How to Integrate Data from Multiple Biological Layers in Mental Health?

8

Rogers F. Silva and Sergey M. Plis

8.1 Overview

The human brain is a massively parallel learning machine that contains multiple highly complex structurally and functionally overlapping subsystems, with processes occurring at different temporal and spatial scales, and interacting with every other bodily system through the peripheral nervous system. In order to gain a more complete understanding of its organization and function, information from various layers of this complex set of biological processes must be evaluated *simultaneously*, in a truly synergistic fashion.

To begin with, collecting such information *directly* often entails invasive procedures that are restricted to very narrow patient populations, such as with electrocorticography (ECoG) and deep brain electrodes. However, in order to be also able to study much broader healthy population baselines, it is necessary to pursue less invasive routes. Specifically, those enabled by means of indirect measurements from secondary biological processes such as cerebral blood flow and induced electromagnetic fields. While noninvasiveness often comes at the cost of blurring some of the true underlying neurological signals, the greater availability of subjects enables normative as well as comparative analyses, with far greater statistical power due to the substantially increased sample sizes. Furthermore, one must also be mindful of inherent sensor and device limitations dictating the temporal and spatial resolutions of the data, which ultimately yield only fragments of the measured processes, adding yet another layer of complexity to the data.

With these in mind, it is sensible to hereon broadly associate the term *biological layer* with different *imaging modalities*, i.e., the signal of some direct or indirect neurobiological process captured by a device. Common examples of

R. F. Silva · S. M. Plis (✉)
The Mind Research Network, Albuquerque, NM, USA
e-mail: splis@mrn.org

I. C. Passos et al., *Personalized Psychiatry*,
https://doi.org/10.1007/978-3-030-03553-2_8

such modalities include, but are not limited to, structural, functional, and diffusion weighted/spectrum magnetic resonance imaging (sMRI, fMRI, and DWI/DSI, respectively), electro- and magneto-encephalography (E/MEG), functional near-infrared spectroscopy (fNIRS), x-rays, computerized tomography (CT), positron emission tomography (PET), single-photon emission CT (SPECT), intracranial electrodes, genetic material information such as DNA microarrays, single nucleotide polymorphism and DNA methylation, as well as metabolomic and microbiome derivatives, etc. Demographic and behavioral information on individuals and populations of interest are also going to be considered modalities for the purposes of this chapter.

Under this broad definition, we will focus on the integration of biological layers by means of *direct joint analysis of all modalities* available. Joint analyses are those which simultaneously utilize data from all modalities in a synergistic way and, thus, can be categorized as *data fusion* approaches. A key requirement for these kinds of analyses is that the information contained in each modality have been collected *on the same subject* so that the data are naturally linked. For the same reason, whenever feasible, simultaneous measurements are also preferred over (and likely more informative than) measures from different sessions since that entails a stronger link between modalities.

The goal of integrating multiple biological layers is to identify the neurobiological processes underlying the measurements recorded in the data in order to understand their function, structure, and interaction. Ideally, we want to make predictions about these processes and be able to explain their causal mechanisms. Each biological layer is itself only a part of the underlying process. For example, blood flow picked up by fMRI and electrical activity of neurons registered by EEG are parts of the same process of neural activity. Only together—plus many other additional pieces of information, such as neural connectivity routes—they provide a complete picture of the underlying mechanism. Available *imaging modalities* provide a (partial) glimpse on many of the individual processes within a functioning brain. When any of them are used, we are dealing not only with the partial nature of the biological layers but also with the fact that each of the layers is measured with uncertainty that is different for each imaging modality. Fortunately, the uncertainty introduced by the employed imaging modality is often different for each biological layer and, optimistically, can cancel if the imaging modalities are properly combined. The difference in uncertainties is illustrated by MEG and fMRI, where the former has arguably greater spatial, while the latter has greater temporal uncertainty relative to the underlying process of neural activity. Given the insufficient nature of each modality, the only way we can build a complete understanding of the brain is by combining these complementary sources. Together, the limited views from each modality allow us to peer into the underlying biostructure. In summary, scientific discovery with data fusion should proceed in cycles: measuring different physical processes at various biological and temporal scales, synthesizing that information using specific methods, understanding the underlying processes identified, and repeating with the gained insights.

In the following sections, we will discuss two principled approaches to fusion of multimodal imaging data. The first is blind source separation (BSS), which deals directly with the problem of identifying underlying sources utilizing statistical (un)correlation and (in)dependence within and across modalities. The second is deep learning, focusing on multimodal architectures for classification, embedding, and segmentation.

8.2 Blind Source Separation Methods

Blind source separation (BSS) deals with the general problem of *blindly* recovering hidden source signals $\mathbf{y}$ from a dataset $\mathbf{x}$, i.e., without any knowledge of the function $\mathbf{f}$ nor the parameters $\boldsymbol{\theta}$ which generate $\mathbf{x} = \mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$. It can be organized into subproblems according to the number of datasets contained in $\mathbf{x}$ and the presence of subsets of $\mathbf{y}$ grouped as multidimensional sources within any single dataset. The following taxonomy arranges BSS subproblems by increasing complexity:

SDU In the single-dataset unidimensional (SDU) subproblem, $\mathbf{x}$ consists of a single dataset whose sources are *not* grouped. This is the seminal and most studied area of BSS, including classical problems such as independent component analysis (ICA) (Comon 1994; Bell and Sejnowski 1995; Hyvärinen and Erkki 1997) and second-order blind identification (SOBI) (Belouchrani et al. 1993; Yeredor 2000).

MDU In the multidataset unidimensional (MDU) subproblem, $\mathbf{x}$ consists of one or more datasets and, while *no* sources are grouped within any dataset, multidimensional sources containing *a single source from each dataset* may occur. Examples in this area include canonical correlation analysis (CCA) (Hotelling 1936), partial least squares (PLS) (Wold 1966), and independent vector analysis (IVA) (Adalı et al. 2014; Kim et al. 2006).

SDM In the single-dataset multidimensional (SDM) subproblem, $\mathbf{x}$ consists of a single dataset with one or more multidimensional sources. Examples include multidimensional ICA (MICA) (Cardoso 1998; Lahat et al. 2012) and independent subspace analysis (ISA) (Hyvärinen and Köster 2006; Szabó et al. 2012).

MDM In the general multidataset multidimensional (MDM) problem, $\mathbf{x}$ contains one or more datasets, each with one or more multidimensional sources that may group further with single or multidimensional sources from the remaining datasets. Examples include multidataset ISA (MISA) (Silva et al. 2014a,b) and joint ISA (JISA) (Lahat and Jutten 2015).

These definitions support a natural hierarchy in which subproblems are contained within one another, with SDU problems being a special case of MDU, SDM, and MDM problems, and MDU and SDM problems being special cases of MDM.

The "blind" property of BSS makes it particularly powerful and attractive in the absence of a precise model of the measured system and with data confounded

by noise of unknown or variable characteristics. These are marked signatures of multimodal fusion applications exploring the extreme complexities of the human brain, with largely heterogeneous noise characteristics and artifacts occurring across data types. This is a clear indicator that BSS is ripe for application in multimodal fusion of human brain data, as we will illustrate in the following sections. To begin with, we present the mathematical notation for the general MDM problem, followed by an example of an application of ICA to fusion of brain MRI and EEG features. We then briefly review other more advanced applications of BSS to multimodal fusion of brain imaging data before moving on to deep learning methods.

8.2.1 General MDM Problem Statement

Given N observations of $M \geq 1$ datasets (or modalities), identify an unobservable latent source random vector (r.v.) $\mathbf{y} = \left[\mathbf{y}_1^T \cdots \mathbf{y}_M^T\right]^T$, $\mathbf{y}_m = [y_1 \cdots y_{C_m}]^T$, that relates to the observed r.v. $\mathbf{x} = \left[\mathbf{x}_1^T \cdots \mathbf{x}_M^T\right]^T$, $\mathbf{x}_m = [x_1 \cdots x_{V_m}]^T$, via a mixture function $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the function parameters. Both $\mathbf{y}$ and the transformation represented by $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ have to be learned blindly, i.e., without explicit knowledge of either of them. In order to make this problem tractable, a few assumptions are required:

1. the number of latent sources C_m in each dataset is known by the experimenter;
2. $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) = \mathbf{A}\mathbf{y}$, i.e., a *linear* transformation, with $\boldsymbol{\theta} = \mathbf{A}$;
3. $\mathbf{A}$ is a $\bar{V} \times \bar{C}$ block diagonal matrix with M blocks, representing a separable layout structure such that $\mathbf{x}_m = \mathbf{A}_m \mathbf{y}_m$, $m = 1 \ldots M$, where $\bar{C} = \sum_{m=1}^{M} C_m$, $\bar{V} = \sum_{m=1}^{M} V_m$, and each block $\mathbf{A}_m$ is $V_m \times C_m$;
4. some of the latent sources in $\mathbf{y}$ are statistically related to each other and this *dependence* is undirected (non-causal), occurring both within or across datasets;
5. related sources establish subspaces (or source *groups*) $\mathbf{y}_k$, $k = 1 \ldots K$, with both K and the specific subspace compositions known by the experimenter and prescribed in an assignment matrix $\mathbf{P}_k$.

Under these assumptions, recovering the sources $\mathbf{y}$ amounts to finding a linear transformation $\mathbf{W}$ of the observed datasets via the unmixing function $\mathbf{y} = \mathbf{W}\mathbf{x}$. This is accurate when $\mathbf{W} = \mathbf{A}^{-}$, the pseudo-inverse of $\mathbf{A}$, which implies $\mathbf{W}$ is also block diagonal, thus satisfying $\mathbf{y}_m = \mathbf{W}_m \mathbf{x}_m$. Source subspaces are then estimated as $\mathbf{y}_k = \mathbf{P}_k \mathbf{W}\mathbf{x}$. In the following, unless noted otherwise, the m-th $V_m \times N$ data matrix is denoted as $\mathbf{X}_m$, containing N observations of $\mathbf{x}_m$ along its columns; $\mathbf{X}$ denotes a $\bar{V} \times N$ matrix concatenating all $\mathbf{X}_m$. Figure 8.1 illustrates this model, starting with its special cases.

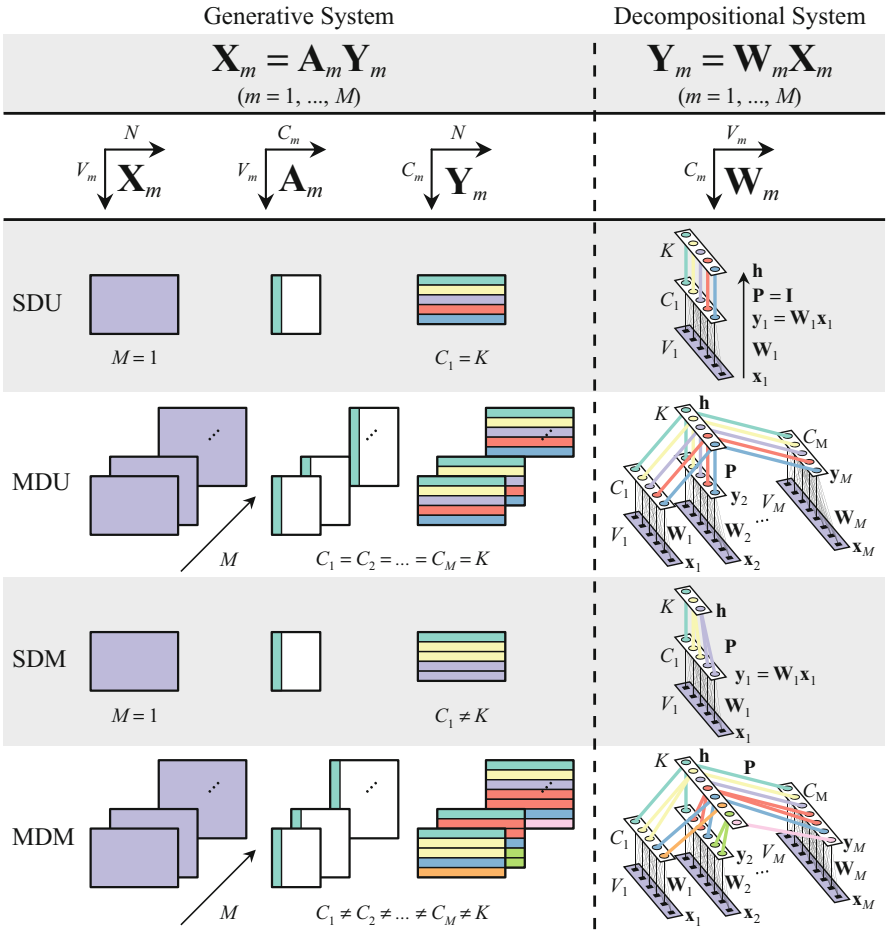

Fig. 8.1 Side-by-side illustration of the generative and decompositional system representations of linear BSS problems. Each of M datasets (or modalities) is represented by a matrix $\mathbf{X}_m$, with the same number of observations N along the columns. A column of $\mathbf{X}_m$ is represented by $\mathbf{x}_m$ (likewise for $\mathbf{Y}_m$ and $\mathbf{y}_m$). The generative system representation describes how each modality is generated from a set of underlying sources, in this case by a linear transformation of the source matrix $\mathbf{Y}_m$ through $\mathbf{A}_m$, the *mixing* matrix. In the general case, both $\mathbf{A}_m$ and $\mathbf{Y}_m$ are unique to each modality. Associations across modalities are represented by subspaces (K), which are collections of statistically *dependent* sources. *This dependence is indicated by coloring sources with the same color*. The linearity of the generative system implies linearity of the decompositional system. The decompositional representation indicates how source estimation occurs, namely by decomposing modalities into their underlying sources via a linear transformation of each modality $\mathbf{X}_m$ through $\mathbf{W}_m$, the *unmixing* matrix. In this representation, each V_m-dimensional column $\mathbf{x}_m$ is linearly transformed into a C_m-dimensional vector $\mathbf{y}_m$, whose elements (the individual sources) are then composed with other sources into subspaces, according to an *assignment* matrix $\mathbf{P}$ and non-linearity $\mathbf{h}(\cdot)$ ensuing from the choice of activation and objective functions

8.2.2 Case Study: Multimodal Fusion with Joint ICA

Here we illustrate a case study of blind source separation applied to multimodal fusion of brain imaging data. Specifically, we focus on joint ICA (jICA) (Calhoun and Adalı 2009), a very attractive model because of its simplicity as an MDU-type model cleverly designed to operate like an SDU-type model. Like ICA, it seeks statistically independent $\mathbf{y}_k$ such that the joint probability density function (pdf) of all sources, $p(\mathbf{y})$, factors as the product of its marginal subspaces: $p(\mathbf{y}) = \prod_{k=1}^{K} p(\mathbf{y}_k)$. Its hallmark assumption, however, is that the same mixing matrix $\mathbf{A}$ generates all modalities. It also assumes *none* of the multimodal sources are statistically related, i.e., $p(\mathbf{y}_k) = \prod_{m=1}^{M} p(y_{mk})$, $\forall k$, and that the pdf $p(\cdot)$ is the same for all sources and modalities. This is equivalent to constraining the block-diagonal structure in the MDU subproblem to $\mathbf{A}_m = \mathbf{A}$, $\forall m$. However, rather than choosing an M-dimensional joint pdf for $\mathbf{y}_k$, jICA combines corresponding sources y_{mk} of $\mathbf{y}_k$ into a single one-dimensional pdf $p(y_i)$, where i is the source number and $i = k$, which conveniently permits an SDU-type solution utilizing any off-the-shelf ICA algorithm after simple side-by-side concatenation of the data matrices from each modality. This also eliminates the requirement that the number of observations N be the same (and corresponding) for all modalities, so N_1 may differ from N_2, yielding $N = N_1 + N_2$ and $V = V_1 = V_2 =$ number of subjects after concatenation. Thorough simulation studies (Silva et al. 2014c) have shown that jICA is fairly robust to violation of the *independence across modalities* and *same pdf* assumptions but *not* so with violation of the *same mixing matrix* $\mathbf{A}$ assumption, which resulted in poorer performance.

Three seminal works have utilized joint ICA for multimodal fusion in brain imaging as a means to draw upon each modality's strengths and provide new information about the brain not offered by either modality alone. Firstly, fusion of multitask fMRI features (Calhoun et al. 2006b) promoted the direct use of data modeled at the subject level in a "unified analytic framework" for joint examination of multitask fMRI activations, leading to interesting, new findings that were missed by traditional analyses. Blood oxygen level dependent (BOLD) fMRI scans from 15 healthy control subjects and 15 outpatients with chronic schizophrenia matched for age, gender, and task difficulty were collected during two separate tasks: an auditory "oddball" task (AOD) and a Sternberg working memory task (SB). For every subject, regressors were created by modeling correct responses to task-specific stimuli as delta functions convolved with a canonical hemodynamic response function (HRF). These regressors plus their temporal derivatives and an intercept were included in a general linear model (GLM) of multiple regression fit to every voxel timeseries. The resulting AOD target-versus-standard contrast and SB recognition (or recall) contrast against baseline from each subject (averaged over all levels of difficulty) were corrected for amplitude bias due to spatially varying latencies using derivative boost and then arranged into matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ (AOD and SB features, respectively). Both matrices were normalized to have the same average sum-of-squares before concatenation, followed by (joint) PCA

data reduction and ICA, using the extended Infomax algorithm to adaptively allow some flexibility on the combined source pdfs $p(y_i)$ and, thus, mitigate potential side effects of violations to the same pdf assumption. Finally, rather than testing thousands of voxels, two-sample t-tests on each column of the shared subject expression profiles $\mathbf{A}$ were conducted to identify sources with significant group differences in coupling (regarded as a relative measure of the degree of group-level functional connectivity difference). For the identified source (Fig. 8.2), the joint probability of the multitask *data* $p(x_1(n_1), x_2(n_2))$ was assessed by means of subject-specific joint histograms, where n_m were the voxel indexes for modality m sorted from largest to smallest by their *source* values y_{mn} over all $n = 1, \ldots, N$, on voxels surviving an arbitrary $|Z| > 3.5$ threshold.

Secondly, fusion of fMRI and sMRI features (Calhoun et al. 2006a) enabled a direct study of the interactions and associations between changes in fMRI activation and changes in brain structure contained in sMRI data. Utilizing probabilistic segmentation (soft classification) maps of gray matter (GM) concentration derived from T_1-weighted sMRI images and the AOD target-versus-standard contrast from the same subjects described above, feature matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ were created, respectively. The sign of alternating voxels was flipped in GM maps to yield zero-mean maps for each subject (this step was undone after jICA estimation and before histogram computation and visualizations). Before concatenation of $\mathbf{X}_1$ and $\mathbf{X}_2$, both matrices were normalized to have the same average sum-of-squares. Joint PCA data reduction and ICA followed, using the extended Infomax algorithm to adaptively allow some flexibility on the combined source pdfs $p(y_i)$ and, thus, mitigate potential side effects of violations to the same pdf assumption. Like in the multitask case, two-sample t-tests on each column of the shared subject expression profiles $\mathbf{A}$ were conducted to identify sources with significant group differences and, for the identified source (Fig. 8.3), the joint probability of the multimodal *data* $p(x_1(n_1), x_2(n_2))$ was assessed by means of subject-specific joint histograms.

Lastly, fusion of EEG and fMRI features (Calhoun et al. 2006c) from 23 healthy control subjects enabled an attempt to resolve neuronal source activity with both high temporal and spatial resolution without needing to directly solve hard, untractable inverse problems. Event related potentials (ERP) were generated by time-locked averaging target epochs of the EEG signals from the midline central electrode (Cz) 200ms before to 1200ms after each target stimulus in an auditory "oddball" task. Also, t-statistic maps were obtained from fitting a GLM of regression to every voxel timeseries of a BOLD fMRI scan during the same oddball task, for a target-versus-standard contrast. Both features (ERPs ($\mathbf{X}_1$) and t-statistic maps ($\mathbf{X}_2$)) were computed on the same subjects for both modalities, with ERPs being interpolated to a number of ERP timepoints (N_1) that matched the number of fMRI voxels (N_2). Joint estimation of the ERP temporal sources ($\mathbf{Y}_1$) and t-map spatial sources ($\mathbf{Y}_2$) was carried out with jICA. High temporal and spatial resolution "snapshots" were then estimated by combining the multimodal sources, first as rows of $\mathbf{F}_{N_1 \times N_2} = \left|\mathbf{Y}_1^\top\right| \mathbf{Y}_2$ (an fMRI movie at high temporal resolution—Fig. 8.4), then as rows of $\mathbf{E}_{N_2 \times N_1} = \left|\mathbf{Y}_2^\top\right| \mathbf{Y}_1$ (a set of voxel-specific ERPs at

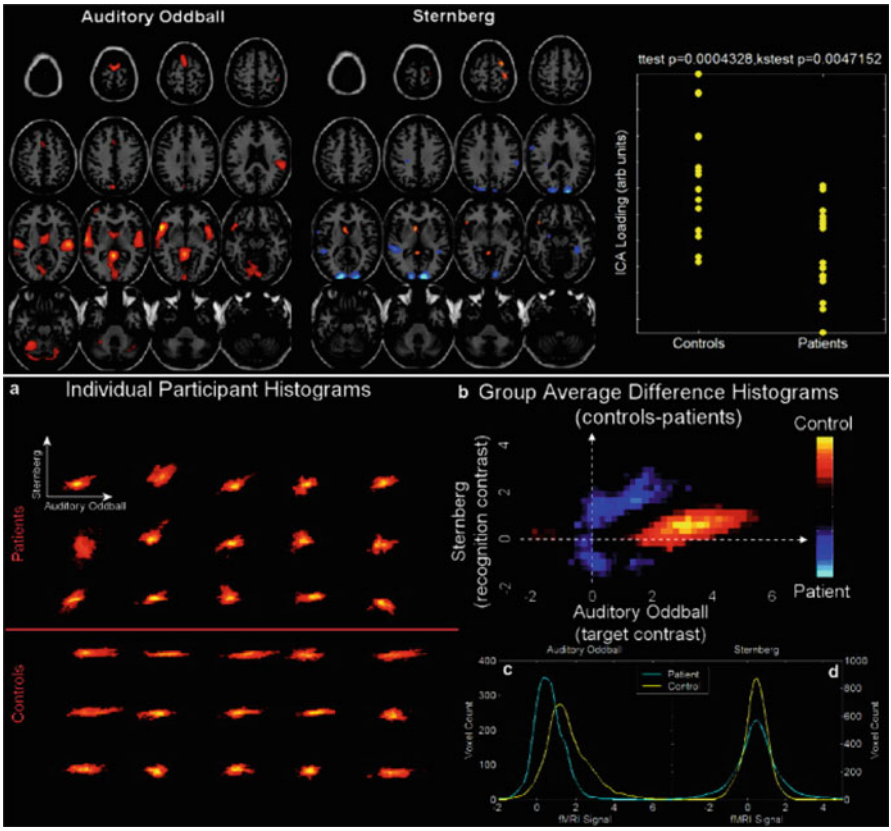

Fig. 8.2 Joint patterns of multitask group differences in schizophrenia. Top panel: Coupled joint source (network of co-varying maximally spatially independent maps) with significant difference in mixing coefficients between healthy controls and schizophrenic patients. Schizophrenia patients demonstrated lower mixing coefficient values **A** (the ICA loadings), which was interpreted as decreased functional connectivity in the joint network, particularly in temporal lobe, cerebellum, thalamus, basal ganglia, and lateral frontal regions, consistent with the cognitive dysmetria and frontotemporal disconnection models. Lower panel: (**a**) Subject-specific joint histograms: the correlation between the two tasks was significantly higher in patients than in controls, suggesting they activated "more similarly" on both tasks than controls; (**b**) Difference of group average histograms; (**c**,**d**) Marginal histograms: more AOD task voxels were active in controls and the SB task showed heavier tails in patients. Overall, the authors concluded that "patients are activating less, but also activating with a less-unique set of regions for these very different tasks." This suggested "both a global attenuation of activity as well as a breakdown of specialized wiring between cognitive domains." Copyright (2005) Wiley. Used with permission from V. D. Calhoun, *A method for multitask fMRI data fusion applied to schizophrenia*, Human Brain Mapping, John Wiley and Sons

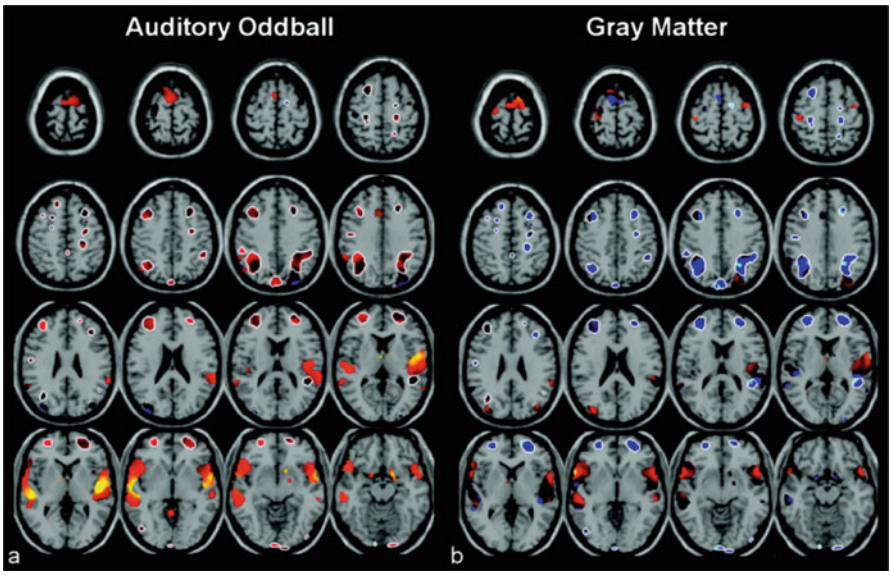

Fig. 8.3 Joint patterns of structural and functional group differences in schizophrenia. A joint multimodal independent source (not shown) with significant difference in mixing coefficients between patients and controls (higher for controls than for patients). Healthy controls showed mostly higher AOD activation in bilateral temporal lobe structures and cerebellum, associated with lower GM concentrations in bilateral frontal and parietal, as well as right temporal regions (not shown). A hypothesis of GM regions serving as "a morphological substrate for changes in AOD functional connectivity in schizophrenia" was suggested based on the coupling of those modalities via their shared mixing coefficients. The figure illustrates the t-values of a voxel-wise two-sample t-test for controls vs. patients of the data ($\mathbf{X}_1$ and $\mathbf{X}_2$) within the source regions surviving a $|Z| > 3.5$ threshold: (**a**) group differences in the AOD data over regions detected in the AOD part of the joint source (no outline) and GM part of the joint source (outlined in white), showing "more AOD activation in controls than patients." (**b**) group differences in the GM data over regions detected in the AOD part of the joint source (no outline) and GM part of the joint source (outlined in white), showing "GM values are increased in controls" over the AOD-detected regions, and decreased over the GM-detected regions (more so on the left than on the right). Orange: controls > patients; blue: the opposite. Copyright (2005) Wiley. Used with permission from V. D. Calhoun, *Method for Multimodal Analysis of Independent Source Differences in Schizophrenia: Combining Gray Matter Structural and Auditory Oddball Functional Data*, Human Brain Mapping, John Wiley and Sons

high spatial resolution—not shown), where $|\cdot|$ is the element-wise absolute value function. Overall, the results provide compelling evidence of the utility of such descriptive representation of the spatiotemporal dynamics of the auditory oddball target detection response, allowing the visualization, in humans, of the involved neural systems including participatory deep brain structures.

In summary, these results corroborate with previous evidence that methods combining the strengths of both techniques may reveal unique information and provide new insights into human brain function.

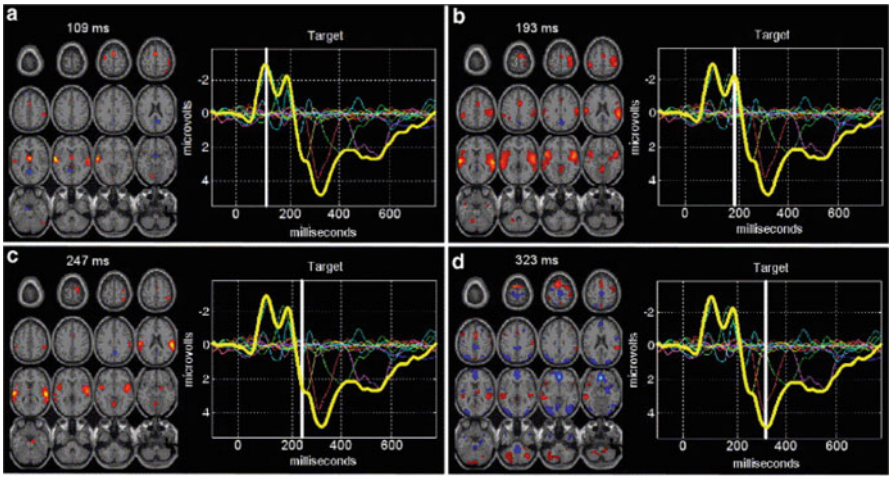

Fig. 8.4 Spatiotemporal dynamics of the auditory oddball target response. The N1 peak for the ERP data corresponded to primary and secondary auditory regions of the temporal lobe, and motor planning regions, as was expected following the initial auditory stimulus and the ensuing preparatory motor activity for the button press. Similarly, the N2 peak showed correspondence with extensive temporal lobe areas, including heteromodal association cortex, with motor planning, primary motor, and cerebellar regions also present, consistent with regions typically involved in the execution of the motor response. The P3a peak corresponded with additional temporal lobe regions, somatosensory cortex, and brain stem activity, consistent with what would be expected. In particular, the reported association of brain stem activity was evidence supportive of a previously hypothesized role for the locus coeruleus norepinephrine (LC-NE) system in generating the P3. This led to the conclusion that jICA can "reveal electrical sources which may not be readily visible to scalp ERPs and expose brain regions that have participatory roles in source activity but may not themselves be generators of the detected electrical signal." The image shows positive (orange) and negative (blue) Z values. Reprinted from NeuroImage, Vol 30 (1), V. D. Calhoun et al., *Neuronal chronometry of target detection: Fusion of hemodynamic and event-related potential data*, Pages 544–553, Copyright (2006), with permission from Elsevier

8.2.3 Advanced Blind Source Separation

The vast majority of approaches for multimodal analysis with BSS are rooted on MDU models. Their key strength is in the ability to not only utilize uncorrelation (or independence) between hidden sources for separation, like separate SDU models for each modality would do, but also leverage the correlation (or dependence) among corresponding multimodal sources to help steer the estimation procedure, automatically identifying linked sources. This increases the overall source separation power by leveraging information in one modality to improve estimation in the other modalities and vice-versa. In the following, we briefly review a number of MDU models and their applications to brain data analysis. The reader is encouraged to explore a recent review (Silva et al. 2016) which outlines further details on the models discussed below.

When (un)correlation, i.e., *linear* (in)dependence, is the sole mechanism for identification and separation of the sources, the models are categorized as second-order statistics (SOS) models. Classical algorithms such as CCA (Hotelling 1936) and PLS (Wold 1966), as well as more recent models such as multiset CCA (mCCA) (Kettenring 1971) and second-order IVA (IVA-G) (Anderson et al. 2010, 2012; Adalı et al. 2014) fall under this category. CCA maximizes the *correlation* between related source pairs $\mathbf{y}_{k=i} = [y_{1i},\ y_{2i}]^\top$ within the same subspace k, where $y_{1i} = \mathbf{W}_{1i}\mathbf{x}_1$ and $y_{2i} = \mathbf{W}_{2i}\mathbf{x}_2$ for $i = 1 \ldots C$ sources, and $\mathbf{W}_{mi}$ is the i-th row of $\mathbf{W}_m$, while PLS maximizes their *covariance* instead. Some extensions of these approaches have focused on expanding these notions beyond just 2 datasets (or modalities), like multi-set CCA (mCCA) (Correa et al. 2009), as well as leveraging higher-order statistics (HOS) to exploit source independence rather than uncorrelation, as in higher-order IVA (Anderson et al. 2013).

CCA's closed form solution for $M = 2$ datasets was utilized by Correa et al. (2008) to identify highly correlated subject expression profiles across fMRI+ERP and fMRI+sMRI datasets (with $N =$ number of subjects). For three modalities, mCCA based on sum of squared correlations (SSQCOR) was utilized for 3-way fusion of fMRI+ERP+sMRI (Correa et al. 2009), also seeking correlated subject expression profiles. In the case of fusion of simultaneous (concurrent) fMRI+EEG, efforts have been made to identify correlated temporal profiles ($N =$ time points) using mCCA across modalities and subjects (one downsampled, HRF-convolved single-trial ERP dataset and one fMRI dataset per subject: $M = 2 \times$ number of subjects) (Correa et al. 2010). In all cases above, the mixing matrix was estimated as $\mathbf{A}_m = \mathbf{X}\mathbf{Y}_m^-$, motivated by least squares projection. A CCA-type analysis was also pursued in source power comodulation (SPoC) (Dähne et al. 2014a), seeking associations between windowed variance profiles (neuronal oscillations from EEG) in $\mathbf{y}_1$ and a single known fixed reference source (behaviorally relevant parameters) y_{21} (considered to be already unmixed). Extensions of this method include canonical SPoC (cSPoC) (Dähne et al. 2014b), which pursued CCA between "envelope" transformations (instantaneous amplitudes) of $\mathbf{y}_m$, where $\mathbf{x}_m$ were rest EEG data from the same subject filtered at different frequency bands, and multimodal SPoC (mSPoC) (Dähne et al. 2013), which pursued CCA between simultaneously measured EEG (or MEG) temporal sources $\mathbf{y}_1$ and temporally filtered windowed variance profiles of fNIRS (or fMRI) temporal sources $\mathbf{y}_2$. The key differences between CCA and SPoC-type approaches are that $\mathbf{y}_1$ and $\mathbf{y}_2$ can have different number of observations and at least one set of sources undergoes a *non-linear* transformation. Another recent variant of CCA for multimodal fusion in neuroimaging is structured and sparse CCA (ssCCA) (Mohammadi-Nejad et al. 2017). This approach also identifies highly correlated subject expression profiles from multimodal data but imposes non-negativity, sparsity, and neighboring structure constraints on each row of $\mathbf{W}_m$. These constraints are expected to improve the interpretability of the resulting features directly from $\mathbf{W}_m$ (i.e., with no estimation of $\mathbf{A}_m$). The approach was utilized for fusion of eigenvector centrality maps of rest fMRI and T1-weighted sMRI from 34 Alzheimer's disease (AD) and 42 elderly healthy controls from the

Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, identifying two sets of multimodal regions highly associated to the disease label.

For PLS, Chen et al. (2009) utilized PLS regression to analyze GM concentration images from sMRI and ^{18}F-fluorodeoxyglucose (FDG) PET in two ways: (1) defining $\mathbf{X}_1$ as the GM maps from N subjects, $\mathbf{X}_2$ as the FDG maps from the same N subjects, and utilizing the (multivariate) PLS2 deflation strategy (Silva et al. 2016) to predict the FDG maps from the GM maps; and (2) defining $\mathbf{X}_1 = [\mathbf{X}_{FDG}^\top, \mathbf{X}_{GM}^\top]^\top$, i.e., the $(V_1 + V_2) \times N$ spatial concatenation of FDG and GM maps, and $\mathbf{X}_2$ as the $1 \times N$ age group label (younger or older), using (univariate) PLS1 for deflation (Silva et al. 2016), deflating only $\mathbf{X}_2$ (but not $\mathbf{X}_1$, for the sake of better interpretability). The latter approach is akin to jICA in the sense that the joint spatial features "share" similar expression levels over subjects, although here data reduction occurs at the feature dimension (V_m) instead of the subject dimension (N). The same approach was recently used with 3 modalities on mild cognitive impairment (MCI) patients, some of which had converted to Alzheimer's disease (AD) and some who had not (Wang et al. 2016). A similar study on a larger population is also available (Lorenzi et al. 2016).

In the case of modalities whose data can be arranged into multidimensional arrays, it is possible to utilize multilinear algebra to extend PLS into multi-way[1] PLS (N-PLS). This was utilized to fuse simultaneous EEG and fMRI recordings of subjects resting with eyes closed (Martínez-Montes et al. 2004). The data was organized into a 3-way tensor $\mathbf{X}_1$ with the $V_1 \times N \times D$ EEG data and a matrix (2-way tensor) $\mathbf{X}_2$ with the $V_2 \times N$ fMRI data, where N was the number of timepoints (and corresponding EEG 'segments'), V_1 was the number of frequencies in the EEG spectrum of each EEG segment, V_2 was the number of fMRI voxels, and D was the number of EEG electrode channels. For the EEG data, the frequencies of each electrode were convolved with the HRF over the time dimension to yield temporal "envelopes" of the EEG signal that were comparable to the fMRI timeseries. The model used for the EEG tensor was equivalent to $\mathbf{X}_{1,d} = \mathbf{A}_1 \text{diag}(\mathbf{b}_d)\mathbf{Y}_1$, $d = 1, \ldots, D$, where $\text{diag}(\mathbf{b}_d)$ is a diagonal matrix with $\mathbf{b}_d$ in the diagonal, i.e., the same decomposition $\mathbf{A}_1\mathbf{Y}_1$ was estimated in every EEG channel except for a set of scaling values $\mathbf{b}_d$ specific to each channel, which can be interpreted as a model of shared (i.e., same) sources $\mathbf{Y}_1$ with electrode-specific mixing $\mathbf{A}_{1,d} = \mathbf{A}_1 \text{diag}(\mathbf{b}_d)$. The covariance between the temporal EEG envelope sources $\mathbf{Y}_1$ and fMRI time course sources $\mathbf{Y}_2$ was then maximized, utilizing an extension of the PLS2 deflation strategy, which accommodates tensors, to predict the fMRI timeseries $\mathbf{X}_2$ from the EEG envelope sources $\mathbf{Y}_1$. This procedure yielded an fMRI map (a column of $\mathbf{A}_2$) whose time course (row of $\mathbf{Y}_2$) covaried highly with an EEG envelope (row of $\mathbf{Y}_1$) corresponding to an alpha band spectrum (column of $\mathbf{A}_1$) and a topographical map described by the electrode-specific scalars $\mathbf{b}_d$. This topographical map was

[1] While here "multi-way" refers to the order of a tensor (i.e., the number of data dimensions), the term multi-way has also been used in the literature to refer to the number of modalities being fused.

also studied using current source localization to identify the generators of the "EEG alpha rhythm".

For IVA, in comparison to mCCA, there are two key differences: (1) $\mathbf{W}$ is not constrained to have orthogonal rows,[2] and (2) HOS can be utilized to identify the sources. Together, these differences allow IVA to generalize mCCA, attaining more compact representations in $\mathbf{A}$ (Adalı et al. 2015) and leveraging HOS *dependence* between linked sources for improved separation.[3] Moreover, in a comparison with jICA, Adalı et al. (2015) noted that although IVA is more flexible when the subject expression profiles differ across a subset of the datasets (i.e., when the "same mixing matrix" assumption of jICA is violated), in very small N (number of subjects) regimes HOS estimation is unreliable and, thus, infeasible. Therefore, IVA-G was utilized instead, since it relies exclusively on SOS, just like mCCA. In the study, a GLM contrast map from fMRI, a GM concentration map from sMRI, and an ERP timeseries from EEG were obtained from 22 healthy controls and 14 schizophrenic patients ($N = 36$ subjects) performing an AOD task. Results from single and pairwise combinations of modalities were compared against the three-modality case. The study concluded that, for this particularly small dataset, "jICA provides a more desirable solution" using a flexible density matching ICA algorithm, a result likely driven by the drastically larger number of observations in the jICA model versus that of IVA for this study.

Another class of data fusion algorithms is based on two-step approaches that pursue BSS of either $\mathbf{A}$ or $\mathbf{Y}$ separately, after fitting an initial BSS model on $\mathbf{X}$. Two models that stand out in this class are "spatial" CCA+jICA (Sui et al. 2010) and mCCA+jICA (Sui et al. 2011). Spatial CCA+jICA uses CCA to initially identify correlated sources $\mathbf{Y}_1^{\text{CCA}} = \mathbf{W}_1^{\text{CCA}}\mathbf{X}_1$ and $\mathbf{Y}_2^{\text{CCA}} = \mathbf{W}_2^{\text{CCA}}\mathbf{X}_2$ in the usual way. However, within each modality, these CCA sources are just uncorrelated, and their separation is not guaranteed if the underlying source (canonical) correlations are equal or very similar (Sui et al. 2010). Thus, jICA on the concatenated *source* matrices $\mathbf{Y}_1^{\text{CCA}}$ and $\mathbf{Y}_2^{\text{CCA}}$ is utilized to further identify joint *independent* sources $\mathbf{Y}_1^{\text{jICA}} = \mathbf{W}^{\text{jICA}}\mathbf{Y}_1^{\text{CCA}}$ and $\mathbf{Y}_2^{\text{jICA}} = \mathbf{W}^{\text{jICA}}\mathbf{Y}_2^{\text{CCA}}$, where $\mathbf{W}^{\text{jICA}}$ is shared across modalities. The final mixing matrix of the spatial CCA+jICA model is then estimated as $\mathbf{A}_m = \left(\mathbf{W}^{\text{jICA}}\mathbf{W}_m^{\text{CCA}}\right)^-$. This model was utilized on multitask fMRI contrast maps derived from subject-level GLM (see Sect. 8.2.2), with $V =$ subjects and $N =$ feature dimensionality (here, voxels), resulting in interpretable multitask independent sources with similar (i.e. highly correlated) spatial map configurations (Sui et al. 2010). To note, such property should also be attainable with IVA directly applied to $\mathbf{X}_m$ and is worth of further investigation. The mCCA+jICA approach (Sui et al. 2011), on the other hand, utilizes mCCA to initially identify highly correlated *subject expression profiles* (rather than features) across m

[2]IVA-G is identical to mCCA with the GENVAR cost, except it also allows non-orthogonal $\mathbf{W}$.

[3]The IVA cost is a sum of M separate ICAs (one per dataset) with an additional term to increase/retain the mutual information between corresponding sources across datasets.

modalities, $\mathbf{Y}_{\text{CCA},m}^{\top} = \mathbf{X}_m \mathbf{W}_{\text{CCA},m}^{\top}$, where $\mathbf{X}_m$ is $V \times N_m$ (number of subjects (V) by feature dimensionality (N_m)). Notice the multiplication from the right of $\mathbf{X}_m$ and the matrix transposes resulting from V being treated as the observations. Thus, the mCCA $V \times N_m$ mixing matrices constitute the features estimated by least squares as $\mathbf{A}_{\text{CCA},m}^{\top} = (\mathbf{Y}_{\text{CCA},m}^{\top})^{-}\mathbf{X}_m$. Joint ICA is then performed on the concatenated *mixing matrices* $\mathbf{A}_{\text{CCA},m}^{\top}$ (along the feature dimension N_m) to identify joint sources $\mathbf{Y}_{\text{jICA},m} = \mathbf{W}_{\text{jICA}}\mathbf{A}_{\text{CCA},m}^{\top}$, where the $V \times V$ matrix $\mathbf{W}_{\text{jICA}}$ is shared across modalities. The final mixing matrix of the mCCA+jICA model is then estimated as $\mathbf{A}_m = \mathbf{Y}_{\text{CCA},m}^{\top}\mathbf{W}_{\text{jICA}}^{-1}$. This model was used by Sui et al. (2011) to perform fusion of GLM-derived fMRI contrast maps and DWI fractional anisotropy (FA) maps from each subject, yielding good separation across 62 healthy control (HC), 54 schizophrenic (SZ), and 48 bipolar (BP) disorder subjects, as indicated by pairwise two-sample t-tests of the group mixing coefficients in each column of each $\mathbf{A}_m$. Source maps for each group and modality were obtained by back-reconstruction, partitioning $\mathbf{A}_m$ into three blocks, $\mathbf{A}_{g,m}$, $g \in \{\text{HC}, \text{SZ}, \text{BP}\}$, one from each group respectively, and computing $\mathbf{Y}_{g,m} = (\mathbf{A}_{g,m})^{-}\mathbf{X}_{g,m}$. In a 3-way study, Sui et al. (2013) explored this approach to study group differences between 116 healthy controls and 97 schizophrenic patients, fusing GLM-derived contrast maps for the tapping condition of a block-design auditory sensorimotor task, together with FA maps and GM concentration maps from each subject. Finally, a very large study by Miller et al. (2016) on $V = 5{,}034$ subjects from the UK Biobank cohort defined $\mathbf{X}_1$ as a collection of $N_1 = 2{,}501$ image-derived phenotype (IDP) variables (individual measures of brain structure from T1-, T2-, and susceptibility-weighted sMRI, brain activity from task and rest fMRI, and local tissue microstructure from diffusion MRI), and $\mathbf{X}_2$ as a collection of $N_2 = 1{,}100$ non-imaging phenotype (non-IDP) variables extracted from the UK Biobank database (grouped into 11 categories) on the same subjects. In this study, the subject expression profiles were combined into a single *shared* profile, $\mathbf{Y}_{\text{CCA}}^{\top} = \mathbf{Y}_{\text{CCA},1}^{\top} + \mathbf{Y}_{\text{CCA},2}^{\top}$, which was used to estimate the modality-specific CCA mixing matrices, i.e., the features[4] $\mathbf{A}_{\text{CCA},m}^{\top} = (\mathbf{Y}_{\text{CCA}}^{\top})^{-}\mathbf{X}_m$. Moreover, rather than estimating mixing matrices with the form above, a final *shared* mixing matrix of the mCCA+jICA model is estimated as $\mathbf{A} = \mathbf{Y}_{\text{CCA}}^{\top}\mathbf{A}_{\text{jICA}}$, where $\mathbf{A}_{\text{jICA}} = \left[\mathbf{A}_{\text{CCA},1}^{\top}, \mathbf{A}_{\text{CCA},2}^{\top}\right] \cdot \left[\mathbf{Y}_{\text{jICA},1}, \mathbf{Y}_{\text{jICA},2}\right]^{-}$ ($[\,\cdot\,,\,\cdot\,]$ indicates matrix concatenation).[5]

[4]The MATLAB code used for this study (available at http://www.fmrib.ox.ac.uk/ukbiobank/nnpaper/ukb_NN.m) actually implements this step as $\left[\mathbf{A}_{\text{CCA},1}, \mathbf{A}_{\text{CCA},2}\right] = F(\mathbf{R}^{\mathbf{yx}})$, where $F(\cdot) = \text{atanh}(\cdot)$ is the element-wise Fisher transform of the $C \times (N_1 + N_2)$ cross-correlation matrix $\mathbf{R}^{\mathbf{yx}} = \text{diag}\left(\mathbf{Y}_{\text{CCA}}\mathbf{Y}_{\text{CCA}}^{\top}\right)^{-\frac{1}{2}}(\mathbf{Y}_{\text{CCA}}\mathbf{X})\text{diag}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-\frac{1}{2}}$ between $\mathbf{y}_{\text{CCA}}$ and $\mathbf{x}^{\top}$, $\text{diag}(\mathbf{B})$ is a diagonal matrix containing only the diagonal elements of $\mathbf{B}$, and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ is a matrix concatenation. Equivalence to the form indicated in the main text is claimed but not proven.

[5]Note that the implementation of mCCA+jICA in that work utilized simple matrix transpose instead of the pseudo-inverses indicted above, possibly presuming that the columns of $\mathbf{Y}_{\text{CCA}}^{\top}$ and rows of $\left[\mathbf{Y}_{\text{jICA},1}, \mathbf{Y}_{\text{jICA},2}\right]$ are orthonormal due to uncorrelation and independence, respectively.

Finally, approaches such as Parallel ICA (Liu et al. 2007) make up a unique class of BSS methods that seek to attain multiple goals simultaneously in an adaptive fashion. Specifically, rather than pursuing a decomposition into two sequential steps like with mCCA+jICA, Parallel ICA carries out separate ICA decompositions of each modality (i.e., in "parallel") while simultaneously identifying and reinforcing associations (in the form of correlations) among specific rows/columns of $\mathbf{A}_m$, $\mathbf{Y}_m$, or both, depending on how the modalities are treated/organized (i.e., if one or more of the datasets is transposed or not). The most widely used implementation simultaneously optimizes for maximal independence among sources $\mathbf{y}_m$ for each modality, treating the columns of $\mathbf{Y}_m$ as observations (like multiple separate SDU models), and maximal correlation among corresponding *mixing coefficients* $\mathbf{a}_k = [a_{1k}, a_{2k}, \ldots, a_{Mk}]^\top$ over modalities, treating the rows of $\mathbf{A}_m$ as observations (like an MDU model, but operating on pair-wise correlations individually rather than as a cohesive correlation matrix). These are typically competing objectives, leading to a trade-off between them (Vergara et al. 2014). Parallel ICA has been widely used in imaging genetics, offering a direct approach to identify neuroimaging endophenotypes related to various mental illnesses by fusing modalities such as fMRI and SNP (Liu et al. 2009), sMRI and SNP (Meda et al. 2012), as well as fMRI, sMRI, and SNP in a 3-way analysis (Vergara et al. 2014). It has also found use in fusion of resting-state networks (RSN) and behavioral measures (Meier et al. 2012).

While BSS has proven to be very fruitful for multimodal fusion thus far, it has mostly been focused on MDU methods. Much stands to be gained from subspaces that span multiple sources within a single dataset in terms of both improved representation power of complex features and, especially, subject-specific characterizations. Such MDM approaches are poised to move multimodal fusion analyses much further and address some of the current challenges and limitations of the area. Indeed, MDM models can be seen as two-layer-deep multimodal networks with fixed connections at the second layer. Thus, one interpretation of MDM models is that they have the ability to recover certain non-linear mixtures of the sources. Given the nature of complex systems such as the brain, sources are highly likely to be non-linearly mixed, which also serves as motivation to the deep learning methods described in Sect. 8.3.

8.2.4 Further Reading

For a unifying BSS modeling framework and discourse on the connections between various additional BSS methods applied to multimodal and unimodal brain imaging data, see Silva et al. (2016).

For a general review on multimodal fusion for brain imaging data, see Calhoun and Sui (2016).

For an overview of methods, challenges, and prospects of multimodal fusion beyond the scope of brain imaging, see Lahat et al. (2015).

For a broader discussion of methods beyond BSS and their application to multimodal brain imaging integration, see Biessmann et al. (2011).

For a clear, generalized description of tensor analysis and fusion as coupled matrix-tensor factorization methods, see Karahan et al. (2015).

For a comprehensive and mathematically oriented account of SDU models, see the Handbook of BSS (Comon and Jutten 2010).

Finally, the less experienced reader interested in a smooth introduction to the preprocessing strategies leading into ICA (and beyond) are recommended to check out the excellent ICA book from Hyvärinen et al. (2002). Those readers might also enjoy the numerous insights contained in the chapter about methods grounded on information theory (including ICA) by Haykin (2008).

8.3 Deep Learning Methods

In the previous section we presented blind source separation approaches in the context of multimodal fusion, particularly those based on MDU models, which may be construed as items of a more general area of *unsupervised learning*. Naturally, the models considered thus far utilize only a single level of *linear* transformation of sources (for generation) or data (for decomposition). However, if deeper chains of linear transformations are considered, each followed by a *nonlinear activation function* of its outputs (Goodfellow et al. 2016), much more powerful and flexible models can be obtained, naturally allowing compositions of multiple modalities, all while resorting to just simple stochastic gradient descent (SGD) for optimization (Goodfellow et al. 2016, Section 8.3.1). While these deeper models are able to approximate arbitrarily complex nonlinearities in the data, simple SOS or HOS does not suffice to attain the typical "blind" property that is characteristic of *linear* BSS (Comon and Jutten 2010, Chapter 14). Thus, for the purposes of this section, we forfeit this property in favor of *supervised* deep models, which, in neuroimaging, constitute the majority of successful deep learning results obtained from real multimodal brain imaging data.

Feedforward Neural Networks, or multilayer perceptrons (MLPs), are a classic model for function approximation, such as for classifiers, where $\mathbf{y} = \mathbf{G}(\mathbf{x})$ maps an input data sample $\mathbf{x}$ to output labels $\mathbf{y}$. The mapping $\mathbf{G}(\cdot)$ can be approximated by an L-layer network $\mathbf{g}(\mathbf{x}, \boldsymbol{\Phi}) = \mathbf{g}^L(\mathbf{g}^{L-1}(\cdots(\mathbf{g}^1(\mathbf{x}))))$ with parameters $\boldsymbol{\Phi}$. Each function $\mathbf{g}^l$ is defined as a linear model $\mathbf{W}_l\mathbf{g}^{l-1} + b_l$, with weights $\mathbf{W}_l$ and bias b_l, followed by nonlinear functions $\mathbf{h}$ (the activation functions), such that:

$$\mathbf{g}^l = \mathbf{h}(\mathbf{W}_l\mathbf{g}^{l-1} + b_l), \tag{8.1}$$

where $\mathbf{g}^0 = \mathbf{x}$, and $\boldsymbol{\Phi} = \{\mathbf{W}_l, b_l\ ;\ l = 1 \ldots L\}$.

In the case of the increasingly popular *convolutional neural networks* (CNNs), instead of a matrix multiplication $\mathbf{W}_l\mathbf{x}$, convolution with some kernel $\mathbf{W}_l$ is utilized at each layer:

$$\mathbf{g}^l = \mathbf{h}(\mathbf{W}_l * \mathbf{g}^{l-1} + b_l). \tag{8.2}$$

In this case, it is common to also define $\mathbf{g}^l$ at certain layers as other operations such as pooling, for example "max pooling" (Zhou and Chellappa 1988), normalization, for example batch normalization (Ioffe and Szegedy 2015), or dropout (Srivastava et al. 2014).

CNNs have multiple advantages (Goodfellow et al. 2016) over MLPs when the input data contains local correlations. CNNs exploit that with their local and, as such, sparse connections. If in MLPs we are connecting every input with every output, here we are applying a kernel to only a small region of input defined by the kernel size. Yet, in deeper layers, neurons are still indirectly connected to larger regions of the input. The size of the region a neuron connects to within its input layer is determined by the size of its receptive field, which depends on the CNN's hyperparameters and architecture. Overall, local connectivity reduces the number of parameters, computational complexity and memory requirements. All that is achieved via parameter-tying, i.e., when the same parameters are (re)used for multiple locations of the input. Furthermore, convolving the same parameter kernel with the input yields translation invariance property of images.

When the CNN is used as a classifier, in which use it has arguably revived increased interest to neural networks and started the ongoing deep learning revolution (Krizhevsky et al. 2012), then the convolutional layers are followed by a few feed forward layers with the softmax prediction at the end. However, for some applications, such as segmentation, it is preferable to stay within convolution layers only and in this case the network is called fully convolutional (Long et al. 2015)

Both CNN types are shown in Fig. 8.5 and in the following sections we will give a short overview of the use of these models.

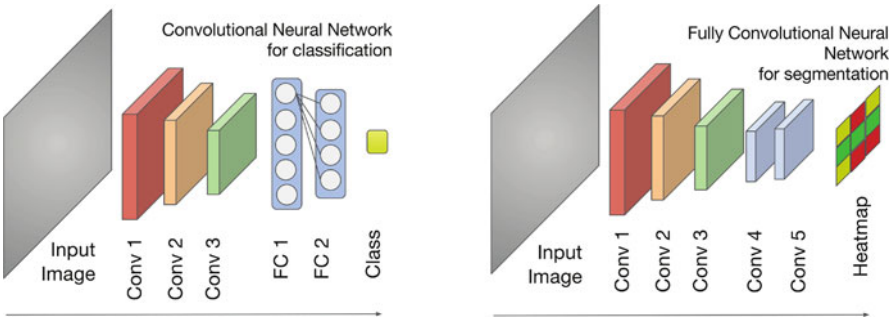

Fig. 8.5 Convolutional and fully convolutional neural networks. When used for classification tasks, CNNs typically feed directly into fully connected (FC) layers before classification. In segmentation tasks, however, fully convolutional networks can better retain the spatial structure of the data

8.3.1 Multimodal Classification

Feed forward neural networks are powerful classifiers that can achieve superior accuracy when trained on representative data. Their flexible and extensible architecture can be adjusted to handle cases that arise in practice. Ulloa et al. (2018) have built a multimodal classifier which combines structural and functional data to predict schizophrenia from brain imaging data (see Fig. 8.6). However, typical brain imaging datasets are comprised of fairly small numbers of subjects. To overcome the large data size requirements for training deep models, synthetic data generation approaches based on SDU models such as ICA have been proposed for augmenting these small datasets (Castro et al. 2015; Ulloa et al. 2015). Expanding on this idea, Ulloa et al. (2018) proposed to augment the training sets of datasets originating from different modalities. The augmentation process involves training a spatial ICA model for each modality (N = number of voxels) to learn both mixings $\mathbf{A}_m$ and sources $\mathbf{Y}_m$. Then, using only the labels of the training set, multidimensional sampling generates multiple new instances of mixing matrices $\mathbf{A}_m^r$ similar to $\mathbf{A}_m$. These are then combined with the ICA estimated sources $\mathbf{Y}_m$ to generate new synthetic examples of labeled data $\mathbf{X}_m^r$.

Initially, deep MLPs were trained separately for each modality utilizing only the synthetic data $\mathbf{X}_m^r$. The weights $\mathbf{W}_l$ from each MLP were then utilized to initialize the modality-specific weights of the final multimodal MLP, as indicated in Fig. 8.6. The multimodal MLP was then trained only on real data to classify disease labels using cross-validation. The resulting trained network was then evaluated on the test set in a 10-fold cross validation procedure yielding significantly improved results over other state of the art models, including the same MLP, that were either trained on a single modality or without using synthetic data (see Table 8.1).

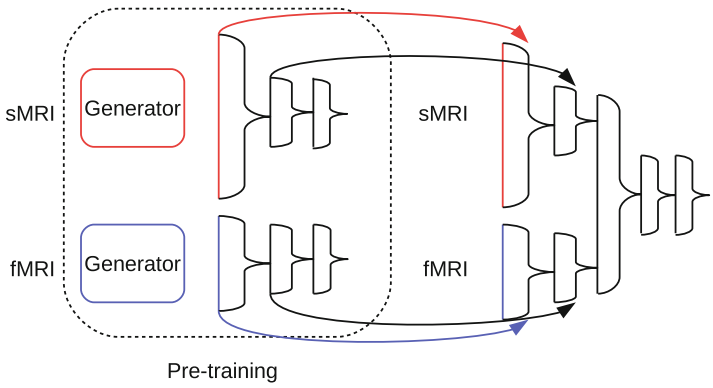

Fig. 8.6 Multimodal classifier. A multimodal MLP is one in which the deeper layers of the unimodal networks are combined (concatenated) together and treated as one. Here, the unimodal networks were trained on synthetic data separately. The weights learned on each modality separately using synthetic data were utilized to initialize the weights of the combined multimodal network, which was then trained using only real data

Table 8.1 Average and standard deviation of the area under the ROC curve (AUC) of an 8-fold cross validation experiment for various classifiers and the proposed methodologies

	sMRI		fMRI		sMRI + fMRI	
Classifier Method	Average AUC	Standard deviation	Average AUC	Standard deviation	Average AUC	Standard deviation
Online learning and synthetic data						
MLP with MVN	0.65	0.05	0.82	0.06	**0.85**	**0.05**
MLP with rejection	**0.74**	**0.07**	**0.83**	**0.05**	0.84	0.05
Raw data						
MLP	0.65	0.09	0.82	0.10	0.80	0.08
Naive Bayes	0.62	0.10	0.71	0.11	0.61	0.07
Logistic Regression	0.69	0.12	0.82	0.07	0.81	0.08
RBF SVM	0.53	0.05	0.82	0.08	0.58	0.15
Linear SVM	0.68	0.09	0.82	0.06	0.80	0.15
LDA	0.73	0.10	0.79	0.09	0.79	0.11
Random Forest	0.65	0.06	0.64	0.05	0.67	0.08
Nearest Neighbors	0.58	0.07	0.68	0.08	0.61	0.12
Decision Tree	0.56	0.11	0.54	0.10	0.53	0.13

8.3.2 Representation Learning for Semantic Embedding

The predictive advantages of multilayered models such as feed forward neural networks come from the powerful representations of the data that they automatically obtain at training. What that means is that the network learns a mapping of input data to the output layer vector space, where the input data samples are easily separable, thus encoding regularities in the data that are not easy to specify upfront. These output layer embeddings can be visualized if the multidimensional vectors are "projected" to a 2D space. Simple linear projections usually do not work well for this purpose, but nonlinear embedding methods such as t-distributed stochastic neighbor embedding (tSNE) (Maaten and Hinton 2008) do.

To obtain an embedding of a set of MRI images one first trains a deep model either for prediction or reconstruction. The obtained model is then used to produce activations at the output layer (or the one prior), which are subsequently represented as points on a 2D plane. Importantly, these points can later be assigned pseudo-color according to any property of interest. Plis et al. (2014) was one of the first to produce individual subject embeddings for MRI data. A deep 3-layer model trained to predict patients from healthy controls, possessing just that information, also learned to segregate disease severity of the patients as shown by the yellow-red spectrum in Fig. 8.7b.

The same approach has been applied to data from the Bipolar-Schizophrenia Network on Intermediate Phenotypes consortium (B-SNIP, http://www.b-snip.org/). The network was trained to predict three diseases from the spectrum (schizophrenia, the most severe, bipolar, and schizo-affective disorders) from healthy controls. After training, this network was used to produce embeddings for the data of subjects from

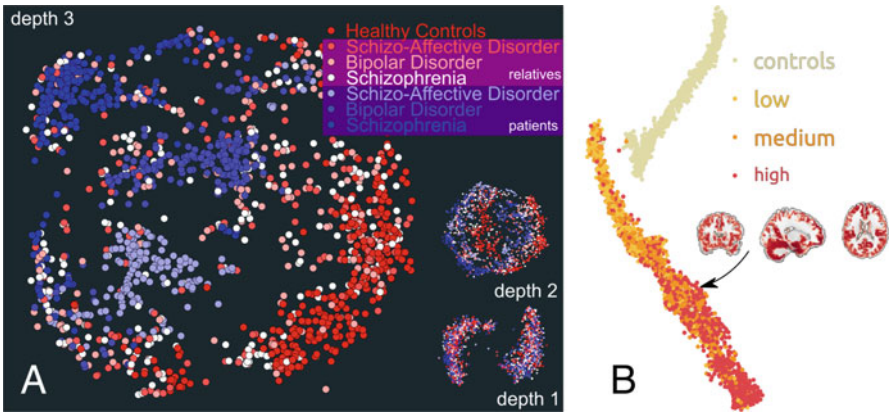

Fig. 8.7 Embedding deep network representations for healthy controls, patients with a spectrum of mental disorders and their unaffected siblings (**a**); for healthy controls and Huntington disease (HD) patients (**b**). Panel (**a**) also demonstrates sensitivity of embeddings to the network depth, where with depth the embedding becomes more interpretable. In panel (**b**), note the emergence of severity spectrum for HD patients despite unavailability of that information to the deep learning algorithm

its training set as well as the unaffected relatives that were previously unseen (shown in Fig. 8.7a). To further illustrate the value of depth in these models, Fig. 8.7a shows embeddings obtained from models of smaller depth: 1 and 2. These do not show such clear segregation spectrum.

8.3.3 Multimodal Tissue Segmentation

The problem of brain tissue segmentation is fundamental to almost any research study on the brain as gray matter volumes and thicknesses are potentially strong biomarkers for a number of disorders. In order to compute these, one needs to first segment the MRI images into various tissue types. Traditionally, a lengthy and computationally heavy process performed in multiple packages and usually relying on multiple sub-stages including skull stripping to rid anything but the brain. Simple gray, white matter and CSF segmentation is widespread enough to be interesting. It can sometimes be completed using simple techniques based on pixel intensity property. However, a much more valuable and yet much harder segmentation is into functional atlases, where each cortical and subcortical region is delineated according to their function relative to some atlas. The problem is challenging as it requires regions to be outlined not just based on voxel intensities alone but also on the relative location of the region within the brain.

Fedorov et al. (2017a) have successfully used a fully convolutional network of a specific kind (dilated convolutional kernels) to quickly (under 3 min, compared to more than 10 h state-of-the-art FreeSurfer (Dale et al. 1999)) partition an MRI in

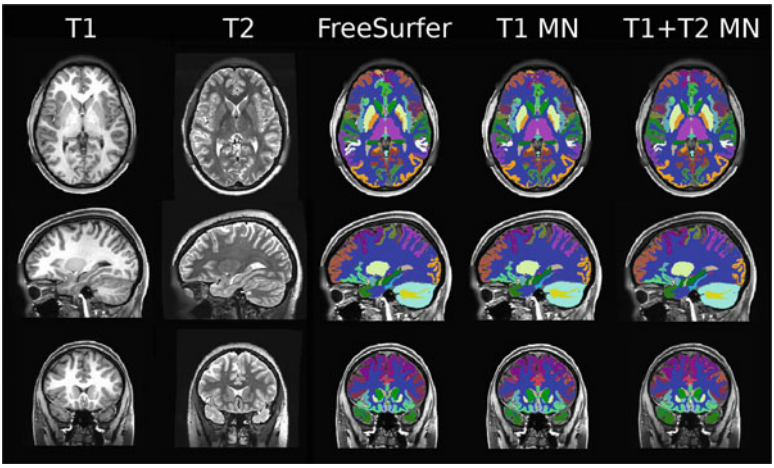

Fig. 8.8 Accelerating conventional approaches to tissue segmentation. Segmentation results produced by FreeSurfer on a single-subject image (center) after 10h of intense processing, using a trained CNN with dilated convolutional kernels (center-right) after 3 min, and using both T1 and T2 contrasts (right). T1 and T2 images included for reference (left and center-left, respectively)

the subject space into tissue types (Fedorov et al. 2017b) and functional regions. What is important for us here is that they have found significant improvements in segmentation accuracy when using multimodal input: not just T1 but also T2 contrast images (see Fig. 8.8). Deep learning models provide very simple mechanisms to use multimodal data without any additional difficulties. Another powerful feature for segmentation models comes from the fact that the learning signal can be produced at each predicted voxel, thus producing significant amounts of training data and reducing sample requirements for training. Çiçek et al. (2016) used just a handful of MRIs to produce a solid model.

8.4 Closing Remarks

Multimodal fusion is indeed a key element for discovery, understanding, and prediction in neuroimaging and mental health. Blind source separation and deep learning approaches have both demonstrated evidence of their ability to recover relevant information from multimodal data in multiple settings. The results presented here support the utility of multimodal approaches for brain imaging data analysis and suggest continued development of these methods, combined with increasingly large datasets, can yield strong, predictive features for both research and clinical settings. In particular, we highlight the current development of MDM approaches for identifying non-trivial hidden subspace structures, as well as deep architectures for unraveling the complex relationships between function and structure in the human brain. The combination of these two strategies holds great promise towards a unified approach for studying both healthy and disease conditions.

Acknowledgements We would like to thank Dr. Vince Calhoun for the useful discussions, as well as Alvaro Ulloa and Aleksandr Fedorov for kindly providing some of the images and results presented here. This work was supported by NIH grants R01EB006841 (SP), 2R01EB005846 (RS), and R01EB020407 (RS), NSF grants IIS-1318759 (SP), 1539067 (RS), and NIH NIGMS Center of Biomedical Research Excellent (COBRE) grant 5P20RR021938/P20GM103472/P30GM122734.

References

Adalı T, Anderson M, Fu GS (2014) Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging. IEEE Signal Process Mag 31(3):18–33. https://doi.org/10.1109/MSP.2014.2300511

Adalı T, Levin-Schwartz Y, Calhoun VD (2015) Multimodal data fusion using source separation: Application to medical imaging. Proc IEEE 103(9):1494–1506. https://doi.org/10.1109/JPROC.2015.2461601

Anderson M, Li XL, Adalı T (2010) Nonorthogonal independent vector analysis using multivariate gaussian model. In: Vigneron V, Zarzoso V, Moreau E, Gribonval R, Vincent E (eds) Proc LVA/ICA 2010, Lecture Notes in Computer Science, vol 6365. Springer, St. Malo, France, pp 354–361. https://doi.org/10.1007/978-3-642-15995-4_44

Anderson M, Adalı T, Li XL (2012) Joint blind source separation with multivariate gaussian model: Algorithms and performance analysis. IEEE Trans Signal Process 60(4):1672–1683. https://doi.org/10.1109/TSP.2011.2181836

Anderson M, Fu GS, Phlypo R, Adalı T (2013) Independent vector analysis, the Kotz distribution, and performance bounds. In: Proc IEEE ICASSP 2013, Vancouver, BC, pp 3243–3247. https://doi.org/10.1109/ICASSP.2013.6638257

Bell A, Sejnowski T (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7(6):1129–1159

Belouchrani A, Abed-Meraim K, Cardoso JF, Moulines E (1993) Second-order blind separation of temporally correlated sources. In: Proc ICDSP 1993, Nicosia, Cyprus, pp 346–351

Biessmann F, Plis S, Meinecke FC, Eichele T, Muller KR (2011) Analysis of multimodal neuroimaging data. IEEE Rev Biomed Eng 4:26–58. https://doi.org/10.1109/RBME.2011.2170675

Calhoun VD, Adalı T (2009) Feature-based fusion of medical imaging data. IEEE Trans Inf Technol Biomed 13(5):711–720. https://doi.org/10.1109/TITB.2008.923773

Calhoun VD, Sui J (2016) Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. Biol Psychiatry Cogn Neurosci Neuroimaging 1(3):230–244. https://doi.org/10.1016/j.bpsc.2015.12.005

Calhoun VD, Adalı T, Giuliani NR, Pekar JJ, Kiehl KA, Pearlson GD (2006a) Method for multimodal analysis of independent source differences in schizophrenia: Combining gray matter structural and auditory oddball functional data. Hum Brain Mapp 27(1):47–62. https://doi.org/10.1002/hbm.20166

Calhoun VD, Adalı T, Kiehl K, Astur R, Pekar J, Pearlson G (2006b) A method for multi-task fMRI data fusion applied to schizophrenia. Hum Brain Mapp 27(7):598–610. https://doi.org/10.1002/hbm.20204

Calhoun VD, Adalı T, Pearlson GD, Kiehl KA (2006c) Neuronal chronometry of target detection: Fusion of hemodynamic and event-related potential data. NeuroImage 30(2):544–553. https://doi.org/10.1016/j.neuroimage.2005.08.060

Cardoso JF (1998) Multidimensional independent component analysis. In: Proc IEEE ICASSP 1998, Seattle, WA, vol 4, pp 1941–1944. https://doi.org/10.1109/ICASSP.1998.681443

Castro E, Ulloa A, Plis SM, Turner JA, Calhoun VD (2015) Generation of synthetic structural magnetic resonance images for deep learning pre-training. In: Proc IEEE ISBI 2015, pp 1057–1060. https://doi.org/10.1109/ISBI.2015.7164053

Chen K, Reiman EM, Huan Z, Caselli RJ, Bandy D, Ayutyanont N, Alexander GE (2009) Linking functional and structural brain images with multivariate network analyses: A novel application of the partial least square method. NeuroImage 47(2):602–610. https://doi.org/10.1016/j.neuroimage.2009.04.053

Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proc MICCAI 2016, pp 424–432. https://doi.org/10.1007/978-3-319-46723-8_49

Comon P (1994) Independent component analysis, a new concept? Signal Process 36(3):287–314. https://doi.org/10.1016/0165-1684(94)90029-9

Comon P, Jutten C (2010) Handbook of blind source separation, 1st edn. Academic Press, Oxford, UK

Correa NM, Li YO, Adalı T, Calhoun VD (2008) Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. IEEE J Sel Topics Signal Process 2(6):998–1007. https://doi.org/10.1109/JSTSP.2008.2008265

Correa NM, Li YO, Adalı T, Calhoun VD (2009) Fusion of fMRI, sMRI, and EEG data using canonical correlation analysis. In: Proc IEEE ICASSP 2009, pp 385–388. https://doi.org/10.1109/ICASSP.2009.4959601

Correa NM, Eichele T, Adalı T, Li YO, Calhoun VD (2010) Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI. Neuroimage 50(4):1438–1445. https://doi.org/10.1016/j.neuroimage.2010.01.062

Dähne S, Bießmann F, Meinecke F, Mehnert J, Fazli S, Müller KR (2013) Integration of multivariate data streams with bandpower signals. IEEE Trans Multimedia 15(5):1001–1013. https://doi.org/10.1109/TMM.2013.2250267

Dähne S, Meinecke F, Haufe S, Höhne J, Tangermann M, Müller KR, Nikulin V (2014a) SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. NeuroImage 86:111–122. https://doi.org/10.1016/j.neuroimage.2013.07.079

Dähne S, Nikulin V, Ramírez D, Schreier P, Müller KR, Haufe S (2014b) Finding brain oscillations with power dependencies in neuroimaging data. NeuroImage 96:334–348. https://doi.org/10.1016/j.neuroimage.2014.03.075

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis: I. segmentation and surface reconstruction. NeuroImage 9(2):179–194. https://doi.org/10.1006/nimg.1998.0395

Fedorov A, Damaraju E, Calhoun V, Plis S (2017a) Almost instant brain atlas segmentation for large-scale studies. arXiv preprint URL http://arxiv.org/abs/1711.00457

Fedorov A, Johnson J, Damaraju E, Ozerin A, Calhoun V, Plis S (2017b) End-to-end learning of brain tissue segmentation from imperfect labeling. In: Proc IJCNN 2017, pp 3785–3792. https://doi.org/10.1109/IJCNN.2017.7966333

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. http://www.deeplearningbook.org

Haykin S (2008) Neural networks and learning machines, 3rd edn. Prentice Hall, Upper Saddle River, NJ

Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3/4):321–377. https://doi.org/10.2307/2333955

Hyvärinen A, Erkki O (1997) A fast fixed-point algorithm for independent component analysis. Neural Comput 9(7):1483–1492. https://doi.org/10.1162/neco.1997.9.7.1483

Hyvärinen A, Köster U (2006) FastISA: A fast fixed-point algorithm for independent subspace analysis. In: Proc ESANN 2006, Bruges, Belgium, pp 371–376

Hyvärinen A, Karhunen J, Oja E (2002) Independent component analysis, 1st edn. Wiley, New York, NY

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc ICML 2015, Lille, France, vol 37, pp 448–456

Karahan E, Rojas-López PA, Bringas-Vega ML, Valdés-Hernández PA, Valdés-Sosa PA (2015) Tensor analysis and fusion of multimodal brain images. Proc IEEE 103(9):1531–1559. https://doi.org/10.1109/JPROC.2015.2455028

Kettenring J (1971) Canonical analysis of several sets of variables. Biometrika 58(3):433–451. https://doi.org/10.2307/2334380

Kim T, Eltoft T, Lee TW (2006) Independent vector analysis: An extension of ICA to multivariate components. In: Proc ICA 2006, Springer, Charleston, SC, Lecture Notes in Computer Science, vol 3889, pp 165–172. https://doi.org/10.1007/11679363_21

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proc NIPS 2012, pp 1097–1105

Lahat D, Jutten C (2015) Joint independent subspace analysis: A quasi-Newton algorithm. In: Proc LVA/ICA 2015, Springer, Liberec, Czech Republic, Lecture Notes in Computer Science, vol 9237, pp 111–118. https://doi.org/10.1007/978-3-319-22482-4_13

Lahat D, Cardoso J, Messer H (2012) Second-order multidimensional ICA: Performance analysis. IEEE Trans Signal Process 60(9):4598–4610. https://doi.org/10.1109/TSP.2012.2199985

Lahat D, Adalı T, Jutten C (2015) Multimodal data fusion: An overview of methods, challenges, and prospects. Proc IEEE 103(9):1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

Liu J, Pearlson G, Calhoun V, Windemuth A (2007) A novel approach to analyzing fMRI and SNP data via parallel independent component analysis. Proc SPIE 6511:651,113–651,113–11. https://doi.org/10.1117/12.709344

Liu J, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero NI, Calhoun VD (2009) Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Hum Brain Mapp 30(1):241–255. https://doi.org/10.1002/hbm.20508

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proc IEEE CVPR 2015, pp 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Lorenzi M, Simpson IJ, Mendelson AF, Vos SB, Cardoso MJ, Modat M, Schott JM, Ourselin S (2016) Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations. Sci Rep 6:22,161. https://doi.org/10.1038/srep22161

Maaten Lvd, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov):2579–2605

Martínez-Montes E, Valdés-Sosa PA, Miwakeichi F, Goldman RI, Cohen MS (2004) Concurrent EEG/fMRI analysis by multiway partial least squares. NeuroImage 22(3):1023–1034. https://doi.org/10.1016/j.neuroimage.2004.03.038

Meda S, Narayanan B, Liu J, Perrone-Bizzozero N, Stevens M, Calhoun VD, Glahn D, Shen L, Risacher S, Saykin A, Pearlson G (2012) A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. NeuroImage 60(3):1608–1621. https://doi.org/10.1016/j.neuroimage.2011.12.076

Meier T, Wildenberg J, Liu J, Chen J, Calhoun VD, Biswal B, Meyerand M, Birn R, Prabhakaran V (2012) Parallel ICA identifies sub-components of resting state networks that covary with behavioral indices. Front Hum Neurosci 6:281. https://doi.org/10.3389/fnhum.2012.00281

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragonu I, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci 19(11):1523–1536. https://doi.org/10.1038/nn.4393

Mohammadi-Nejad AR, Hossein-Zadeh GA, Soltanian-Zadeh H (2017) Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach. IEEE Trans Med Imaging 36(7):1438–1448. https://doi.org/10.1109/TMI.2017.2681966

Plis SM, Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, Johnson HJ, Paulsen JS, Turner JA, Calhoun VD (2014) Deep learning for neuroimaging: a validation study. Front Neurosci 8:229. https://doi.org/10.3389/fnins.2014.00229

Silva RF, Plis SM, Adalı T, Calhoun VD (2014a) Multidataset independent subspace analysis. In: Proc OHBM 2014, Poster 3506

Silva RF, Plis SM, Adalı T, Calhoun VD (2014b) Multidataset independent subspace analysis extends independent vector analysis. In: Proc IEEE ICIP 2014, Paris, France, pp 2864–2868. https://doi.org/10.1109/ICIP.2014.7025579

Silva RF, Plis SM, Adalı T, Calhoun VD (2014c) A statistically motivated framework for simulation of stochastic data fusion models applied to multimodal neuroimaging. NeuroImage 102, Part 1:92–117. https://doi.org/10.1016/j.neuroimage.2014.04.035

Silva RF, Plis SM, Sui J, Pattichis MS, Adalı T, Calhoun VD (2016) Blind source separation for unimodal and multimodal brain networks: A unifying framework for subspace modeling. IEEE J Sel Topics Signal Process 10(7):1134–1149. https://doi.org/10.1109/JSTSP.2016.2594945

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958

Sui J, Adalı T, Pearlson G, Yange H, Sponheim S, White T, Calhoun V (2010) A CCA + ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. NeuroImage 51(1):123–134. https://doi.org/10.1016/j.neuroimage.2010.01.069

Sui J, Pearlson G, Caprihan A, Adalı T, Kiehl K, Liu J, Yamamoto J, Calhoun VD (2011) Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA + joint ICA model. NeuroImage 57(3):839–855. https://doi.org/10.1016/j.neuroimage.2011.05.055

Sui J, He H, Yu Q, Chen J, Rogers J, Pearlson G, Mayer A, Bustillo J, Canive J, Calhoun VD (2013) Combination of resting state fMRI, DTI and sMRI data to discriminate schizophrenia by N-way MCCA+jICA. Front Hum Neurosci 7(235). https://doi.org/10.3389/fnhum.2013.00235

Szabó Z, Póczos B, Lőrincz A (2012) Separation theorem for independent subspace analysis and its consequences. Pattern Recognit 45(4):1782–1791. https://doi.org/10.1016/j.patcog.2011.09.007

Ulloa A, Plis S, Erhardt E, Calhoun V (2015) Synthetic structural magnetic resonance image generator improves deep learning prediction of schizophrenia. In: Proc IEEE MLSP 2015, pp 1–6. https://doi.org/10.1109/MLSP.2015.7324379

Ulloa A, Plis SM, Calhoun VD (2018) Improving classification rate of schizophrenia using a multimodal multi-layer perceptron model with structural and functional MR. arXiv preprint URL http://arxiv.org/abs/1804.04591

Vergara VM, Ulloa A, Calhoun VD, Boutte D, Chen J, Liu J (2014) A three-way parallel ICA approach to analyze links among genetics, brain structure and brain function. NeuroImage 98:386–394. https://doi.org/10.1016/j.neuroimage.2014.04.060

Wang P, Chen K, Yao L, Hu B, Wu X, Zhang J, Ye Q, Guo X (2016) Multimodal classification of mild cognitive impairment based on partial least squares. J Alzheimers Dis 54(1):359–371. https://doi.org/10.3233/JAD-160102

Wold H (1966) Nonlinear estimation by iterative least squares procedures. In: David F (ed) Research papers in statistics. Festschrift for J. Neyman. Wiley, New York, NY, pp 411–444

Yeredor A (2000) Blind separation of gaussian sources via second-order statistics with asymptotically optimal weighting. IEEE Signal Process Lett 7(7):197–200. https://doi.org/10.1109/97.847367

Zhou YT, Chellappa R (1988) Computation of optical flow using a neural network. In: Proc IEEE ICNN 1988, vol 2, pp 71–78. https://doi.org/10.1109/ICNN.1988.23914