBNs), and their results are also more explainable from their mathematical structure. These benefits gain in importance as continual improvements in the spatial and temporal resolution of fMRI reveal more of the hierarchy of spatiotemporal brain architecture. These new models of hierarchical BCNs may also advance the development of fMRI diagnostic and prognostic biomarkers, given the recent recognition of disparities between low-level vs high-level network connectivity across a wide range of neurological and psychiatric disorders.

Keywords: fMRI, Deep Learning, Functional Connectivity, Hierarchical Networks, Linear Algorithms, Resting State

Abbreviations

ADMM: Alternating Direction Method of Multipliers

BCN: Brain Connectivity Network

CNN: Convolutional Neural Network

CPU: Central Processing Unit

DBN: Deep Belief Network

DCAE: Deep Convolutional Auto Encoder

Deep FICA: Deep Fast Independent Component Analysis

Deep MF: Deep Matrix Fitting

Deep NMF: Deep Non-negative Matrix Factorization

Deep SDL: Deep Sparse Dictionary Learning

DNN: Deep Neural Network

fMRI: Functional Magnetic Resonance Imaging

GD: Gradient Descent

GLM: General Linear Model

GPU: Graphics Processing Unit

HD: Hausdorff Distance

ICA: Independent Component Analysis

IS: Intensity Similarity

LASSO: Least Absolute Shrinkage and Selection Operator

RBM: Restricted Boltzmann Machine

RRO: Rank Reduction Operator

rsfMRI: Resting-State Functional MRI

SS: Spatial Similarity

tfMRI: Task-Evoked Functional MRI

TBI: Traumatic Brain Injury

TPU: Tensor Processing Unit

1. Introduction

Functional Magnetic Resonance Imaging (fMRI) has been widely used for the identification of brain connectivity networks (BCNs) (Bartels et al., 2005; Beckmann et al., 2005; Biswal et al., 1995, 2010; Bullmore et al., 2009; Duncan et al., 2010; Stam et al., 2014). A variety of scientific investigations have already demonstrated the hierarchical modular organization of human brain networks (Bassett et al., 2008; Biswal et al., 2005; Bullmore et al., 2009; Sporns et al., 2004). The architecture of cortical BCNs is organized at different spatial scales, from both functional and structural perspectives, ranging from local circuits at the microscale to columns and layers at the mesoscale to areas and areal networks at the macroscale (Bullmore et al., 2009; Power et al., 2011; Stam et al., 2014; Sporns et al., 2004).

In the last two decades, a variety of computational methods have been developed to detect BCNs, e.g., General Linear Modeling (GLM), Graph Theory, Independent Component Analysis (ICA), and Sparse Dictionary Learning (SDL) (Andersen et al., 1999; Calhoun et al., 2001; Lee et al., 2011; Lee et al., 2016; Lv et al., 2015; Zhang et al., 2017; Zhang et al., 2018; Zhang et al., 2019). However, these methods are based on a ‘shallow’ framework that cannot identify in unsupervised data-driven fashion the hierarchical and spatially overlapping organization of BCNs using resting-state fMRI (rsfMRI) or task-evoked fMRI (tfMRI) signals (Hu et al., 2018; Huang et al., 2018; Zhang et al., 2019; Zhang et al., 2020). Traditionally, the hierarchical spatial organization of BCNs has been indicated by varying the number of features in shallow linear models, for example from low to high numbers of independent components in ICA (Iraji et al., 2019; Smith et al., 2009), and noting that smaller networks at the more granular decomposition tend to merge or otherwise recombine to form larger networks at the coarser decomposition. However, there is no principled, unsupervised way to map this hierarchical organization with shallow methods.

Fortunately, with the advent of deep learning, algorithms have been developed that are capable of reconstructing hierarchical network architectures, e.g., the Deep Convolutional Auto Encoder (DCAE), Deep Belief Network (DBN) and Convolutional Neural Network (CNN)

(Bengio et al., 2012; Esteva, et al., 2019; Gurovich et al., 2019; Hannun et al., 2019; LeCun et al., 2015; Plis et al., 2014; Schmidhuber et al., 2015; Suk et al., 2014; Suk et al., 2016; Zhang et al., 2020). The Restricted Boltzmann Machine (RBM) can be used to model fMRI time series signals and effectively reconstruct functional brain networks with impressive accuracy (Hu et al., 2018; Huang et al., 2018). Moreover, other recent studies reported meaningfully hierarchical temporal organization of tfMRI time series, each with corresponding task-evoked BCNs (Hu et al., 2018; Zhang et al., 2019; Zhang et al., 2020) using DCAE, RBM and DBN. In general, these machine learning techniques are considered to be deep nonlinear models, e.g., deep neural networks (DNN). Although these nonlinear models such as DBN have recently proven effective at hierarchical spatiotemporal decomposition of task-evoked fMRI data (Dong et al., 2020), there are several disadvantages: (i) large training samples; (ii) extensive computational resources, e.g., graphics processing units (GPUs) or tensor processing units (TPUs); (iii) manual tuning of hyperparameters; (iv) time-consuming training process; (v) non-convergence to the global optimum; and (vi) "black box" results that lack explainability. Deep linear algorithms can overcome all these shortcomings of nonlinear techniques, since they are fast even on conventional central processing units (CPUs) with hyperparameters that can be automatically determined and with convex optimization functions that are guaranteed to converge. Furthermore, as we show in the theoretical analysis below, important aspects of their behavior can be explained from their relatively simple mathematical structure. For fMRI research, these deep linear models can detect BCNs using data from relatively few experimental subjects compared to deep nonlinear models and may prove especially useful and efficient as the spatial and temporal resolution of fMRI continues to improve, revealing more of the hierarchy of brain organization.

For hierarchical spatial functional connectivity mapping, we adopt a compositional approach to develop multilayer versions of SDL (Deep SDL), Fast ICA (Deep FICA; Seo, 2018) and Non-negative Matrix Factorization (Deep NMF; Trigeorgis et al., 2016), as well as a novel multilayer linear model that we name Deep Matrix Fitting (Deep MF). We contrast these deep linear algorithms for fMRI functional connectivity analysis in two ways. First, we employ theory to investigate the mathematical properties of these models, in order to predict differences in, e.g., network sparsity, connectivity strength and convergence velocity. Second, we conduct *in silico* connectivity reconstruction experiments using simulated fMRI signal time series to test the predictions of the theoretical analyses for the relative performance of the four deep linear models for fMRI brain network mapping. This leads to clear conclusions about the strengths and weaknesses of each method and provides a guide to the research community for applying this novel class of network reconstruction methodologies as well as for developing new ones.

2. Methods

2.1 Shallow versus Deep Linear Models of fMRI Functional Connectivity

The following introductions in Sections 2.2 to 2.5 provide the fundamentals of each deep linear model. Furthermore, these descriptions are prerequisites to analyze the theoretical properties of each model in the succeeding sections. Figure 1 compares the computational steps of a deep linear model versus that of a conventional “shallow” linear model:

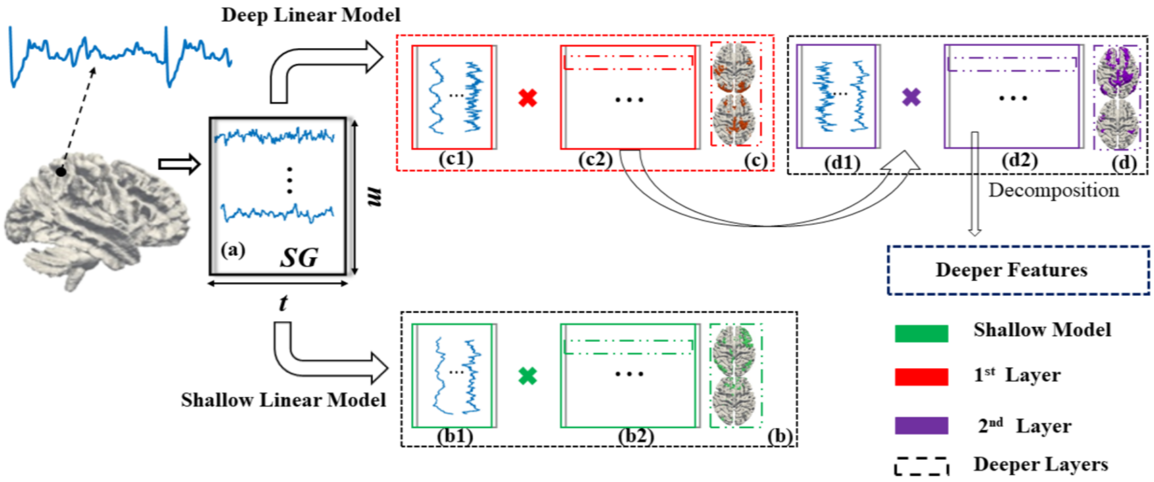


Figure 1. Deep linear model versus shallow linear model. The shallow model has only a single layer, i.e., a single decomposition. The deep linear model is constructed via multiple layers, i.e., continuous decomposition. (**a**) *SG* represents the input fMRI signal matrix; it contains the *t* time points and *m* voxels. (**b**) describes the pipeline of a shallow model in which the original input signal is decomposed into the weight matrix/dictionary (shown as **b1**) and feature matrix, i.e. connectivity networks (shown as **b2**). (**c**) and (**d**) represent the 1st and 2nd layers of the linear deep model, respectively. (**c1**) represents the 1st layer weight matrix/dictionary identified via SG. (**c2**) represents the 1st layer feature matrix, i.e. connectivity networks, recognized via SG. Similarly, (**d1**) and (**d2**) represent the corresponding matrices of the 2nd layer, which are both derived from the 1st layer feature matrix. The dashed blue rectangle indicates the deeper features beyond the 2nd layer that are derived from the 2nd layer feature matrix.

2.2 Deep Matrix Fitting

We propose a novel and efficient deep linear model that we name Deep Matrix Fitting. This algorithm aims to detect the hierarchical and overlapping organization of BCNs better than previously described data-driven functional connectivity reconstruction methods, e.g., ICA, SDL, DCAE and DBN (Calhoun et al., 2001; Lv et al., 2015; Zhang et al., 2020; Hinton and Salakhutdinov, 2006; Hinton et al., 2012). Due to the constraints of spatial independence in ICA, some investigators have reported that ICA cannot easily identify extensively overlapped functional brain networks (Calhoun et al., 2001; McKeown and Sejnowski, 1998; Zhang et al., 2019). Although SDL can efficiently derive spatial features, i.e., functional brain networks, based on rsfMRI and tfMRI, it is very challenging to leverage the dictionary size, sparsity trade off and even number of layers to implement a deep SDL. To be specific, one must heuristically estimate the dictionary size and number of layers. Simply utilizing the same size of dictionary and number of layers can easily result in the vanishing of spatial features of deeper layers, due to iteratively using the ℓ_1 norm. Recent deep nonlinear models, such as DBN, can successfully reveal the architecture of hierarchical spatiotemporal features. Unfortunately, the probabilistic energy-based model of DBNs necessarily requires a large number of training samples to avoid overfitting. Furthermore, DBN requires extensive computational resources such as GPUs and even TPUs (Zhang et al., 2019; Zhang et al., 2020). The novel Deep MF proposed in this work successfully solves these aforementioned problems. Deep MF can automatically estimate the optimal dictionary size, sparsity trade-off and number of layers, using an operator of rank reduction (Wen et al., 2012; Shen et al., 2014). In other words, Deep MF does not require any manual hyperparameter tuning to decompose the rsfMRI signal matrix. Since Deep MF is a deep linear model, it should detect latent features faster than DBN while only requiring conventional CPUs. In general, Deep MF can be approximately considered as a deep SDL (described in Section 2.3) with the additional mechanism to automatically determine all crucial hyperparameters via rank reduction.

The equation governing Deep MF is:

$$min_{X_i, \quad Y_i, \quad S\in\mathbb{R}^{m\times n}} \left\| \prod_{i=1}^{M-1} X_i\, Y_M - SG \right\|_F^2 + \mu \sum_{i=1}^{M} \|Y_i\|_1 + \lambda \sum_{i=1}^{M} \|Z_i\|_1 \qquad (1)$$

$$X_i Y_i \leftarrow \mathcal{R}(Y_{i-1})$$

where $\{X_i\}_{i=1}^{M}$ represents the hierarchical dictionaries, e.g., X_i indicates the dictionary of the *i* th layer. $\{X_i\}_{i=1}^{M}$ is also considered as the time series in GLM and the weight matrix in ICA and DBN. M is the total number of layers. Similarly, $\{Y_i\}_{i=1}^{M}$ represents the hierarchical spatial features, e.g., Y_i indicates the spatial features of *i*th layer. $\{Y_i\}_{i=1}^{M}$ is also denoted as a correlation matrix. $\{Z_i\}_{i=1}^{M}$ are the matrices of background components, which is usually treated as the noise. $\mathcal{R}$ represents a rank reduction operator (RRO) to automatically estimate the hyperparameters and more details will be introduced in the following section. Naturally, we assume the spatial features Y_{i-1} can be decomposed as deeper dictionary X_i and spatial features Y_i, in order to implement the deep linear framework (Figure 1). Therefore, the original input data SG can be decomposed as $\prod_{i=1}^{M-1} X_i\, Y_M$. In Eq. (1), λ and μ are known as the sparse trade off to control the sparsity levels of background components and spatial features, respectively. In addition, in Eq. (1), $\|\cdot\|_F$ and $\|\cdot\|_1$ represent the Frobenius and ℓ_1 norms, respectively.

This optimization function, shown as Eq. (1), consists of more parameters than ICA and SDL. In general, SDL includes two parameters to be optimized: dictionary and correlation matrix. Naturally, it is easier to comprehensively employ alternative optimizer and shrinkage methods (Wen et al., 2012). Before optimizing Eq. (1), we need to convert Eq. (1) to an augmented Lagrangian function. If considering the k^{th} layer, we have:

$$\mathcal{L}_\beta\left(\prod_{i=1}^{k-1} X_i\, , Y_k, Z_k, \mathrm{e}_k\right) \overset{\text{def}}{=} \left\| \prod_{i=1}^{k-1} X_i\, Y_k - SG \right\|_F^2 + \left\langle \prod_{i=1}^{k-1} X_i\, Y_k - SG, \mathrm{e}_k \right\rangle \qquad (2)$$

For Eq. (2), for *k* layers (we assume the total number of layers as *k*), these can be solved using Alternating Direction of Method of Multipliers (ADMM) (Shen et al., 2014), and to solve $\sum_{i=1}^{k}\|Z_i\|_1$, we jointly utilize the shrinkage method. In Eq. (2), all parameters are as discussed before, with e_k defined as the multiplier. The ℓ_1 norm of Y_k and Z_k shown in Eq. (1) can be solved directly using the shrinkage method (Beck et al., 2009).

The iterative format to solve Eq. (2) using ADMM can be organized as follows:

$$X_k^{it+1} = argmin_{X_k^{it+1} \in \mathbb{R}^{m \times h_k}} \ \mathcal{L}_\beta(X_k^{it}, Y_k^{it}, Z_k^{it}, \mathrm{e}_k^{it}) \tag{3-1}$$

$$Y_k^{it+1} = argmin_{Y_k^{it+1} \in \mathbb{R}^{h_k \times n}} \ \mathcal{L}_\beta(X_k^{it+1}, Y_k^{it}, Z_k^{it}, \mathrm{e}_k^{it}) \tag{3-2}$$

$$Z_k^{it+1} = argmin_{Z_k^{it+1} \in \mathbb{R}^{m \times n}} \ \mathcal{L}_\beta(X_k^{it+1}, Y_k^{it+1}, Z_k^{it}, \mathrm{e}_k^{it}) \tag{3-3}$$

$$\mathrm{e}_k^{it+1} = \ \mathrm{e}_k^{it} + \beta(\prod_{i=1}^{k-1} X_i \, Y_k + \sum_{i=1}^{k} Z_k^{it+1} - SG) \tag{3-4}$$

Using ADMM, in each iteration (the current iteration is represented as *it*), we update a single parameter independently, and finally calculate the multiplier, based on the current error. Since, in each single step, only one parameter is optimized and others are fixed, Eq (2) is considered as a convex problem and the global optimum can be obtained via a descent algorithm, e.g. gradient descent (GD) or ADMM. In Eq. (3-4), β denotes the step length.

To automatically estimate the dictionary size and number of layers, we introduce the operator RRO. Briefly, RRO focuses on the identification of major components included in the raw data, and simultaneously determines which components are relatively weak and that will therefore be continuously merged into the background matrices. In general, RRO demonstrates that the number of units, i.e., dictionary size, should be consistently reduced, if considering deeper layers (Hinton et al., 2012; Zhang et al., 2019). In other words, the continuous increase of units in deeper layers can result in lack of convergence. If the number of units/dictionary size, i.e., the estimated rank of the matrix, is reduced to one, that indicates the decomposition should be terminated. Hence, the layer that owns a rank of unity should be considered the final layer. Deep MF employs RRO to continuously reduce the dictionary size and therefore also determine the number of layers. In fact, it does not require any manual design for the essential hyperparameters of deep learning models, such as the number of layers or unit number of each layer that are used in DBN and other peer deep models.

In detail, this rank estimator RRO employs a technique of rank-revealing by continuously using orthogonal decomposition, in this case via *QR* factorization (Wen et al., 2012; Shen et al., 2014). The advantage of *QR* is that it is faster and makes fewer requirements of the input matrix. For example, *QR* performs orthogonal decomposition faster than Singular Value Decomposition (*SVD*) and can solve incomplete (i.e., number of features < number of samples) and over-complete (i.e., number of features > number of samples) matrices.

At the beginning, *r** is denoted as the initial estimated rank of $S^i$and we denote *r* as the optimal rank estimation of input matrix S^i. If *r**≥*r* holds, the detection of the diagonal line of the upper-triangular matrix in the *QR* factorization can be performed using the input matrix S^i. If we can determine the ideal size of *QR* factorization using S^i in the work with permutation matrix *E*, the diagonal matrix *R* is non-increasing in magnitude (Wen et al., 2012; Shen et al., 2014). The *QR* factorization and rank-revealing will eventually provide a reasonable solution using a proper thresholding value introduced in Eq. (2) and Eq. (3) (Wen et al., 2012; Shen et al., 2014). By detecting the diagonal line of matrix *R*, we compute two vectors $d \in \mathbb{R}^r$ and $r \in \mathbb{R}^{r-1}$:

$$d_i = |R_{ii}| \qquad r_i = \frac{d_i}{d_{i+1}} \tag{4}$$

And then examine the value:

$$\xi = \frac{(m-1)r(p)}{\sum_{i \neq p} r_i} \tag{5}$$

where *r(p)* is the maximum element of the vector *r* (with the largest index *p* if the maximum value is not unique). In our current implementation, we reset the rank estimated *r*, if $\xi > 2$, and this adjustment can be successfully done only once (Wen et al., 2012; Shen et al., 2014).

The mathematical definition of RRO is shown below:

$$\mathcal{R}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} = \begin{bmatrix} {a_1}^{(1)} \\ {a_2}^{(1)} \\ \vdots \\ {a_{n-2}}^{(1)} \\ {a_{n-1}}^{(1)} \end{bmatrix} \quad \mathcal{R}^k\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} = \begin{bmatrix} {a_1}^{(1)} \\ {a_2}^{(1)} \\ \vdots \\ {a_{n-k-1}}^{(1)} \\ {a_{n-k}}^{(1)} \end{bmatrix} \tag{6}$$

where $\mathcal{R}$ denotes the RRO operator; and theoretically, we have $\mathcal{R}^k[a_1, a_2, \cdots, a_n] = [\hat{a}]$, if $k \to \infty$. It clearly demonstrates that the RRO can continuously maintain the vital components and reduce the dimensions of the original data. By continuously using the technique of low rank estimation, Deep MF implements automatic estimation of dictionary size and number of layers. Also, we provide a theoretical analysis of each operator and experimental validation of Deep MF, in Sections 3 and 4 respectively, by comparing with three other deep linear models, specifically, Deep SDL, Deep FICA and Deep NMF.

2.3 Deep Sparse Dictionary Learning

In the last decade, sparse dictionary learning (SDL), widely known as the algorithm Online Dictionary Learning (ODL) (Mairal et al., 2010; Liu et al., 2010), has been successfully applied to identify the concurrent BCNs of the human brain and the non-human primate brain from fMRI datasets (Lv et al., 2015; Zhang et al., 2018). In this category, to satisfy the requirements of hierarchical organization of BCNs, we propose a novel Deep SDL algorithm that is a multilayer extension of conventional shallow SDL-based methods. Briefly, fMRI signals from all voxels within the whole brain are extracted and are then organized as an extensive 2D matrix, where the number of columns represents the total brain voxels and the number of rows stands for the time points. For the first layer, the input 2D matrix is decomposed into the product of an incomplete/over-complete dictionary basis matrix (each atom representing a time series) and a feature matrix (representing this network's spatial volumetric distribution). For each successive layer, the current features matrix is treated as an input matrix to be continuously decomposed. A particularly important characteristic of this Deep SDL framework is its ability to carry out over-complete decomposition for all layers; but, considering the finite features, the deep layers only concentrate on the incomplete decomposition.

If considering all layers, the optimization function of Deep SDL is:

$$min_{D_i,\quad Y_i\in\mathbb{R}^{m\times n}} \left\| \prod_{i=1}^{M-1} D_i\, Y_M - SG \right\|_F^2 + [\mu_1 \;\; \mu_2 \cdots \mu_M] \otimes \sum_{i=1}^{M} \|Y_i\|_1 \tag{7}$$

$$D_i Y_i \leftarrow Y_{i-1}$$

In Eq. (7), $\{D_i\}_{i=1}^{M}$ denotes the set of dictionaries; *i* represents the number of current layer, and *M* is the total number of layers, $\{Y_i\}_{i=1}^{M}$ defines the set of hierarchical features, i.e., BCNs. And $\{\mu_i\}_{i=1}^{M}$ represents all sparsity trade-offs for all layers, respectively.

Deep SDL, like ODL, utilizes GD as the optimizer to update all parameters. Simple GD is an efficient optimizer, but only guarantees the convergence of convex problems. In Section 3.1, we also provide the requirements of GD to be a contraction operator. Briefly, the convergence property of GD heavily depends on the step length. The following equation shows the iterative format of simple GD:

$$x_{k+1} \leftarrow x_k - \lambda f_x' \;\; \lambda \in (0,1) \tag{8}$$

Compared with the ADMM optimizer used in Deep MF, the iterative function GD of Deep SDL is very simple. ADMM can utilize the current optima to update, and can be faster than GD, but ADMM requires that we can obtain the derivative of the optimization function.

2.4 Deep Fast Independent Component Analysis

ICA is a very popular and widely used data-driven computational technique, which was introduced to fMRI research over two decades ago (McKeown and Sejnowski 1998). In previous work, investigators have already reported that FICA using the Fixed-Point algorithm as an optimizer can be a very robust method (Hyvarinen, 1999). Inspired by Deep MF and FICA, the novel framework of Deep FICA aims to detect the hierarchically organized components. For simple shallow FICA, the original input matrix is decomposed as the weight matrix and independent component (IC) matrix. Applied to fMRI data, the ICs represent the BCNs. Similar to Deep MF and Deep SDL, in each layer of Deep FICA, the previous IC matrix is considered as the input signal matrix that will be decomposed using PCA and the Fixed-Point algorithm continuously (Figure 1). Deep FICA extracts only spatially independent features and can only solve the incomplete decomposition problem and not over-complete decomposition.

The optimization function of Deep FICA is:

$$max \sum_{k=1}^{M} \sum_{i=1}^{N} J_G(W_i^k) \tag{9}$$

In Eq. (9), W_i^k represents the i^{th} IC from the k^{th} layer. The maximum $J_G(\cdot)$ indicates the independency of each potential IC.

Considering each IC, Deep FICA and FICA both utilize an efficacy Fixed-Point algorithm to update the IC:

$$W_{it+1}^{k} \leftarrow \frac{W_{it}^{k}}{\sqrt{\|W_{it}^{k} c_{it} (W_{it}^{k})^{T}\|}} \tag{10}$$

Compared with Deep MF and Deep SDL, Deep FICA is relatively easy to implement, since it does not include more complex algorithms, e.g., the RRO or the sparsity operator.

2.5 Deep Non-negative Matrix Factorization

Non-negative Matrix factorization (NMF) is a particularly useful family of techniques in data analysis. Before the wide utilization of the Deep CNN, NMF was a crucial technique to identify the features of a human face (Trigeorgis et al., 2016). In recent years, there has been a significant amount of research on deep factorization methods that focus on particular characteristics of both the data matrix and the hierarchical resulting factors. The application area of the family of NMF algorithms has grown significantly during recent years. It has been demonstrated that NMF can be a successful dimensionality reduction technique over a variety of application areas including, but not limited to, environmetrics, microarray data analysis, document clustering, face recognition and more. Moreover, due to its particular non-negative constraints, NMF can also be directly utilized to analyze the fMRI data/signal (Lee & Seung, 1999). Deep NMF provides an opportunity to detect the potentially hierarchical structures of BCNs.

Deep NMF focuses on the decomposition of the non-negative multivariate data matrix into hierarchical factors $\{Z_i\}_{i=1}^{M}$ and $\{H_i\}_{i=1}^{M}$, such that (Trigeorgis et al., 2016):

$$\begin{aligned} &Function_{Deep} \overset{\text{def}}{=} \frac{1}{2}\|SG - Z_1 \cdot Z_2 \cdot \cdots \cdot Z_M H_M\|_F^2 \\ &= tr[SG^T SG - 2SG^T Z_1 \cdot Z_2 \cdot \cdots \cdot Z_M + SG_M^T \cdot Z_1 \cdot Z_2 \cdot \cdots \cdot Z_M \cdot Z_1^T \\ &\cdot Z_2^T \cdot \cdots \cdot Z_M^T \cdot H_M] \end{aligned} \tag{11}$$

In Eq. (11), SG represents the input fMRI signal, and $\{Z_i\}_{i=1}^{M}$ represents the weight matrix; $\{H_i\}_{i=1}^{M}$ denotes the sets of non-negative components. M denotes the total number of layers. To calculate the optimal solutions of $\{Z_i\}_{i=1}^{M}$ and $\{H_i\}_{i=1}^{M}$ requires minimizing the loss function $Function_{Deep}$ in Eq. (11). And tr represents the trace of the matrix.

A key difference between Deep NMF versus Deep SDL and Deep MF is the updating principle. Unlike Deep SDL and Deep MF, Deep NMF employs a fast policy to update these two factors: $\{Z_i\}_{i=1}^{M}$ and $\{H_i\}_{i=1}^{M}$ (Trigeorgis et al., 2016). This principle is shown as follows:

$$\begin{gathered} H_{it+1}^{k} \leftarrow \frac{H_{it}^{k}}{f_{max}^{H}} \\ Z_{it+1}^{k} \leftarrow SG \cdot (H_{it+1}^{k})^{\dagger} \end{gathered} \tag{12}$$

where H_{it}^{k} and Z_{it}^{k} represent the non-negative components and weight matrix from the k^{th} layer, iteration number *it*. And operator $(\cdot)^{\dagger}$ represents the pseudo-inverse of the input matrix (Trigeorgis et al., 2016). The f_{max}^{H} denotes the current maximum value of function *f*, related to H_{it}^{k}.

Intuitively, these four deep linear models are each distinctive. Deep MF can be more intelligent, and automatically determine all hyperparameters. Deep SDL can perform over-complete decomposition for each layer. Deep FICA reveals the spatially independent and hierarchical components, and has a faster convergence velocity. Finally, Deep NMF could detect the non-negative components included in fMRI signals. We theoretically analyze the relative performance of each deep linear model in the following section.

3. Results: Theoretical Analyses

In this section, we employ mathematical theory, specifically real analysis, linear functional analysis and abstract algebra, to explain why different deep linear models have the distinctive characteristics that they do. In particular, we hope to explain:

(i) The advantages of linear deep models over shallow models and deep nonlinear models;

(ii) Why some deep linear models, e.g., Deep MF and Deep SDL, converge slowly while others, e.g., Deep NMF and Deep FICA, converge quickly;

(iii) Why some deep linear models, e.g., Deep MF and Deep SDL, can better estimate connectivity strength, while others, e.g., Deep NMF and Deep FICA, can better estimate the spatial extent of connectivity networks.

3. 1 Fundamental Interpretation of Each Linear Model

All theoretical analyses are based on the vital assumption that all mappings/operators/algorithms must be applied on a finite dimensional space. Please consult Appendix A for the mathematical details (Assumption 1.1, Lemma 1.1, Theorem 1.1). If considering any algorithm and/or process as an operator, Assumption 1.1 and Lemma 1.1 demonstrate that the norm of the operator should be equivalent, in order to dramatically simplify our discussion.

According to Theorem 1.1, if considering the shallow linear, deep linear and deep nonlinear models as approximations of the original function $f(x)$, then, obviously, deeper models can employ more items such as $\{P_n(x)\}_{n=1}^{N}$ rather than just $P(x)$. Thus, the deeper models can more accurately approximate the original function than a shallow model. Meanwhile, nonlinear models can have: $\left\| \lim_{N\to\infty} \{P_n(x)\}_{n=1}^{N} - f(x) \right\| = 0$. But to optimize the infinite items, it will be very time-consuming or even impossible to solve a non-polynomial (NP) complexity problem. Hence, the theorem also answers why nonlinear models require a sampling technique to reduce the complexity, e.g., Gibbs sampling for DBN (Hinton and Salakhutdinov, 2006; Hinton et al., 2012).

According to the discussion in the last section, we can abstractly describe each deep linear model using the combination of several operators. All operators involved in this study are given in Table 1.

Table 1. All definitions of operators and their norms.

Operator	Definition	Operator/Norm	Definition
$\mathfrak{U}$	Deep MF	$\mathcal{U}$	Update Operator of Deep NMF
$\mathfrak{N}$	Deep NMF	$\mathcal{F}$	Operator of Fixed-Point Algorithm
$\mathfrak{L}$	Deep SDL	$\mathcal{C}$	Consistent Operator
$\mathfrak{T}$	Deep FICA	φ	Norm of ADMM
$\mathcal{M}$	Initialization Operator	ρ	Norm of GD
$\mathcal{S}$	Sparsity Operator	μ	Norm of Fixed-Point Algorithm
$\mathcal{P}$	Principal Component Analysis (PCA)	γ	Norm of Normalization in Deep NMF
$\mathcal{A}$	ADMM	δ	Norm of Updating Deep NMF
$\mathcal{G}$	GD	C	Norm of Input fMRI Matrix
$\mathcal{N}$	Normalization	SG	Input fMRI Signal Matrix
$\mathcal{R}$	Rank Reduction Operator	$\mathfrak{C}$	Set of Consistent Operators

In Table 2, we provide the definitions of sets involved in the following sections:

Table 2. All definitions of space and set

Space/Set	Definition
$\mathbb{N}$	Field of Natural Numbers
$\mathbb{R}$	Field of Real Numbers
$\mathbb{K}$	Field of Rational Numbers

3.2 Intensity Similarity

As discussed in Sections 2.2 to 2.5, we have the following definitions:

Definition 2.1 If we denote Deep MF as an operator $\mathfrak{U}$, based on the description of Deep MF, considering the iteration *k*, we can denote $\mathfrak{U} \overset{\text{def}}{=} M \cdot \mathcal{A}^k \cdot \mathcal{S}^k \cdot \mathcal{R}^k$.

Definition 2.2 If we denote Deep SDL as an operator $\mathfrak{L}$, based on the description of Deep

SDL, considering the iteration *k*, we can denote $\mathfrak{L} \overset{\text{def}}{=} M \cdot \mathcal{G}^k \cdot \mathcal{S}^k$.

Definition 2.3 If we denote Deep FICA as an operator $\mathfrak{I}$, based on the description of Deep FICA, considering the iteration *k*, we can denote $\mathfrak{I} \overset{\text{def}}{=} \mathcal{P} \cdot \mathcal{F}^k$.

Definition 2.4 If we denote Deep NMF as an operator $\mathfrak{N}$, based on the description of Deep NMF, considering the iteration *k*, we can denote $\mathfrak{N} \overset{\text{def}}{=} M \cdot \mathcal{U}^k \cdot \mathcal{N}$.

According to Definitions 2.1 to 2.4, as well as Corollaries 1.2 to 1.3 and Theorems 2.1 to 2.11 as proved in Appendix B, using the inequality of norm, considering the iteration *k*, and let *SG* be the input matrix; for any operator applied on *SG*, we can derive the features as: F_1, F_2, F_3, F_4; then we have:

$$\textit{Deep MF:}\quad \|F_1\| = \left\|\mathfrak{A}^k \cdot SG\right\| \leq \|M\| \cdot \left\|\mathcal{A}^k\right\| \cdot \left\|\mathcal{S}^k\right\| \cdot \left\|\mathcal{R}^k\right\| \cdot \|SG\| \tag{13-1}$$

$$\textit{Deep SDL:}\quad \|F_2\| = \left\|\mathfrak{L}^k \cdot SG\right\| \leq \|M\| \cdot \left\|\mathcal{G}^k\right\| \cdot \left\|\mathcal{S}^k\right\| \cdot \|SG\| \tag{13-2}$$

$$\textit{Deep FICA:}\quad \|F_3\| = \left\|\mathfrak{I}^k \cdot SG\right\| \leq \left\|\mathcal{P}^k\right\| \cdot \left\|\mathcal{F}^k\right\| \cdot \|SG\| \tag{13-3}$$

$$\textit{Deep NMF:}\quad \|F_4\| = \left\|\mathfrak{N}^k \cdot SG\right\| \leq \|M\| \cdot \left\|\mathcal{U}^k\right\| \cdot \|\mathcal{N}\| \cdot \|SG\| \tag{13-4}$$

According to Lemma 1.1 and Theorems 2.1 to 2.6, operators $\mathcal{A}^k$, $\mathcal{U}^k$, $\mathcal{N}$, $\mathcal{G}^k$, and $\mathcal{F}^k$ can be treated as contraction operators, which indicates that the norm of each operator should be larger than zero and smaller than one. Other operators are constant values, according to Theorems 2.1 to 2.11.

If we denote the norm of contraction operators as:

$$\left\|\mathcal{A}^k\right\| = \ \varphi < 1 \tag{14-1}$$

$$\left\|\mathcal{G}^k\right\| = \rho < 1 \tag{14-2}$$

$$\left\|\mathcal{F}^k\right\| = \ \mu < 1 \tag{14-3}$$

$$\left\|\mathcal{U}^k\right\| = \delta < 1;\ \ \|\mathcal{N}\| = \gamma < 1 \tag{14-4}$$

Despite the fact that the norms of operators $\mathcal{A}, \mathcal{U}, \ \mathcal{G}, \ \mathcal{F}$ are not equivalent, according to Theorems 3.1 and 3.2 (see Appendix C), we consider an extreme condition $k \to \infty$, and then we have: $\varphi = \delta = \rho = \mu$ that indicates convergence to the global optimum. Then we can rewrite all equations 13-1 to 13-4 as:

$$\textit{Deep MF:} \quad \|F_1\| \leq \varphi \cdot \|M\| \cdot \|\mathcal{S}^k\| \cdot \|\mathcal{R}^k\| \cdot C \tag{15-1}$$

$$\textit{Deep SDL:} \quad \|F_2\| \leq \rho \cdot \|M\| \cdot \|\mathcal{S}^k\| \cdot C \tag{15-2}$$

$$\textit{Deep FICA:} \quad \|F_3\| \leq \mu \cdot \|\mathcal{P}^k\| \cdot C \tag{15-3}$$

$$\textit{Deep NMF:} \quad \|F_4\| \leq \delta \cdot \gamma \cdot \|M\| \cdot C \tag{15-4}$$

Obviously, based on Eqs. (15-1) to (15-4), we have the conclusion:

$$\|F_4\| \leq \|F_3\| \leq \|F_2\| \leq \|F_1\| \tag{16}$$

Since all features $\{F_i\}_{i=1}^{4}$ have the same dimensions, this inequality Eq. (16) can clearly explain why the intensity of features, i.e., the connectivity strength of voxels in the networks, varies based on the different models. In particular, F_4 (the features derived from Deep NMF) should have the smallest intensity and F_1 (the features obtained via Deep MF) should have the largest intensity. Meanwhile, Eqs. (15-1 to 15-4) also reveal the convergence velocity of each model. Since Deep MF contains the most operators with the complex optimization function ADMM, it should be slowest. Because Deep SDL uses a sparsity operator as well as GD, which is relatively slow, it is comparable in speed to Deep MF, even given a perfect step-length. Theoretically, Deep FICA and Deep NMF should have faster convergence.

3.3 Spatial Similarity

Spatial matching is another important way to measure the similarity between identified components and templates. To examine this property, we use Assumptions 3.1 to 3.2 and Lemma 3.1 to prove Theorem 3.1 in Appendix C:

Theorem 3.1 If we denote the following sets:

$$Deep\ \ MF: D = \{\mathfrak{A}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\}$$

$$Deep\ \ SDL:\ L = \{\mathfrak{L}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\}$$

$$Deep\ \ FICA:\ I = \{\mathfrak{I}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\}$$

$$Deep\ \ NMF:\ \Theta = \{\mathfrak{N}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\}$$

And considering the iteration *k*, and $k > K$, it implies:

$$0 < \frac{\left|\mathfrak{A}^k V\right|}{|V \cup D|} \leq \frac{\left|\mathfrak{L}^k V\right|}{|V \cup L|} \leq \frac{\left|\mathfrak{T}^k V\right|}{|V \cup I|} \leq \frac{\left|\mathfrak{N}^k V\right|}{|V \cup \Theta|} \tag{17}$$

where $|\cdot|$ denotes the number of positive elements. Θ represents the set that only contains the element 0. T represents the functional regions of brain.

Since the convergence of deep models is a vital issue when solving real world problems (Topol, 2019), Eq. (16) and Theorem 3.1 can explain the convergence of all deep linear models, considering enough iterations. Clearly, if we examine the spatial similarity between two BCNs, according to Theorem 3.1, we can conclude: with the same number of iterations, Deep NMF has the best performance on spatial matching, Deep FICA has the next best performance, and Deep SDL and Deep MF have the least. That is, the norm of the operator of Deep NMF is very small, and is iteratively applied on functional regions and background noise, which causes the intensity of functional areas to decrease very rapidly. However, since the intensity of background is very small, the noise can be reduced to near zero much faster than Deep MF and Deep SDL. The performance of Deep FICA on spatial matching should be comparable to Deep NMF; since a normalization operator is involved in Deep NMF, the intensity of components identified by Deep NMF should be smaller than Deep FICA. Theorem 3.2 included in Appendix C also explains that all deep linear models finally converge given enough iterations.

To test these theoretical analyses, in the next section, a simulated experimental reconstruction will be introduced as the ground truth templates for the first layer BCNs. These templates will be employed to construct the simulated fMRI signal and all deep linear models will be applied on the simulated data and their 1st layer results will be compared to the templates. By examining the intensity similarity and spatial similarity to the ground truth templates, the correctness of the theoretical conclusions can be investigated.

4. Results: Experimental Validation

4.1 Simulated fMRI Data

In this work, we employ an *in silico* fMRI simulation method proposed previously (Zhang et al., 2018, 2019), using templates of BCNs (Smith et al., 2009) to test these proposed deep linear models. Specifically, we selected 12 BCNs (Table 3) that were originally derived using conventional shallow ICA (Smith et al., 2009).

Table 3. All abbreviations of BCNs in simulation

Name/Number	Abbreviation	Name	Abbreviation
Primary Visual Network/1	VIS-1	Auditory Network/7	AUD
Perception Visual Shape Network/2	VIS-2	Executive Control Network/8	ECN
Perception Visual Motion Network/3	VIS-3	Left Frontoparietal Network/9	FP-L
Default Mode Network/4	DMN	Right Frontoparietal Network/10	FP-R
Brainstem & Cerebellum Network/5	B/C	Dorsal Attention Network/11	DAN
Sensorimotor Network/6	SM	Salience Network/12	SN

These template BCNs derived from resting-state fMRI have been released publicly and are considered to be functional brain areas covering a large part of cerebral cortex (Smith et al., 2009). Since all deep linear models should be evaluated equally using a known ground truth, we employ a simulation of resting-state fMRI signals. Following the fMRI simulation pipeline for spatially independent networks named Experiment 1 in the previous study (Zhang et al., 2019), the 12 templates (Smith et al., 2009) are collected as components/spatial features; and we adopt 12 time series, including 200 time points, derived from a previous study

(Lv et al., 2015). The final simulation data is a matrix obtained as the product of time series and components.

The detailed parameters of each template are: 91×109 matrix, 91 slices, 2.0 mm isotropic voxels. The number of mask voxels is 262,309 and the number of time points is 200. All templates are registered to standard MNI space at 2.0 mm. This pipeline contains the steps of spatial artifact cleanup, distortions removal and cortical surfaces generation. After that, different subjects are aligned to the standard MNI space (Lv et al., 2015; Zhang et al., 2019).

Table 4 provides the main hyperparameter settings of the four proposed deep linear models, including: the number of components of the 1st and 2nd layers, the number of iterations and the step length of gradient descent, where applicable. Since the 1st layer may include noise components, we choose a larger number of components than the expected number of features, which in this case is at least a dozen ground truth template BCNs. For the 2nd layer, which should have fewer high-level features, the number of components should be less than in the 1st layer. Lv et al. (2015) introduce an experimental method to search for the best number of components, but, in fact, heuristically tuning the hyperparameters of deep models is very difficult. Since Deep MF is capable of estimating all these hyperparameters automatically, only the maximum number of iterations is given. For the other three methods, the hyperparameter values were chosen heuristically based on best matching to the ground truth templates for the 1st layer and for perceived quality of the derived networks for higher layers.

Table 4. Important Hyperparameter Settings of Four Deep Linear Models

	Deep MF	Deep SDL	Deep FICA	Deep NMF
Number of Components of 1st layer	N/A	15	20	30
Number of Components of 2nd layer	N/A	13	10	15
Number of Iterations	100	100	100	100
Step Length	N/A	0.01	N/A	N/A

4.2 Investigating the First Layer Reconstructions of Each Deep Linear Model via Intensity, Spatial and Hausdorff Distances

We can quantitatively compare the identified components, i.e., BCNs, with the original ground truth, i.e., templates, in three distinct ways. First, the similarity can be calculated spatially, largely independent of the intensity of each voxel of the identified components. The definition of spatial similarity is:

$$SimilaritySpatial = \frac{|Component \cap Template|}{|Component \cup Template|} \qquad (18)$$

where |·| represents binarization, which represent the voxels above a given intensity threshold. The spatial similarity is measuring the ratio of intersection and union of identified component and template.

In contradistinction, only considering the intensity of each voxel of the derived components, it is useful to calculate the distance between the intensities of components and templates. The definition of intensity similarity is:

$$Similarity_{Intensity} = \left(\sum_{i=1}^{N} \frac{|x_i - y_i|}{|x_i| + |y_i|}\right)^{-1} \qquad (19)$$

where |·| represents the absolute value. Given a threshold, the intensity similarity is calculated via summed absolute value of intensity of component (denoted as x_i) and template (denoted as y_i) divided by the absolute value of their difference. N denotes the total number of voxels. Obviously, if all intensity values of identified component and template are equal, the intensity similarity approaches infinity.

Finally, to jointly consider both spatial and intensity matching, we use the Hausdorff Distance (HD):

$$X = \sum_{i=1}^{N} 2 \times min(x_i, y_i)\, , if\ x_i,\ y_i \in C \cap T$$
$$Y = \sum_{i=1}^{M} x_i + y_i\, ,\ if x_i,\ y_i \in C \cup T \qquad (20)$$
$$HD = \frac{X}{Y}$$

Briefly, X represents two times the minimum intensity value of intersection of component and template; and Y represents the summed intensity value of union of component and template. C and T represent the sets of components and templates, respectively. Therefore, HD includes the influences of intensity similarity and spatial overlap simultaneously.

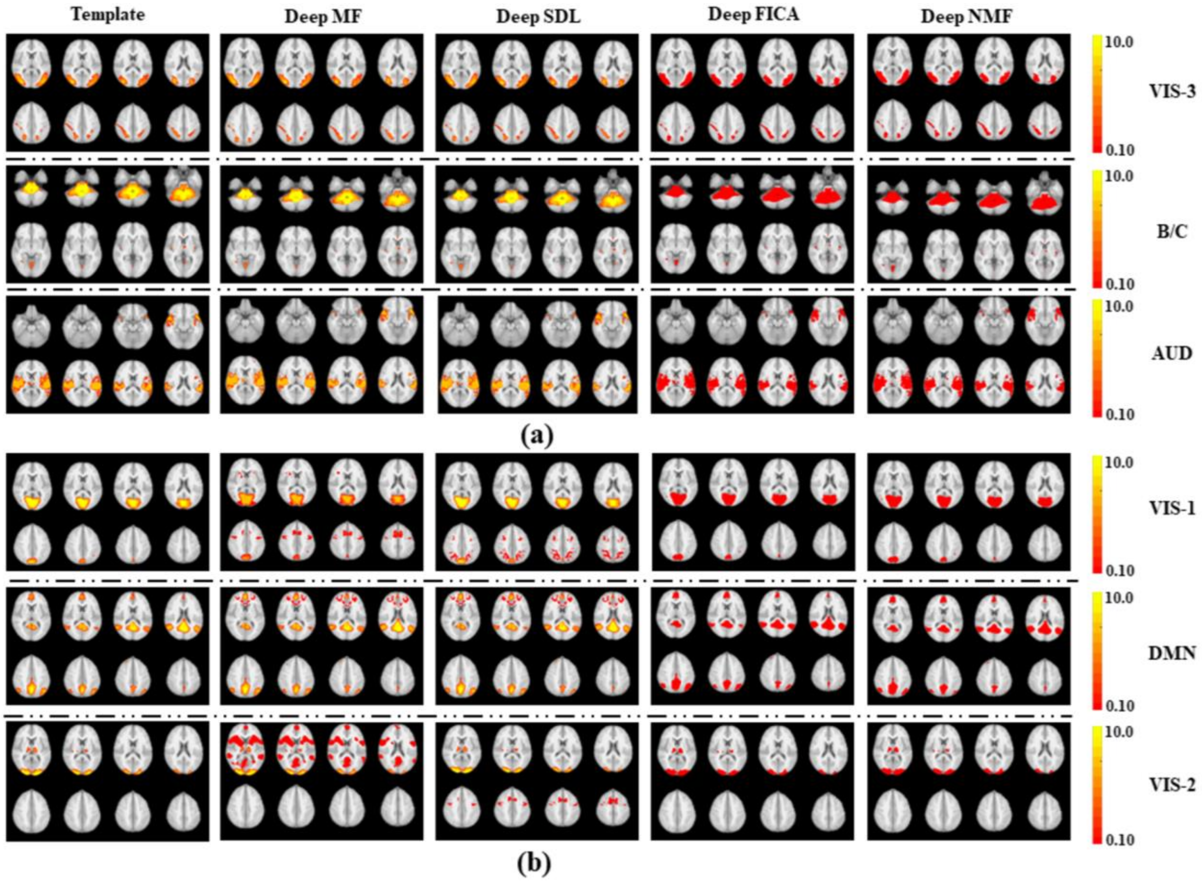


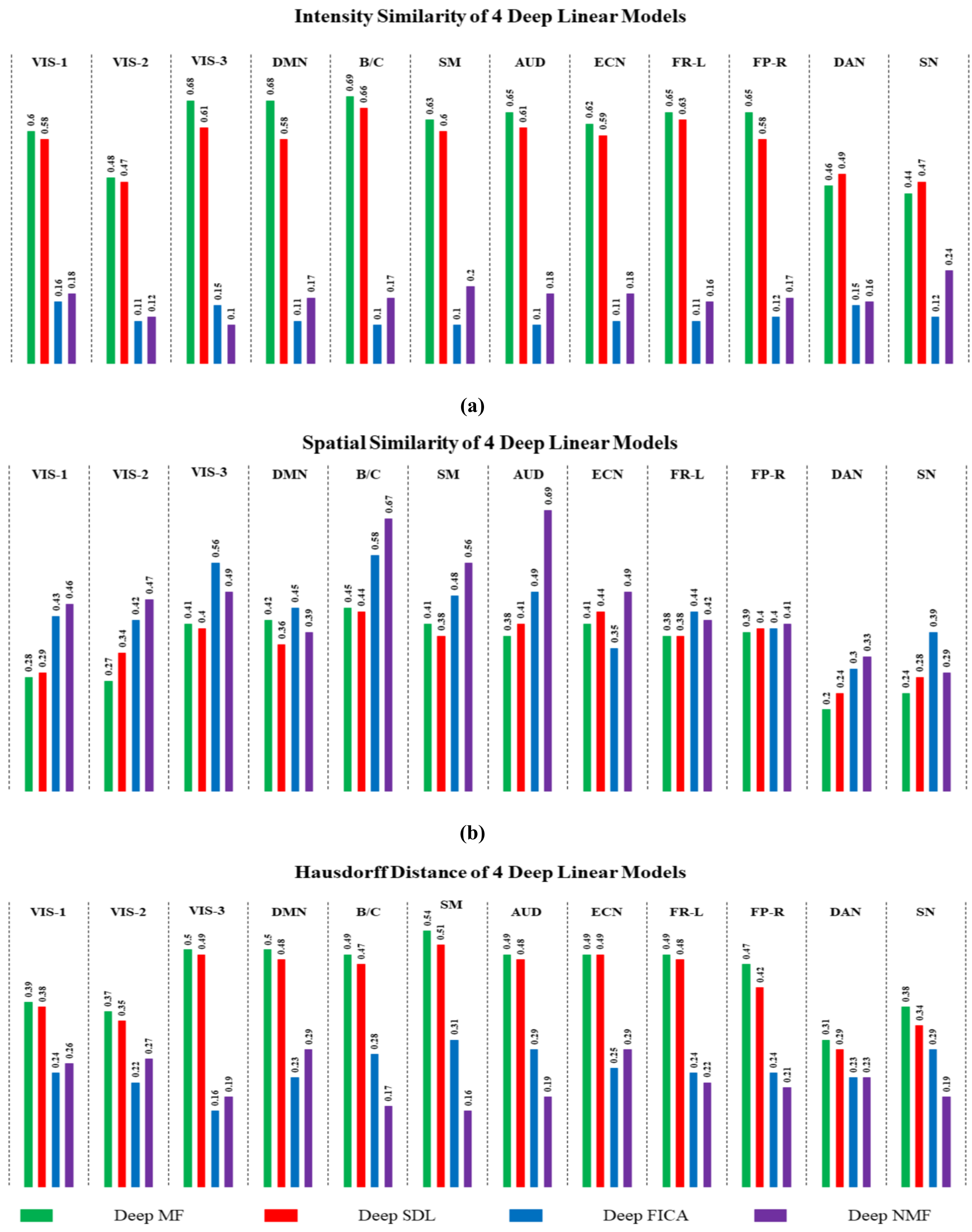
Figure 2. Comparison of six 1st layer networks from all four deep linear models with the ground truth templates from simulated fMRI data (see Table 3 for network abbreviations). The first column presents eight representative slices from each of six representative template networks. The second to fifth columns show the corresponding slices from the networks identified via Deep MF, Deep SDL, Deep FICA and Deep NMF, respectively. **(a)** The AUD, B/C and VIS-3 networks illustrate better intensity matching to the templates by Deep MF and Deep SDL than by Deep FICA or Deep NMF (see color bar of intensities measuring connectivity strength on the right). **(b)** The VIS-1, VIS-2 and DMN networks also show this same disparity among the deep linear models for intensity matching, but also show better spatial similarity to the templates for Deep FICA and Deep NMF compared to Deep MF or Deep SDL.

The results show that Deep NMF and Deep FICA produce smaller network intensities than the templates, whereas Deep MF and Deep SDL yield larger intensities that better match the templates (Figure 2a). In contradistinction, there are generally more noisy areas detected from Deep MF and Deep SDL due to the larger norms of their iterative operators, compared to Deep

NMF and Deep FICA (Figure 2b). Hence, Deep NMF and Deep FICA have better spatial similarity to the templates than the other two methods. This illustrates the trade-off between intensity matching and spatial matching. To view reconstructions for all 12 examined BCNs and for further details, please see Figure S1 included in the Supplemental Materials.

As defined by Eq. (18) to Eq. (20) in Section 4.2, the quantitative comparisons among the four deep linear models for intensity similarity, spatial similarity and the Hausdorff distance are provided by Figure 3. These quantitative results clearly demonstrate that Deep MF and Deep SDL provide the best intensity matching (Figure 3a), since their convergence velocity is relatively slow. Therefore, Deep MF and Deep SDL can reconstruct the most accurate connectivity strengths of each component from input fMRI signals, consistent with theory (Section 3). Considering spatial similarity, due to the fastest convergence velocity and non-negative normalization of Deep NMF, the intensity is reduced rapidly across iterations. Since the noise has smaller intensity than the signal, it is reduced much faster, which helps account for Deep NMF yielding the best spatial similarity results for most networks (Figure 3b). This result is also predicted by theory in Section 3. A rigorous proof is presented in the Appendices. It should be noted that Deep FICA has an inherent advantage over the other models for spatial matching since it is most similar to the shallow ICA analysis used to generate the ground truth templates for the BCNs.

All proposed deep linear models can be evaluated by HD to consider both intensity and spatial similarity (Figure 3c). Deep MF generated the best performance for all 12 BCNs with Deep SDL running close behind. Hence, the additional RRO in Deep MF does yield advantages over the other three deep linear models. Similarly, the sparsity operator of Deep SDL and Deep MF help them outperform Deep FICA and Deep NMF, which both lack that capability.



Intensity Similarity of 4 Deep Linear Models

(a)

Spatial Similarity of 4 Deep Linear Models

(b)

Hausdorff Distance of 4 Deep Linear Models

(c)

Figure 3. Comparisons of the twelve 1st layer networks of the four deep linear models for **(a)** intensity similarity to the ground truth templates; **(b)** spatial similarity to the ground truth templates; and **(c)** the Hausdorff Distance to the ground truth templates that jointly considers intensity and spatial similarity.

4.3 The 2nd Layer Networks of Each Deep Linear Model

Compared with the shallow 1st layer features, it is difficult to successfully investigate the features of deeper layers because there is no widely accepted ground truth for those more complex higher-level networks. The 2nd layer features can be comprehended as the recombination of 1st layer features. Another challenge for testing deeper networks is that the deep linear models differ with regard to how many layers can be reconstructed from a given dataset. For example, Deep FICA can only decompose the simulated fMRI into two layers, but Deep MF can decompose the simulated signal into four layers. Given these constraints, as well as the altered connectivity strengths in the 2nd layer relative to the 1st layer, we limit the analysis of deeper networks to examining the spatial similarity between 2nd layer networks of each of the four deep linear models with the shallow ground truth templates (Figure 4).

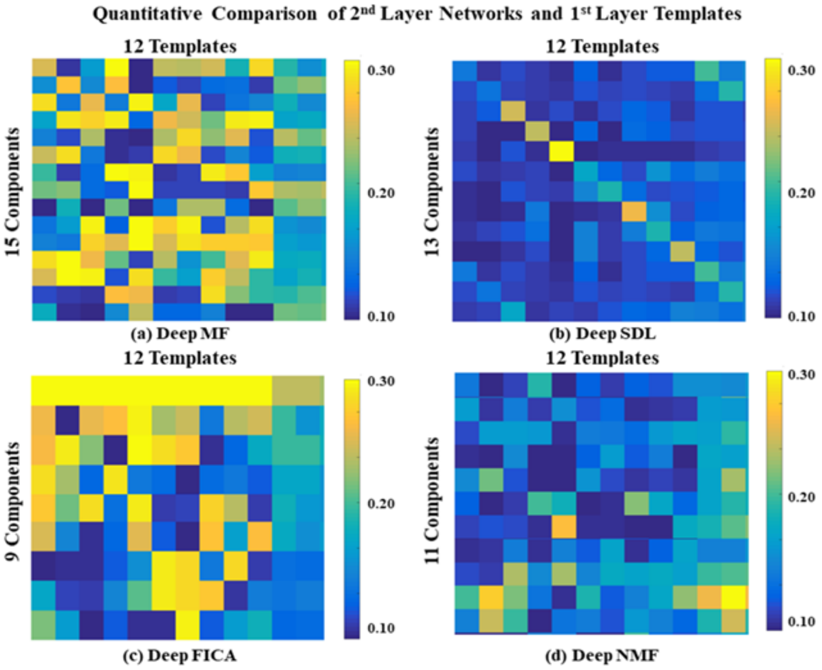


Figure 4. Comparisons of BCNs from 2nd layer of Deep MF, Deep SDL, Deep FICA and Deep NMF. Each element represents the spatial similarity of the identified component and the ground truth templates; (a), (b), (c) and (d) are Deep MF, Deep SDL, Deep FICA and Deep NMF, respectively. The rows represent the identified 2nd layer BCNs and the columns represent the ground truth templates of the simulated experiment.

Based on these preliminary comparisons, it is clear that the four deep linear models can produce different higher-level features. Most notable is that Deep SDL produces 2nd layer networks that are the most spatially similar to the shallow ground truth templates, as shown by the larger main diagonal elements in its similarity matrix and the smaller off-diagonal elements (Figure 4b). Hence, Deep SDL does relatively little recombination of the 1st layer features in its 2nd layer. In contradistinction, the first component of the Deep FICA 2nd layer (top row of its similarity matrix in Figure 4c) is very strongly correlated with 10 of the 12 ground truth templates and therefore appears to be a spatially "global" network. The 2nd layer features of Deep NMF have overall the least spatial similarity with the ground truth templates (Figure 4d) whereas Deep MF produces the greatest variation in the correlations between its 2nd layer features and the ground truth templates (Figure 4a).

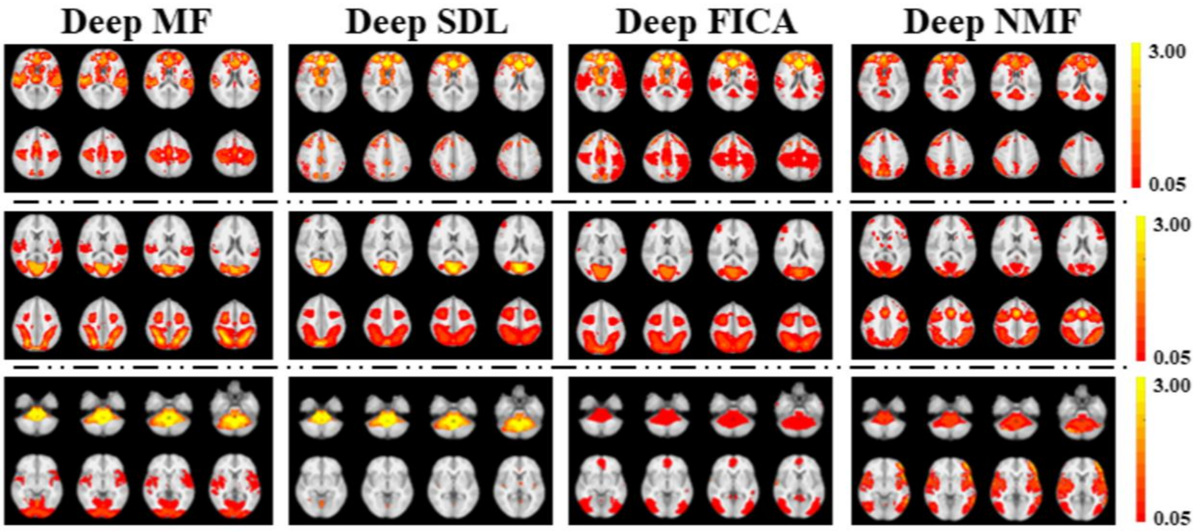


Figure 5. Comparisons of BCNs derived from the 2nd layer of Deep MF, Deep SDL, Deep FICA and Deep NMF. Each column includes three representative 2nd layer networks from a deep linear model, matched across models in each row.

Three representative 2nd layer BCNs matched for each deep linear model are presented in Figure 5. The full set of non-noise 2nd layer networks are given in Figure S2 of the Supplemental Materials. The top row of Figure 5 shows that the nodes of the ECN, including anterior cingulate cortex and medial prefrontal cortex, are represented in that 2nd layer network for all four models. For both Deep MF and Deep FICA, the ECN is combined with nodes of the SN, including the insulae, pre-supplementary motor areas (pre-SMA), and premotor areas.

For Deep NMF and Deep SDL, however, the ECN is joined instead with nodes of the DMN, including the precuneus, posterior cingulate cortex and the superior parietal lobules. This higher-level connectivity between ECN and DMN is weaker for Deep SDL than Deep NMF, whereas connectivity within ECN is stronger for Deep SDL than Deep NMF, in keeping with the observations from Figure 4 that Deep SDL preserves 1st layer networks the most of all four algorithms, whereas Deep NMF preserves 1st layer networks the least (Figures 4 & S2). It can be observed that the same 2nd layer network of Deep FICA also contains parts of the DMN, most notably the posterior cingulate cortex, although to a lesser extent than Deep NMF. Therefore, Deep FICA recombines nodes of three different 1st layer spatially independent components (DMN, ECN & SN) into a single 2nd layer independent component. Figure S3 of the Supplementary Materials provides a spatial similarity matrix for the non-noise 2nd layer networks for Deep SDL, Deep FICA and Deep NMF with reference to Deep MF.

The middle row of Figure 5 shows that all four models produced a 2nd layer network consisting of VIS-1 and DAN. However, Deep NMF additionally included parts of VIS-2, VIS-3 and pre-SMA whereas Deep MF additionally included VIS-3 and posterior perisylvian regions. The bottom row of Figure 5 illustrates links of the brainstem and cerebellum with visual cortex. However, Deep MF finds correlations of B/C with VIS-1 & VIS-2 whereas Deep FICA and Deep NMF finds correlations with VIS-3 instead. Both Deep MF and Deep NMF include perisylvian regions in this 2nd layer network as well. Given the relatively slow convergence of Deep SDL via gradient descent, these features might be expected in its 3rd layer instead.

We also compare computation time for the four deep linear models presented in this work on our computing cluster (Figure 6), which demonstrates that Deep FICA is the fastest and Deep SDL is by far the slowest. Deep MF provides the best trade-off between speed and performance as judged by reconstruction accuracy for 1st layer networks (Figure 3).

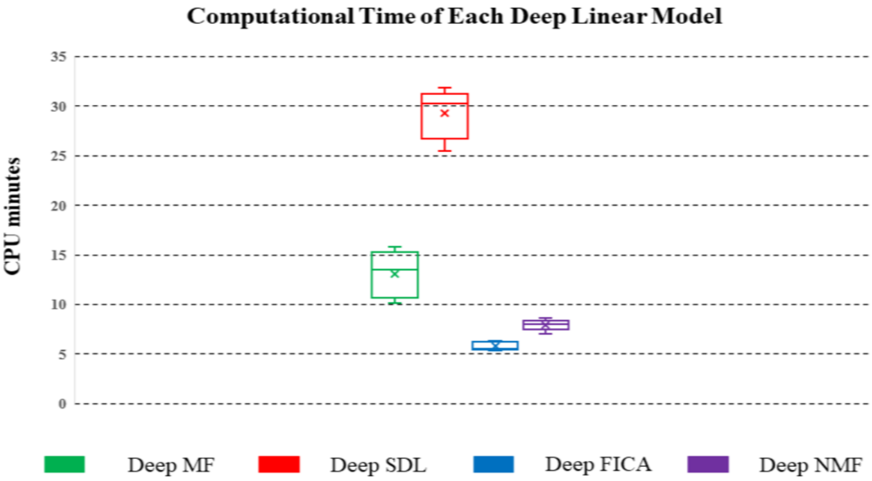


Figure 6. Comparisons of computation time of 10 independent runs, using the same number of iterations and the same simulated fMRI dataset for Deep FICA (blue), Deep SDL (red), Deep MF (green) and Deep NMF (purple). The box plots give the mean and standard deviation of the CPU time in minutes for the 10 runs.

5. Discussion

We have introduced novel deep linear models that integrate multiple operators to extract hierarchical spatial features in fMRI data. These models bridge the gap between traditional shallow linear models (Andersen et al., 1999; Beckmann et al., 2005; Calhoun et al., 2001; Hyvarinen, 1999; Lee et al., 2011; Lee et al., 2016; Mairal et al., 2010; McKeown & Sejnowski, 1999) and newer deep nonlinear models (Hu et al., 2018; Huang et al., 2018; Dong et al., 2020; Zhang et al., 2020). The primary advantages of the proposed algorithms over more complex deep nonlinear models are to quickly and easily map the hierarchical organization of BCNs without requiring large amounts of fMRI data or HPC clusters with GPUs or TPUs. The behavior of deep linear models is also more explainable than are, for example, CNNs and DBNs, as we show through theoretical predictions of their relative performance (Section 3) that are validated via simulations (Section 4). Furthermore, convergence to the global optimum can be guaranteed for deep linear models with convex optimization functions, unlike deep nonlinear models where such convergence is rarely achieved in practice. This is important given the recent realization that real-world imaging applications often suffer from underspecification, resulting in wildly unpredictable performance from any particular deep nonlinear network due to convergence to different local optima from different random initial conditions despite identical training data and hyperparameters (D’Amour et al., 2020).

Deep MF employs ADMM, which is a distributed optimization algorithm particularly well suited to compositional analysis of hierarchical modular systems, and also utilizes RRO for data-driven determination of all hyperparameters, which can be considered an intelligent factorization method. This is a major advantage over many conventional shallow data-driven fMRI connectivity reconstruction methods and the other three deep linear models presented here, as well as more complex deep nonlinear models, all of which must be manually tuned for hyperparameter settings. Deep SDL can explore more potential components than other models, even more than the number of original time points, via over-complete decomposition. Deep NMF converges very rapidly and recognizes the non-negative constraints of BCNs in fMRI. Finally, Deep FICA efficiently maps hierarchically spatially independent BCNs and is even easier to implement than the other peer deep linear models, especially given the wide usage of shallow ICA models for unsupervised fMRI mapping.

In this research, we also introduce an innovative framework for studying the relative performance of deep linear models, both theoretically by comparing their mathematical structure as well as *in silico* via fMRI simulations. Evaluating the 1st layer reconstructions of these deep linear models using simulated fMRI data from widely accepted ground truth BCNs, we find that Deep MF and Deep SDL are clearly superior for computing connectivity strength whereas Deep NMF and Deep FICA are modestly better for mapping spatial extent. These results were predicted from the unique mix of mathematical operators used in each of the four methods (Section 3). Overall, Deep MF provided the most robust combination of intensity matching, spatial matching and computational efficiency of all four techniques. This can be attributed to its joint use of sparsity and rank reduction operators in conjunction with the distributed ADMM optimization function. We also discovered that deeper features such as the 2nd layer BCNs are recombinations of the 1st layer networks and that these can vary among the four deep linear models. For example, Deep SDL produces the least recombination of the 1st layer networks in its 2nd layer. This can be attributed to its relatively slow convergence velocity using gradient descent optimization; therefore, more low-level network recombination is seen in its 3rd layer instead. Another important factor that differs among the deep linear models is the number of spatial features that can be accommodated at each level of the hierarchy and the maximum number of layers for any given dataset. For example, Deep FICA supports the fewest number of meaningful components at the 2nd layer (nine) and only two layers total; therefore, its 2nd layer networks would be the least sparse. This can be seen in the top row of Figure 5 in which Deep FICA combines nodes of DMN, ECN and SN into a single 2nd layer network, unlike the other models that only incorporate two of the three 1st layer networks, but which can instead generate even deeper networks beyond the 2nd layer. Hence, the choice of deep linear model matters for exploring higher-level BCNs. The mathematical evaluation framework and the fMRI simulation procedure provided in this work should enable further development of deep linear models that are optimized for different types of real-world applications in biomedical imaging, with Deep MF as the current best algorithm for fMRI hierarchical functional connectivity mapping.

One shortcoming of the current work is that the ground truth templates for testing the 1st layer networks were generated using conventional shallow ICA (Smith et al., 2009), which is currently the most widely accepted technique for data-driven analysis of functional connectivity. Aside from giving Deep FICA an inherent advantage, these spatially independent BCNs do not adequately evaluate the ability to reconstruct overlapping networks that is a property of methods such as shallow or Deep SDL. We also do not comprehensively investigate the properties of the deeper layers of these four models, which is an extensive topic that is beyond the scope of this paper, especially considering the absence of gold standards for these more complex high-level networks as well as the wide variation among deep linear models in key attributes such as convergence velocity and enforcement of sparsity.

In this initial exploratory work, many of the derived 2nd layer BCNs demonstrate neurobiological face validity. For example, the SN is known to modulate the anticorrelated connectivity of the DMN and the ECN (Menon & Toga, 2015); hence the linkage of their nodes into a single higher-level network (Figure 5, top row). The functional coupling of vision networks with the DAN shown in Figure 5 (middle row) is also well known, given the role that the DAN plays in visual attention and eye movements (Vossel et al., 2014). Future neuroscientific studies will be required to empirically validate the deep features of these models using demographic, clinical, cognitive, behavioral and/or electrophysiological data.

Since these deep linear models do not require large training datasets nor specialized computing infrastructure, they can be easily applied to clinical research with the potential to generate novel functional connectivity biomarkers of neurodevelopmental, neurodegenerative, and psychiatric disorders (Parkes et al., 2020), including for diagnosis, prognosis and treatment monitoring. This is particularly significant given the recent observation that neuropathology and psychopathology often affect low-level network connectivity differently than high-level network connectivity. For example, many different psychiatric disorders have been found to decrease lower-order sensory and somatomotor network connectivity in a uniform manner across patients (Elliott et al., 2018; Kebets et al., 2019), while increasing distinctiveness among patients in networks at higher levels of the hierarchy (Kauffman et al., 2017; Parkes et al., 2020). In fMRI studies of mild traumatic brain injury (TBI), altered functional connectivity has been found early after concussion both within individual BCNs, such as the SN, DMN and ECN, as well as between different BCNs (Palacios et al., 2017). Interactions of BCNs, such as that of the SN with the DMN, are thought to be especially important for outcome after TBI and can be used to guide personalized treatment (Jilka et al., 2014; Li et al., 2019). Disordered coupling of the SN with the DMN and ECN has also been shown in mild cognitive impairment (Chand et al., 2017). Hence, prevalent neurological disorders such as head trauma and neurodegenerative disease are thought to affect multiple levels of the human brain's hierarchical organization. Such high-level interactions between DMN, ECN and SN can be investigated with deeper layers of these hierarchical linear models that integrate their spatially distinct gray matter nodes into a single larger-scale network, as seen in Figure 5 (top row). These examples show how more principled data-driven characterization of this hierarchy, particularly at its higher levels, holds great promise for providing clinically actionable biomarkers of neurological and psychiatric diseases.

The benefits of deep linear models gain in importance as the spatial and temporal resolution and sensitivity of fMRI continue to increase with improved MR imaging hardware and pulse sequences, e.g., the advent of SLice Dithered Enhanced Resolution Simultaneous

MultiSlice (SLIDER-SMS) imaging (Vu et al., 2018) and MultiBand MultiEcho (MBME) imaging (Boyacioğlu et al., 2015; Cohen et al., 2020). Higher fMRI sensitivity and spatial resolution will enable mesoscale functional imaging that supports more 1st layer components of deep linear models to uncover subnetworks of the BCN templates used in this work. This will also permit the use of deeper models to extract more levels of the hierarchy of functional connectivity. Whereas many widely used methods for performing time-varying fMRI analysis are heuristic rather than data-driven, such as those with arbitrary time windows (Iraji et al., 2020), advances in fMRI temporal resolution can be combined with deep linear models that perform joint spatiotemporal decomposition for principled unsupervised dynamic functional connectivity mapping that reveals ever more of the human brain's hierarchical organization.

6. Acknowledgements

This work was supported by the U.S. National Institutes of Health [R01 MH116950, U01 EB025162] and U.S. Department of Defense [W81XWH-14-2-0176].

References

Andersen, A.H., Gash, D.M., Avison, M.J. (1999). Principal component analysis of the dynamic response measured by fMRI: a generalized linear systems framework. *Magnetic Resonance Imaging*, 17:795-815.

Bartels, A., Zeki, S. (2005). Brain dynamics during natural viewing conditions - a new guide for mapping connectivity in vivo. *Neuroimage*, 24:339-349.

Bassett, D. S., Bullmore, E., Verchinski, B. A., Mattay, V. S., Weinberger, D. R., & Meyer-Lindenberg, A. (2008). Hierarchical organization of human cortical networks in health and schizophrenia. *Journal of Neuroscience*, 28:9239-9248.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*:183-202.

Beckmann, C.F., Smith, S.M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *Neuroimage*, 25:294-311.

Bengio, Y., Courville, A.C., Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. CoRR, abs/1206.5538, 1.

Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med*, 34(4):537-41.

Biswal, B.B., Maarten, M., Xi-Nian, Z., Suril, G., Clare, K., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Stan, C. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107:4734-4739.

Boyacioğlu, R., Schulz, J., Koopmans, P. J., Barth, M., Norris, D. G. (2015). Improved sensitivity and specificity for resting state and task fMRI with multiband multi-echo EPI compared to multi-echo EPI at 7 T. *Neuroimage*, 119:352-361.

Bullmore, E., Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10:186-198.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14:140–151.

Chand, G. B., Wu, J., Hajjar, I., Qiu, D. (2017). Interactions of the Salience Network and Its

Subsystems with the Default-Mode and the Central-Executive Networks in Normal Aging and Mild Cognitive Impairment. *Brain Connect*, 7:401-412.

Cohen, A. D., Yang, B., Fernandez, B., Banerjee, S., Wang, Y. (2020). Improved resting state functional connectivity sensitivity and reproducibility using a multiband multi-echo acquisition. *Neuroimage*, 225:117461.

D'Amour, A., Heller, K., Moldovan, D., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv*:2011.03395v2

Dong, Q., Ge, F., Ning, Q., Zhao, Y., Lv, J., Huang, H., Yuan, J., Jiang, X., Shen, D., Liu, T. (2020). Modeling Hierarchical Brain Networks via Volumetric Sparse Deep Belief Network. IEEE Trans Biomed Eng, 67:1739-1748.

Dummit, D. S., & Foote, R. M. (2004). *Abstract algebra* (Vol. 3). Hoboken: Wiley.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behavior. *Trends in Cognitive Sciences*, 14:172-179.

Elliott, M. L., Romer, A., Knodt, A. R., Hariri, A. R. (2018). A connectome-wide functional signature of transdiagnostic risk for mental illness. *Biol Psychiatry*, 84:452-459.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25:24-29.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., ... & Bird, L. M. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25:60.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65.

Hinton, G.E., Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313:504-507.

Hinton, G.E., Osindero, S., Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527-1554.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29:82-97.

Hu, X., Huang, H., Peng, B., Han, J., Liu, N., Lv, J., ... & Liu, T. (2018). Latent source mining in FMRI via restricted Boltzmann machine. *Human Brain Mapping*, 39:2368-2380.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., ... & Liu, T. (2018). Modeling task fMRI data via deep convolutional autoencoder. *IEEE Transactions on Medical Imaging*, 37(7).

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10:626-634.

Iraji, A., Fu, Z., Damaraju, E., et al. (2019). Spatial dynamics within and between brain functional domains: A hierarchical approach to study time-varying brain function. *Hum Brain Mapp*, 40:1969-1986.

Iraji, A., Faghiri, A., Lewis, N., Fu, Z., Rachakonda, S., Calhoun, V. D. (2020). Tools of the trade: Estimating time-varying connectivity patterns from fMRI data. *Soc Cogn Affect Neurosci*, nsaa114. doi: 10.1093/scan/nsaa114.

Jilka, S.R., Scott, G., Ham, T., Pickering, A., Bonnelle, V., Braga, R. M., Leech, R., Sharp, D.J. (2014). Damage to the Salience Network and interactions with the Default Mode Network. *J Neurosci*, 34:10798-107807.

Kadison, R. V., & Ringrose, J. R. (1997). *Fundamentals of the theory of operator algebras* (Vol. 2). American Mathematical Soc..

Kaufmann, T., Alnæs, D., Doan, N. T,, Brandt, C. L., Andreassen, O. A,, Westlye, L.T. (2017). Delayed stabilization and individualization in connectome development are related to psychiatric disorders. *Nat Neurosci*, 20:513-515.

Kebets, V., Holmes A. J., Orban, C., Tang, S., Li, J., Sun, N., Kong, R., Poldrack, R. A., Yeo, B. T. T. (2019). Somatosensory-motor dysconnectivity spans multiple transdiagnostic dimensions of psychopathology. *Biol Psychiatry*, 86:779-791.

LeCun, Y., Bengio, Y., Hinton, G.E. (2015). Deep learning. *Nature*, 521:436-444.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788-791.

Lee, K., Tak, S., Ye, J.C. (2011). A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion. *IEEE Transactions on Medical Imaging*, 30:1076-1089.

Lee, Y.-B., Lee, J., Tak, S., Lee, K., Na, D.L., Seo, S.W., Jeong, Y., Ye, J.C., Initiative, A.s.D.N. (2016). Sparse SPM: Group Sparse-dictionary learning in SPM framework for resting-state functional connectivity MRI analysis. *Neuroimage*, 125:1032-1045.

Li, L. M., Violante, I. R., Zimmerman, K., Leech, R., Hampshire, A., Patel, M., Opitz, A., McArthur, D., Jolly, A., Carmichael, D. W., Sharp, D. J. (2019). Traumatic axonal injury influences the cognitive effect of non-invasive brain stimulation. *Brain*,142:3280-3293.

Liu, J., Yuan, L., & Ye, J. (2010). An efficient algorithm for a class of fused lasso problems. *In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 323-332). ACM.

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., Chen, H., Zhang, T., Hu, X., Han, J. (2015). Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *Biomedical Engineering, IEEE Transactions on*, 62:1120-1131.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1).

Mckeown, M. J., Sejnowski, T. J. (1998). Independent component analysis of fMRI data: Examining the assumptions. *Human Brain Mapping*, 6:368-372.

Menon, V., Toga, A. (2015). *Salience Network*. Elsevier. pp. 597–611. ISBN 978-0-12-397316-0.

Palacios, E. M., Yuh, E. L., Chang, Y. S., Yue, J. K., Schnyer, D. M., Okonkwo, D. O., Valadka, A. B., Gordon, W. A., Maas, A. I. R., Vassar, M., Manley, G. T., Mukherjee, P. (2017). Resting-State Functional Connectivity Alterations Associated with Six-Month Outcomes in Mild Traumatic Brain Injury. *J Neurotrauma*, 34:1546-1557.

Parkes, L., Satterthwaite, T. D,, Bassett, D. S. (2020). Towards precise resting-state fMRI biomarkers in psychiatry: synthesizing developments in transdiagnostic research, dimensional models of psychopathology, and normative neurodevelopment. *Curr Opin Neurobiol*, 65:120-128.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., ... &

Calhoun, V. D. (2014). Deep learning for neuroimaging: a validation study*. Frontiers in Neuroscience*, 8:229.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., … & Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron*, 72:665-678.

Royden, H. L. (1968). *Real analysis*. Krishna Prakashan Media.

Rudin, W. (1973). *Functional analysis*.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61:85-117.

Seo, J. D. (2018). “Deep” Independent Component Analysis in Tensorflow. https://towardsdatascience.com/deep-independent-component-analysis-in-tensorflow-manual-back-prop-in-tf-94602a08b13f

Shen, Y., Wen, Z., & Zhang, Y. (2014). Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods and Software*, 29:239-263.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., ... & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106:13040-13045.

Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9), 418-425.

Stam, C. J. (2014). Modern network science of neurological disorders*. Nature Reviews Neuroscience*, 15:683.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A.s.D.N. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage*, 101:569-582.

Suk, H.-I., Wee, C.-Y., Lee, S.-W., Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *Neuroimage*, 129:292-307.

Trigeorgis, G., Bousmalis, K., Zafeiriou, S., & Schuller, B. W. (2016). A deep matrix factorization method for learning attribute representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:417-429.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44.

Vossel, S., Geng, J. J., Fink, G. R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *Neuroscientist*, 20:150-159.

Vu, A. T., Beckett, A., Setsompop, K., Feinberg, D. A. (2018). Evaluation of SLIce Dithered Enhanced Resolution Simultaneous MultiSlice (SLIDER-SMS) for human fMRI. *Neuroimage*, 164:164-171.

Wen, Z., Yin, W., & Zhang, Y. (2012). Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4:333-361.

Zhang, W., Jiang, X., Zhang, S., Howell, B. R., Zhao, Y., Zhang, T., ... & Liu, T. (2017). Connectome-scale functional intrinsic connectivity networks in macaques. *Neuroscience*, *364*:1-14.

Zhang, W., Lv, J., Zhang, S., Zhao, Y., & Liu, T. (2018). Modeling resting state fMRI data via longitudinal supervised stochastic coordinate coding. *In Biomedical Imaging (ISBI 2018), IEEE 15th International Symposium on* (pp. 127-131). IEEE.

Zhang, W., Lv, J., Li, X., Zhu, D., Jiang, X., Zhang, S., ... & Liu, T. (2019). Experimental Comparisons of Sparse Dictionary Learning and Independent Component Analysis for Brain Network Inference from fMRI Data, *IEEE Transactions on Biomedical Engineering*, 66:289-299.

Zhang, W., Zhao, S., Hu, X., Dong, Q., Huang, H., Zhang, S., ... & Liu, T. (2020). Hierarchical Organization of Functional Brain Networks Revealed by Hybrid Spatiotemporal Deep Learning. *Brain Connectivity*, 10:72-82.

Appendix A

Assumption 1.1 For any operator discussed in this study, we have: $\forall \mathcal{C} \in \mathfrak{C}$, $\mathcal{C}: \mathbb{R}^{S\times T} \to \mathbb{R}^{S\times T}$. This assumption demonstrates that all operators are mapping from the finite dimensional space to another finite dimensional space, which is also reasonable in the real world.

Lemma 1.1 **(Norm Equality)** Given any arbitrary norm $\|\cdot\|$ and/or their finite linear combination $\sum_{i=1}^{n} k_i\|\cdot\|$ denoted based on any finite set, this norm or their finite linear combination is equivalent to ℓ_2 norm (e.g., $\|\cdot\|_2$).

Proof: We denote ℓ_1 and ℓ_2 norm in finite dimensions, such as:

$$\|a\|_1 = \sum_{i=1}^{n} |a_i|$$

$$\|a\|_2 = (\sum_{i=1}^{n} a_i^2)^{\frac{1}{2}} \tag{A.1}$$

$$a = [a_1, a_2, \cdots, a_n]$$

Obviously, since all norms are non-negative, according to Eq. (A.1), we have:

$$\sum_{i=1}^{n} a_i^2 \leq (\sum_{i=1}^{n} |a_i|)^2 \tag{A.2}$$

Eq. (A.2) implies:

$$\|a\|_2 \leq \|a\|_1 \tag{A.3}$$

And, based on Cauchy-Schwarz inequality, we have:

$$\|a\|_1^2 = (\sum_{i=1}^{n} |a_i| \cdot 1)^2 \leq \sum_{i=1}^{n} a_i^2 \cdot \sum_{j=1}^{n} 1^2 = \|a\|_2^2 \cdot n \tag{A.4}$$

It implies:

$$\frac{1}{\sqrt{n}}\|a\|_1 \leq \|a\|_2 \tag{A.5}$$

According to the theorem of norm equality (Rudin, 1973), given an arbitrary finite dimensional space, if and only if the following inequality holds:

$$c\|\cdot\|_2 \leq \|\cdot\| \leq C\|\cdot\|_2 \tag{A.6}$$

Thus, the norm $\|\cdot\|$ is equivalent to $\|\cdot\|_2$. Since Eq. (B.3) and Eq. (B.4) hold, we have:

$$c\frac{1}{\sqrt{n}}\|a\|_1 \leq \|a\|_2 \leq \|a\|_1 \tag{A.7}$$

It implies $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent. Similarly, we can prove $\sum_{i=1}^{n} k_i\|\cdot\|$ is also equivalent to $\|\cdot\|_2$.

***Theorem 1.1* (Superiority of Deep Linear Models)** Given a real function $f(x)$ and $m(\{x \in [a,b]: |f(x)| = \pm\infty\}) = 0$. If considering the series of polynomials $\{P_n(x)\}_{n=1}^{N}$, we have: if N is large enough, we have: $\forall\varepsilon > 0$ $\left\|\{P_n(x)\}_{n=1}^{N} - f(x)\right\| \leq \varepsilon$; if $N \to \infty$, we have: $\lim_{N\to\infty}\{P_n(x)\}_{n=1}^{N} = f(x)$; however, for any shallow model, since N should be bounded, we only have: $\left\|\{P_n(x)\}_{n=1}^{N} - f(x)\right\| \leq M$.

Proof: According to Лузин (Luzin) Theorem (Royden, 1968), we have a close set:

$$\begin{gathered} F_n \subset F_{n+1} \subset \cdots \subseteq [a,b] \\ m([a,b]\backslash F_n) = \frac{1}{n} \\ f \in C(F_n) \end{gathered} \tag{A.8}$$

Then we have a consistent real function $g(x)$, and obviously we have:

$$g(x) = f(x) \tag{A.9}$$

Since for any continuous real function, we have:

$$|g(x) - P_n(x)| < \frac{1}{n} \tag{A.10}$$

Let $\mathcal{F} = \bigcup_{n=1}^{\infty} F_n$, and obviously we have:

$$m([a,b]\backslash\mathcal{F}) = 0 \tag{A.11}$$

If $\mathfrak{F}$ is a real function denoted on set $\mathcal{F}$, it indicates:

$$\lim_{N\to\infty}\left|\mathfrak{F} - \{P_n(x)\}_{n=1}^{N}\right| < \varepsilon \tag{A.12}$$

then we have $\lim_{N\to\infty}\{P_n(x)\}_{n=1}^{N} = \mathfrak{F}$,

Moreover, it is easy to prove $\{P_n\}_{n=1}^{\infty}$ denoted on $[a,b]$ as a ring $(\{P_n\}_{n=1}^{\infty}, +, \times)$ (Dummit, 2004; Kadison, 1997). And $m(\cdot)$ represents a Lebesgue measure.

***Lemma 1.2* (Contraction of Operators Combination)** Given two contraction mappings Φ_1 and Φ_2, we have the composite of two contraction mapping as $\Phi_2 \cdot \Phi_1$. The composite mapping $\Phi_2 \cdot \Phi_1$ must be contractive.

***Proof*:** According to the definition of contraction linear operator, we have:

$$\begin{gathered} \exists\zeta \in (0,1) \\ \rho \overset{\text{def}}{=} \|\Phi x - \Phi \mathrm{y}\| \\ \rho(\Phi x, \Phi y) \leq \zeta\rho(x,y) \end{gathered} \tag{A.13}$$

Obviously, and we have:

$$\rho(\Phi_1 u, \Phi_1 v) \le \zeta\rho(u,v) \;\; \forall\zeta \in (0,1) \\ \rho(\Phi_2 x, \Phi_2 y) \le \eta\rho(x,y) \;\; \forall\eta \in (0,1) \tag{A.14}$$

If we set:

$$x = \Phi_1 u, y = \Phi_1 v \tag{A.15}$$

the inequality below holds:

$$\rho(\Phi_2 x, \Phi_2 y) \le \eta\rho(\Phi_1 u, \Phi_1 v) \le \zeta\eta\rho(u,v) \tag{A.16}$$

Since the definition as

$$\forall\zeta,\eta \in (0,1),\; \rho(\Phi_2\Phi_1 u, \Phi_2\Phi_1 y) \le \zeta\eta\rho(u,v) \tag{A.17}$$

***Corollary 1.1* (General Contraction Operator)** According to Lemma 1.2, if denote the operators $\{\Phi_i\}_{i=1}^{K}$, $\forall\Phi_i \;\; i \in \mathbb{N}$, $\Phi_i\colon \mathbb{R}^{S\times T} \to \mathbb{R}^{S\times T}$; considering any combination of operators: $\Phi_K \cdot \cdots \cdot \Phi_2 \cdot \Phi_1$, if at least a single operator Φ_i is contraction operator, and other operators are bounded, such as $\forall i \ne k \;\; \|\Phi_i\| \le M$. If and only if $\prod_{i=1}^{K}\|\Phi_i\| < 1$, the combination of operator series $\Phi_K \cdot \cdots \cdot \Phi_2 \cdot \Phi_1$ is a contraction operator.

Proof: Obviously, according to Lemma 1.2, use a series as $\{\zeta_i\}_{i=1}^{K}$ to replace $\zeta,\eta \in (0,1)$,

Obviously, we have:

$$\zeta_i \in (0,1) \;\; i \in \mathbb{N} \\ \rho(\Phi_K \cdot \cdots \cdot \Phi_2\Phi_1 u, \Phi_K \cdot \cdots \cdot \Phi_2\Phi_1 y) \le \zeta_K \cdot \cdots \zeta_2 \cdot \zeta_1 \cdot \rho(u,v) \tag{A.18}$$

Since $\zeta_K \cdot \cdots \zeta_2 \cdot \zeta_1 < 1$, we have proved this corollary.

***Corollary 1.2* (Iterative Contraction Operator)** According to Lemma 1.2, if denote the operators $\{\Phi_i\}_{i=1}^{K}$, $\forall\Phi_i \;\; i \in \mathbb{N}$, $\Phi_i\colon \mathbb{R}^{S\times T} \to \mathbb{R}^{S\times T}$; considering any combination of operators: $\Phi_K \cdot \cdots \cdot \Phi_2 \cdot \Phi_1$, if at least a single operator Φ_i is contraction operator, and other operators are bounded, such as $\forall i \ne k, \|\Phi_i\| \le M$. If and only if $\lim_{n\to\infty} \prod_{i=1}^{K}\|\Phi_i\|^n = c < 1$, the combination of operator series $\Phi_K^n \cdot \cdots \cdot \Phi_2^n \cdot \Phi_1^n$.

Proof: Obviously, according to Lemma 1.2 and Corollary 1.1 and 1.2, use a series as $\{\zeta_i\}_{i=1}^{K}$ to replace $\zeta,\eta \in (0,1)$,

And we have:

$$\forall\zeta_i \in (0,1) \;\; i \in \mathbb{N} \\ \rho(\Phi_K^n \cdot \cdots \cdot \Phi_2^n \cdot \Phi_1^n u, \Phi_K^n \cdot \cdots \cdot \Phi_2^n \cdot \Phi_1^n y) < \zeta_i^n \cdot \cdots \cdot \zeta_2^n \cdot \zeta_1^n \cdot \rho(u,v) \tag{A.19}$$

Since $0 < \zeta_i^n \cdot \cdots \cdot \zeta_2^n \cdot \zeta_1^n < 1$, we have proved this corollary.

Appendix B

Definition 2.1 If we denote Deep MF as an operator $\mathfrak{A}$, based on the description of Deep MF, considering the iteration k, we can denote $\mathfrak{A} \overset{\text{def}}{=} M \cdot \mathcal{A}^{\mathrm{k}} \cdot \mathcal{S}^{\mathrm{k}} \cdot \mathcal{R}^{\mathrm{k}}$.

Definition 2.2 If we denote Deep SDL as an operator $\mathfrak{L}$, based on the description of Deep SDL, considering the iteration k, we can denote $\mathfrak{L} \overset{\text{def}}{=} M \cdot \mathcal{G}^{\mathrm{k}} \cdot \mathcal{S}^{\mathrm{k}}$.

Definition 2.3 If we denote Deep FICA as an operator $\mathfrak{T}$, based on the description of Deep FICA, considering the iteration k, we can denote $\mathfrak{T} \overset{\text{def}}{=} \mathcal{P} \cdot \mathcal{F}^{\mathrm{k}}$.

Definition 2.4 If we denote Deep NMF as an operator $\mathfrak{N}$, based on the description of Deep NMF, considering the iteration k, we can denote $\mathfrak{N} \overset{\text{def}}{=} M \cdot \mathcal{U}^{\mathrm{k}} \cdot \mathcal{N}$.

Theorem 2.1 **(Contraction of ADMM Operator)** ADMM could be considered as contraction operator. It can be treated as a general iterative contraction operator in finite dimensionality space. We have ADMM $\overset{\text{def}}{=} \mathcal{A}$. If denote the $\|\mathcal{A}^{k+1}\| = \alpha\|\mathcal{A}^{k}\|$, and β should be step length, i.e., penalty parameter, if $n \to \infty$ $0 < (\alpha\beta)^n \|BN\| < 1$, $\mathcal{A}$ can be considered as a contraction operator. And $\|BN\|$ denotes the norm of different residual error, considering two distinctive input matrices.

Proof: X and Y, represent the two input matrices.

Consider the iterative format of ADMM as

$$\mathcal{A}_{\,k+1} \leftarrow \mathcal{A}_{\,k} - \min(f_{\mathcal{A}}) \tag{B.1}$$

And it also can imply:

$$\begin{gathered} \|\mathcal{A}_{\,k+1}\| = \alpha\|\mathcal{A}_{\,k}\|, \\ 0 < \alpha < 1 \end{gathered} \tag{B.2}$$

According to the definition of contraction operator, we have:

$$\begin{aligned} \|\mathcal{A}\mathrm{X} - \mathcal{A}\mathrm{Y}\| \le \alpha \Bigg\| \Bigg(& \beta(\ \mathrm{e}_k^t + \prod_{i=1}^{k-1} X_i\, Y_k + \sum_{i=1}^{k} Z_k^{t+1} - SG) - \alpha\beta(\widehat{\mathrm{e}}_k^t + \prod_{i=1}^{k-1} \hat{X}_i\, \hat{Y}_k \\ & + \sum_{i=1}^{k} \hat{Z}_k^{t+1} - \widehat{SG}) \Bigg) \Bigg\| \end{aligned} \tag{B.3}$$

And we also have:

$$\begin{aligned} \|\mathcal{A}\mathrm{X} - \mathcal{A}\mathrm{Y}\| \le \alpha\beta \Bigg\| & \ \mathrm{e}_k^t - \widehat{\mathrm{e}}_k^t + \prod_{i=1}^{k-1} X_i\, Y_k - \prod_{i=1}^{k-1} \hat{X}_i\, \hat{Y}_k + \sum_{i=1}^{k} Z_k^{t+1} - \sum_{i=1}^{k} \hat{Z}_k^{t+1} \\ & + \widehat{SG} - SG \Bigg\| \end{aligned} \tag{B.4}$$

Since $\mathrm{e}_k^t, \widehat{\mathrm{e}}_k^t, \prod_{i=1}^{k-1} X_i\, Y_k, \prod_{i=1}^{k-1} \hat{X}_i\, \hat{Y}_k, \sum_{i=1}^{k} Z_k^{t+1}, \sum_{i=1}^{k} \hat{Z}_k^{t+1}, \hat{S}, S \in \mathbb{R}^{m\times n}$, they are obviously bounded; and using Corollary 1.1 and 1.2, we have:

$$\|\mathcal{A}\mathrm{X} - \mathcal{A}\mathrm{Y}\| \le \alpha\beta\|BN\| \tag{B.5}$$

Obviously, it demonstrates:

$$\|\mathcal{A}^n\mathrm{A} - \mathcal{A}^n\mathrm{B}\| \le (\alpha\beta)^n\|BN\| < 1 \tag{B.6}$$

If and only if $0 < (\alpha\beta)^n < 1$, or $0 < \alpha\beta < 1$, $\mathfrak{A}^n$ is equivalent to a contraction operator. According to Lemma 1.2 and Corollary 1.1, 1.2, it also indicates: when *n* is large enough, $n > N$, we have:

$$\lim_{n\to\infty}\|\mathcal{A}^n\mathrm{A} - \mathcal{A}^n\mathrm{B}\| \le \lim_{n\to\infty}(\alpha\beta)^n\|BN\| \tag{B.7}$$

Obviously, if and only if $\lim_{n\to\infty}(\alpha\beta)^n\|BN\| < 1$, the iterative ADMM operator can be equivalent to a contraction operator.

***Theorem 2.2* (Initialization Operator is bounded)** If we denote the sparse operator as $\mathcal{M}$: $\mathbb{R}^{S\times T} \to \mathbb{R}^{S\times T}$, we have $\|\mathcal{M}\| < \infty$.

Proof: according to the definition of operator norm (Rudin 1973), $\|\mathcal{M}\| \le sup\frac{\|MX\|}{\|X\|}$; obviously, $\|\mathcal{M}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$X = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} \mathcal{M} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} \|X\| < \infty \ \|\mathcal{M}X\| < \infty \tag{B.8}$$

Obviously, $\|\mathcal{M}\| < \infty$.

***Theorem 2.3* (Sparsity Operator is bounded)** If we denote the sparse operator as $\mathcal{S}$: $\mathbb{R}^{S\times T} \to \mathbb{R}^{S\times T}$, we have $\|\mathcal{S}\| < \infty$.

Proof: according to the definition of operator norm (Rudin, 1973), $\|\mathcal{S}\| \le sup\frac{\|\mathcal{S}X\|}{\|X\|}$; obviously, $\|\mathcal{S}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$\mathcal{S}X = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ 0 \\ a_n \end{bmatrix} \ \mathcal{S}Y = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} \tag{B.9}$$

and we examine:

$$\mathcal{S}X-\mathcal{S}Y=\begin{bmatrix} a_1-b_1 \\ a_2 \\ \vdots \\ -\ \ b_{n-1} \\ a_n-b_n \end{bmatrix};\ \ X-Y=\begin{bmatrix} a_1-b_1 \\ a_2-b_2 \\ \vdots \\ a_{n-1}-b_{n-1} \\ a_n-b_n \end{bmatrix},\ \|\mathcal{S}X-\mathcal{S}Y\|\le s\ \ \|X-Y\|, \tag{B.10}$$

Without loss of generality, and based on Lemma 1.2, we calculate the ℓ_2 norm, and we have:

$$\infty > s \ge \frac{\sum_{i=u}^{n}(a_i-b_i)^2+\sum_{i=v}^{p}(a_i)^2+\sum_{i=w}^{t}(b_i)^2}{\sum_{i=1}^{n}(a_i-b_i)^2} \tag{B.11}$$

This inequality demonstrates that $\|\mathcal{S}\|$ is a bounded. And $\mathcal{S}$ is a bounded operator.

***Theorem 2.4* (Rank Reduction Operator is bounded)** If we denote the sparse operator as $\mathcal{R}$: $\mathbb{R}^{S\times T}\to\mathbb{R}^{S\times T}$, we have $\|\mathcal{R}\|<\infty$.

Proof: According to the definition of operator norm (Rudin, 1973), $\|\mathcal{R}\|\le sup\frac{\|\mathcal{R}X\|}{\|X\|}$; obviously, $\|\mathcal{R}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$X=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix},\ \mathcal{R}X=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \\ \vdots \\ 0 \end{bmatrix} \tag{B.12}$$

Eq. (B.31) implies:

$$sup\frac{\|\mathcal{R}X\|}{\|X\|}=\frac{\sum_{i=1}^{n}a_i^2}{\sum_{i=u}^{p}(a_i-b_i)^2+\sum_{i=v}^{q}a_i^2}<\infty. \tag{B.13}$$

***Theorem 2.6* (Normalization Operator of Deep NMF is bounded)** If we denote the normalization operator of Deep NMF as $\mathcal{N}$: $\mathbb{R}^{S\times T}\to\mathbb{R}^{S\times T}$, we have $\|\mathcal{N}\|\le 1$.

Proof: according to the definition of operator norm (Rudin, 1973), $\|\mathcal{N}\|\le sup\frac{\|\mathcal{N}X\|}{\|X\|}$; obviously, $\|\mathcal{N}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$X=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix}\ \ \mathcal{N}X=\begin{bmatrix} b_1 \\ 0 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} \tag{B.14}$$

According to Eq. (B.33), we need to notice: $\{a_i\}_{i=1}^{K}\subseteq[-q,q]$, $1\le q<\infty$; $\{b_i\}_{i=1}^{K}\subseteq[0,1]$. Obviously, $\|\mathcal{N}X\|<\|X\|$. Finally, we have: $\|\mathcal{N}\|<1$.

Theorem 2.7 **(Contraction of Updating Operator Deep NMF)** If we denote the updating operator as $\mathcal{U}\colon\ \mathbb{R}^{S\times T}\rightarrow\mathbb{R}^{S\times T}$, we have $\|\mathcal{U}\|<1$.

Proof: according to the definition of operator norm (Rudin, 1973), $\|\mathcal{U}\|\leq sup\frac{\|\mathcal{U}X\|}{\|X\|}$; obviously, $\|\mathcal{U}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$X=\begin{bmatrix}a_1\\a_2\\\vdots\\a_{n-1}\\a_n\end{bmatrix}\quad\mathcal{U}X=\begin{bmatrix}b_1\\0\\\vdots\\b_{n-1}\\b_n\end{bmatrix}\tag{B.15}$$

According to the iterative format of Deep NMF, we need to notice: $b_i=\frac{a_i}{\max f(a_i)}$; obviously, $\|\mathcal{U}X\|<\|X\|$. Finally, we have: $\|\mathcal{U}\|<1$. Otherwise, if $\|\mathcal{U}\|>1$, when $k\rightarrow\infty$, we have: $\|\mathcal{U}X\|=\infty$.

Theorem 2.8 **(Contraction of GD Operator)** Gradient Descent (GD) is a bounded contraction operator, if and only if the derivative of target function is bounded:

$|f''(\varsigma)|<\frac{1}{\sigma}<\infty$, σ is the step length.

Proof: The standard iteration format is:

$$x_{k+1}=x_k-\sigma f'(x_k)\tag{B.16}$$

Using the definition of operator, we have:

$$\tau(x_k)=x_k-\sigma f'(x_k)\ \ \forall\sigma\in(0,1)\tag{B.17}$$

And we have:

$$\tau\|\tau X-\tau Y\|=\big\|(X-Y)-\sigma\big(f'(X)-f'(Y)\big)\big\|\tag{B.18}$$

Using Mean value theorem, we have:

$$\tau\|\tau X-\tau Y\|=|1-\sigma f''(\varsigma)|\|X-Y\|\tag{B.19}$$

According to the definition of contraction operator (Rudin, 1973), if and only if:

$$|1-\sigma f''(\varsigma)|<1,\ |1-\sigma f''(\varsigma)|\in\mathbb{K}\tag{B.20}$$

It also implies, when the following inequality holds:

$$|f''(\varsigma)|<\frac{1}{\sigma}<\infty\tag{B.21}$$

GD is considered as a contraction mapping/operator. Without generality, we can set $\sigma<\frac{1}{|f''(x)|+1}$.

And obviously, using multiplicative inequality, we have:

$$\|\tau X - \tau Y\| \leq \|\tau\|\|X - Y\| \tag{B.22}$$

Since X and Y both denote in finite ℓ^2 space, we have:

$$\|\tau\|\|X - Y\| \leq \infty \tag{B.23}$$

Using Uniformly bounded theorem, we have:

$$\|\tau\| \leq M,\ M \in \mathbb{K} \tag{B.24}$$

GD is a bounded mapping/operator.

According to Lemma 1.2, and Corollary 1.1-1.2, obviously, for n iterations for an operator, and if we set the accuracy level as ε, we have:

$$\|\tau^n X - \tau^{n+1} Y\| = \sigma^n \|X - \tau Y\| < \varepsilon \tag{B.25}$$

Since X and Y is both denoted in finite ℓ^2 space, we have:

$$\sigma^n \|X - \tau Y\| \leq \sigma^n (\|X\| + \|\tau Y\|) \tag{B.26}$$

Obviously, $\|X - Y\|_{\ell^2}$ is bounded, and we have:

$$\begin{gathered}\sigma^n (\|X\| + \|\tau Y\|) \leq \sigma^n (\|X\| + \|\tau\|\|Y\|) \leq \sigma^n (\|X\| + \|Y\|) \leq \sigma^n \cdot 2\|X\| \\ 0 < \sigma^n \cdot 2\|X\| < \varepsilon \\ n > log \frac{\varepsilon}{2\|X\|} \,/\, log\, \sigma > 0\end{gathered} \tag{B.27}$$

We provide the infimum of iteration as $log \frac{\varepsilon}{2\|X\|} \,/\, log\, \sigma$ to approach the accuracy level ε.

Theorem 2.9 **(Operator PCA is bounded)** If we denote the updating operator as $\mathcal{P}$: $\mathbb{R}^{S \times T} \to \mathbb{R}^{S \times T}$, we have $\|\mathcal{P}\| < \infty$.

Proof: According to the definition of operator norm (Rudin, 1973), $\|\mathcal{P}\| \leq sup \frac{\|\mathcal{P}X\|}{\|X\|}$; obviously, $\|\mathcal{U}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$X = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} \quad \mathcal{P}X = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-k} \\ \vdots \\ 0 \end{bmatrix} \tag{B.29}$$

According to the dimensional reduction of PCA, we have: $sup \frac{\|\mathcal{P}X\|}{\|X\|} = sup \frac{(\sum_{i=1}^{n-k} b_i^2)^{\frac{1}{2}}}{(\sum_{i=1}^{n} a_i^2)^{\frac{1}{2}}} < \infty$. It demonstrates: $\|\mathcal{P}\| < \infty$.

***Theorem 2.10* (Contraction of Fixed-Point Operator)** If we denote the updating operator as $\mathcal{F}\colon\ \mathbb{R}^{S\times T} \to \mathbb{R}^{S\times T}$, we have $\|\mathcal{F}\| < 1$.

Proof: according to the definition of operator norm (Rudin, 1973), $\|\mathcal{F}\| \leq sup\frac{\|\mathcal{F}X\|}{\|X\|}$; obviously, $\|\mathcal{U}X\|$ and $\|X\|$ is bounded, since both of norms are based on finite dimensional matrix. And if we denote:

$$X = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} \quad \mathcal{F}X = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} \tag{B.30}$$

According to the iterative format of Deep NMF, we need to notice: $b_i = \frac{a_i}{\sqrt{\|a_i\|c_i\|a_i^T\|}}$; obviously, $\|\mathcal{F}X\| < \|X\|$. Finally, we have: $\|\mathcal{F}\| < 1$. Otherwise, if $\|\mathcal{F}\| > 1$, when $k \to \infty$, we have: $\|\mathcal{F}X\| = \infty$.

***Theorem 2.11* (Inequality of Operator Norms)** According to Theorem 2.1, 2.7, 2.8 and 3.0, if we assume: $\|\mathcal{A}^{k+1}\| = \alpha_1\|\mathcal{A}^k\|$, $\|\mathcal{G}^{k+1}\| = \alpha_2\|\mathcal{G}^k\|$, $\|\mathcal{U}^{k+1}\| = \alpha_3\|\mathcal{U}^k\|$, $\|\mathcal{F}^{k+1}\| = \alpha_4\|\mathcal{F}^k\|$, we have: $\alpha_1 \neq \alpha_3,\ \alpha_4$; $\alpha_2 \neq \alpha_3,\ \alpha_4$;

Proof: Proof by contradiction. In general, we assume $\|\mathcal{A}\| = \|\mathcal{U}\|$, according to the iterative formats of Deep MF and Deep NMF, and considering:

$$\mathcal{A}_{\ k+1} \leftarrow \mathcal{A}_{\ k} - \min(f_{\mathcal{A}}) \tag{B.31}$$

If we employ the $\alpha\mathcal{A}_{\ k} = \mathcal{A}_{\ k+1}$ to replace $\mathcal{A}_{\ k+1}$:

$$\alpha\mathcal{A}_{\ k} = \mathcal{A}_{\ k} - \min(f_{\mathcal{A}}) \tag{B.32}$$

And we can reformat this equality as:

$$(1-\alpha)\mathcal{A}_{\ k} = \min(f_{\mathcal{A}}) \tag{B.33}$$

Considering the iterative format of Deep NMF:

$$\mathfrak{N}_{k+1} \leftarrow \mathfrak{N}_k / max(f_{\mathfrak{N}}) \tag{B.34}$$

Let we denote:

$$\left|\frac{1}{ma\,x(f_{\mathfrak{N}})}\right| \leq \varepsilon \tag{B.35}$$

And considering an extreme condition, $\forall\varepsilon_i \leq \varepsilon$, for each iteration i, and $\lim\limits_{i\to\infty} \varepsilon_i = \varepsilon$;

$$\lim_{i\to\infty}(1-\varepsilon_i)\mathcal{A}_{\ k} = \min(f_{\mathcal{A}}) \tag{B.36}$$

Then we have the conclusion:

$$\exists\ n \ll k, \mathcal{A}_{\ k} = \min(f_{\mathcal{A}}) \tag{B.37}$$

It demonstrates for the iterative format of Deep MF, before convergence, the iteration can be terminated, since a very small norm of operator $\mathcal{A}$. $\mathcal{A}$ cannot guarantee the convergence. It obviously disobeys the property of ADMM.

Similarly, we can also prove $\alpha_2 \neq \alpha_3$, α_4; and $\alpha_1 \neq \alpha_4$.

Appendix C

Assumption 3.1 For all operators, these operators should be considered as linear operators, and we have:

$$\Phi \cdot (\mathrm{X} + \mathrm{Y}) = \Phi \cdot \mathrm{X} + \Phi \cdot \mathrm{Y} \tag{C.1}$$

Assumption 3.2 For any input matrix, we can successfully separate the vital information and background noise. If we denote: $V = \{\bigcup_{i=1}^{P} voxel_i,\ voxel_i \in BN\}$, and $N = \{\bigcup_{i=1}^{Q} voxel_i,\ voxel_i \notin BN\}$. BN represents the functional areas, i.e., potentially activated areas of brain. We have some crucial assumptions: $V \cap N = \emptyset,\ V \geqslant 0,\ B \geqslant 0,\ \|V\| \gg \|N\|$.

***Lemma 3.1* (Continuous Operators)** For all operators analyzed in this study, if $k > K, \forall k \in \mathbb{N}$, these iterative operators can be considered as consistent operator. It means: if we have $\|V - \hat{V}\| \leq \varepsilon$, $\|\mathfrak{A}^k V - \mathfrak{A}^k \hat{V}\| \to 0$.

Proof: We denote: $\mathfrak{A},\ \mathfrak{L},\ \mathfrak{T},\ \mathfrak{N} \in \mathfrak{C}$: $\mathbb{R}^{s\times t} \to \mathbb{R}^{s\times t}$

For $V,\ \hat{V} \subseteq \mathbb{R}^{s\times t}$, we assume that:

$$\|V - \hat{V}\| \leq \frac{\varepsilon}{M} \tag{C.2}$$

If $k > K$, For any operator belongs to $\mathfrak{C}$ can be considered as a contraction operator, and we have:

$$\|\mathfrak{A}^k V - \mathfrak{A}^k \hat{V}\| \leq \|\mathfrak{A}^k\| \cdot \|V - \hat{V}\| \leq M \cdot \frac{\varepsilon}{M} = \varepsilon \tag{C.3}$$

This inequality demonstrates that all operators of $\mathfrak{C}$, if k is large enough, can be treated as the consistent operators (Rudin, 1973). Similarly, it also demonstrates: $\|\mathfrak{A}^k V - \mathfrak{L}^k V\| \leq \varepsilon$

***Theorem 3.1* (Distinctive Spatial Similarity)** If we denote the following set:

$$\begin{gathered}
Deep\ \ MF\!: D = \{\mathfrak{A}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\} \\
Deep\ \ SDL\!:\ L = \{\mathfrak{L}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\} \\
Deep\ \ FICA\!:\ I = \{\mathfrak{T}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \notin T\} \\
Deep\ \ NMF\!:\ \Theta = \{\mathfrak{N}^k N, N \in \bigcup_{i=1}^{M} voxel_i,\ voxel_i \to 0\}
\end{gathered} \tag{C.4}$$

And considering the iteration k, it implies:

$$\frac{|\mathfrak{A}^k V|}{|V \cup D|} \leq \frac{|\mathfrak{L}^k V|}{|V \cup L|} \leq \frac{|\mathfrak{T}^k V|}{|V \cup I|} \leq \frac{|\mathfrak{N}^k V|}{|V \cup \Theta|} \tag{C.5}$$

where $|\cdot|$ denotes the number of positive elements.

Proof: Based on assumptions 3.1 and 3.2, if $\forall\ k \in \mathbb{N}$, we have:

$$\begin{gathered}
\mathfrak{A}^k C = \mathfrak{A}^k V + (\mathfrak{A}^k N) \\
\mathfrak{L}^k C = \mathfrak{L}^k V + (\mathfrak{L}^k N) \\
\mathfrak{T}^k C = \mathfrak{T}^k V + (\mathfrak{T}^k N) \\
\mathfrak{N}^k C = \mathfrak{N}^k V + \Theta
\end{gathered} \tag{C.6}$$

According to Corollary 1.1 and 1.2, $k > K$, we have:

$$0 = \|\Theta\| \leq \|\mathfrak{T}^k N\| \leq \|\mathfrak{L}^k N\| \leq \|\mathfrak{A}^k N\| < \infty \tag{C.7}$$

We can also rewrite it as:

$$0 = |\Theta| < |I| \leq |L| \leq |D| \tag{C.8}$$

And, according to the spatial similarity, we also have:

$$\begin{aligned}
Deep\ MF_{Similarity} &\overset{\text{def}}{=} \frac{|(\mathfrak{A}^k \mathrm{V} \cup D) \cap V|}{|V \cup (\mathfrak{A}^k \mathrm{A} \cup D)|} = \frac{|\mathfrak{A}^k V|}{|V \cup D|} \\
Deep\ SDL_{Similarity} &\overset{\text{def}}{=} \frac{|(\mathfrak{L}^k V \cup L) \cap A|}{|V \cup (\mathfrak{L}^k \mathrm{V} \cup L)|} = \frac{|\mathfrak{L}^k \mathrm{A}|}{|V \cup L|} \\
Deep\ FICA_{Similarity} &\overset{\text{def}}{=} \frac{|(\mathfrak{T}^k V \cup I) \cap V|}{|V \cup (\mathfrak{T}^k \mathrm{A} \cup I)|} = \frac{|\mathfrak{T}^k V|}{|V \cup I|} \\
Deep\ NMF_{Similarity} &\overset{\text{def}}{=} \frac{|(\mathfrak{N}^k V \cup \Theta) \cap V|}{|V \cup (\mathfrak{N}^k \mathrm{A} \cup \Theta)|} = \frac{|\mathfrak{N}^k V|}{|V|}
\end{aligned} \tag{C.10}$$

Again, considering $k > K$, and Corollary 1.1 to 1.2, and Theorem 3.2, we have:

$$|\mathfrak{N}^k V| = |\mathfrak{T}^k V| = |\mathfrak{L}^k V| = |\mathfrak{A}^k V| \tag{C.11}$$

Obviously, we have:

$$0 < |V| = |V \cup \Theta| \leq |V \cup I| \leq |V \cup L| \leq |V \cup D| < \infty \tag{C.12}$$

Finally, the following inequality holds, such that:

$$0 < \frac{|\mathfrak{A}^k V|}{|V \cup D|} \leq \frac{|\mathfrak{L}^k V|}{|V \cup L|} \leq \frac{|\mathfrak{T}^k V|}{|V \cup I|} \leq \frac{|\mathfrak{N}^k V|}{|V \cup \Theta|} \tag{C.13}$$

***Theorem 3.2* (Bounded Iterative Operators)** For all operators analyzed in this study, if $k > K, \forall k \in \mathbb{N}$, these iterative operators can be considered as consistent operator. If we have: $\|\hat{V}\| \leq \varepsilon$, it means: $\|\mathfrak{N}^k V - \mathfrak{A}^k V\| \to 0$, $\|\mathfrak{L}^k V - \mathfrak{A}^k V\| \to 0$ and $\|\mathfrak{T}^k V - \mathfrak{A}^k V\| \to 0$.

Proof: We denote: $\mathfrak{A}, \mathfrak{L}, \mathfrak{T}, \mathfrak{N} \in \mathfrak{C}$: $\mathbb{R}^{s \times t} \to \mathbb{R}^{s \times t}$

If $k > K$, For any operator belongs to $\mathfrak{C}$ can be considered as a contraction operator, according to Lemma 3.1, and we have:

$$\begin{aligned}\left\|\mathfrak{N}^k V - \mathfrak{A}^k V\right\| &= \left\|\mathfrak{N}^k V - \mathfrak{N}^k \hat{V} + \mathfrak{N}^k \hat{V} - \mathfrak{A}^k \hat{V} + \mathfrak{A}^k \hat{V} - \mathfrak{A}^k V\right\| \\ &\leq \left\|\mathfrak{N}^k V - \mathfrak{N}^k \hat{V}\right\| + \left\|\mathfrak{N}^k \hat{V} - \mathfrak{A}^k \hat{V}\right\| + \left\|\mathfrak{A}^k \hat{V} - \mathfrak{A}^k V\right\|\end{aligned} \tag{C.14}$$

According to Lemma 3.1, and we have:

$$\begin{gathered}\left\|\mathfrak{N}^k V - \mathfrak{N}^k \hat{V}\right\| \leq \frac{\varepsilon}{3} \\ \left\|\mathfrak{A}^k \hat{V} - \mathfrak{A}^k V\right\| \leq \frac{\varepsilon}{3}\end{gathered} \tag{C.15}$$

Considering the inequality:

$$\left\|\mathfrak{N}^k \hat{V} - \mathfrak{A}^k \hat{V}\right\| \leq \left\|\mathfrak{N}^k - \mathfrak{A}^k\right\| \cdot \left\|\hat{V}\right\| \tag{C.16}$$

Obviously, $\left\|\mathfrak{N}^k - \mathfrak{A}^k\right\| \leq M$, and we choose $\left\|\hat{V}\right\| \leq \frac{\varepsilon}{3M}$; it implies:

$$\left\|\mathfrak{N}^k \hat{V} - \mathfrak{A}^k \hat{V}\right\| \leq \frac{\varepsilon}{3} \tag{C.17}$$

And we have:

$$\left\|\mathfrak{N}^k V - \mathfrak{A}^k V\right\| \leq \varepsilon \tag{C.18}$$