Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects

Dana Lahat, Tülay Adalı, Christian Jutten

Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects

Dana Lahat, Tülay Adalı, *Fellow, IEEE,* and Christian Jutten, *Fellow, IEEE*

***Abstract*—In various disciplines, information about the same phenomenon can be acquired from different types of detectors, at different conditions, in multiple experiments or subjects, among others. We use the term "modality" for each such acquisition framework. Due to the rich characteristics of natural phenomena, it is rare that a single modality provides complete knowledge of the phenomenon of interest. The increasing availability of several modalities reporting on the same system introduces new degrees of freedom, which raise questions beyond those related to exploiting each modality separately. As we argue, many of these questions, or "challenges", are common to multiple domains. This paper deals with two key questions: "why we need data fusion" and "how we perform it". The first question is motivated by numerous examples in science and technology, followed by a mathematical framework that showcases some of the benefits that data fusion provides. In order to address the second question, "diversity" is introduced as a key concept, and a number of data-driven solutions based on matrix and tensor decompositions are discussed, emphasizing how they account for diversity across the datasets. The aim of this paper is to provide the reader, regardless of his or her community of origin, with a taste of the vastness of the field, the prospects and opportunities that it holds.**

***Index Terms*—Keywords: data fusion, multimodality, multiset data analysis, latent variables, tensor, overview.**

I. Introduction

Information about a phenomenon or a system of interest can be obtained from different types of instruments, measurement techniques, experimental setups, and other types of sources. Due to the rich characteristics of natural processes and environments, it is rare that a single acquisition method provides complete understanding thereof. The increasing availability of multiple datasets that contain information, obtained using different acquisition methods, about the same system, introduces new degrees of freedom that raise questions beyond those related to analysing each dataset separately.

The foundations of modern data fusion have been laid in the first half of the 20th century [1], [2]. Joint analysis of multiple datasets has since been the topic of extensive research, and earned a significant leap forward in the late 1960's–early 1970's with the formulation of concepts and techniques such as multi-set canonical correlation analysis (CCA) [3], parallel factor analysis (PARAFAC) [4], [5], and other tensor decompositions [6], [7]. However, until rather recently, in most cases, these data fusion methodologies were confined within the limits of psychometrics and chemometrics, the communities in which they evolved. With recent technological advances, in a growing number of domains, the availability of datasets that correspond to the same phenomenon has increased, leading to increased interest in exploiting them efficiently. Many of the providers of multi-view, multirelational, and multimodal data are associated with high-impact commercial, social, biomedical, environmental, and military applications, and thus the drive to develop new and efficient analytical methodologies is high and reaches far beyond pure academic interest.

Motivations for data fusion are numerous. They include obtaining a more unified picture and global view of the system at hand; improving decision making; exploratory research; answering specific questions about the system, such as identifying common vs. distinctive elements across modalities or time; and in general, extracting knowledge from data for various purposes. However, despite the evident *potential* benefit, and massive work that has already been done in the field (see, for example, [8]–[16] and references therein), the knowledge of *how* to actually exploit the additional diversity that multiple datasets offer is still at its very preliminary stages.

Data fusion is a challenging task for several reasons [8]–[11], [17]–[19]. *First*, the data are generated by very complex systems: biological, environmental, sociological, and psychological, to name a few, driven by numerous underlying processes that depend on a large number of variables to which we have no access. *Second*, due to the augmented diversity, the number, type and scope of new research questions that can be posed is potentially very large. *Third*, working with heterogeneous datasets such that the respective advantages of each dataset are maximally exploited, and drawbacks suppressed, is not an evident task. We elaborate on these matters in the following sections. Most of these questions have been devised only in the very recent years, and, as we show in the sequel, only a fraction of their potential has already been exploited. Hence, we refer to them as "challenges".

A rather wide perspective on challenges in data fusion is presented by [8], which discusses linked-mode decomposition models within the framework of chemometrics and psychometrics, and [9], which focusses on "automated decision making" with special attention to multisensor information fusion. In practice, however, challenges in data fusion are most often brought up within a framework dedicated to a specific application, model and dataset; examples will be given in the sections that follow.

In this paper, we bring together a comprehensive (but definitely not exhaustive) list of challenges in data fusion.

D. Lahat and Ch. Jutten are with GIPSA-Lab, UMR CNRS 5216, Grenoble Campus, BP46, F-38402 Saint Martin d'Hères, France. T. Adalı is with the Department of CSEE, University of Maryland, Baltimore County, Baltimore, MD 21250, USA. email:{Dana.Lahat, Christian.Jutten}@gipsa-lab.grenoble-inp.fr, adali@umbc.edu. This work is supported by the project CHESS, 2012-ERC-AdG-320684 (D. Lahat and Ch. Jutten) and by the grants NSF-IIS 1017718 and NSF-CCF 1117056 (T. Adalı). GIPSA-Lab is a partner of the LabEx PERSYVAL-Lab (ANR–11-LABX-0025).

Following from [8], [9], [16], [19] (and others), and further emphasized by our discussion in this paper, it is clear that at the appropriate level of abstraction, the same challenge in data fusion can be relevant to completely different and diverse applications, goals and data types. Consequently, a solution to a challenge that is based on a sufficiently data-driven, model-free approach may turn out to be useful in very different domains. Therefore, there is an obvious interest in opening up the discussion of data fusion challenges to include and involve disparate communities, so that each community could inform the others. Our goal is to stimulate and emphasize the relevance and importance of a perspective based on challenges to advanced data fusion. More specifically, we would like to promote data-driven approaches, that is, approaches with minimal and weak priors and constraints, such as sparsity, non-negativity, low-rank and independence, among others, that can be useful to more than one specific application or dataset. Hence, we present these challenges in quite a general framework that is not specific to an application, goal or data type. We also give examples and motivations from different domains.

In order to contain our discussion, we focus on setups in which a phenomenon or a system is observed using multiple instruments, measurement devices or acquisition techniques. In this case, each acquisition framework is denoted as a *modality* and is associated with one dataset. The whole setup, in which one has access to data obtained from multiple modalities, is known as *multimodal*. A key property of multimodality is *complementarity*, in the sense that each modality brings to the whole some type of added value that cannot be deduced or obtained from any of the other modalities in the setup. In mathematical terms, this added value is known as *diversity*. Diversity allows to reduce the number of degrees of freedom in the system by providing constraints that enhance uniqueness, interpretability, robustness, performance, and other desired properties, as will be illustrated in the rest of this paper. Diversity can be found in a broad range of scenarios and plays a key role in a wide scope of mathematical and engineering studies. Accordingly, we suggest the following *operative* definition for the special type of diversity that is associated with multimodality:

Definition I.1: Diversity (due to multimodality) is the property that allows to enhance the uses, benefits and insights (such as those discussed in Section II), in a way that cannot be achieved with a single modality.

Diversity is the key to data fusion, as will be explained in Section III. Furthermore, in Section III, we demonstrate how a diversity approach to data fusion can provide a fresh new look on previously well-known and well-founded data and signal processing techniques.

As already noted, "data fusion" is quite a diffuse concept that takes different interpretations with applications and goals [8], [9], [20]. Therefore, within the context of this paper, and in accordance with the types of problems on which we focus, our emphasis is on the following tighter interpretation [21]:

Definition I.2: Data fusion is the analysis of several datasets such that different datasets can interact and inform each other.

This concept will be given a more concrete meaning in Sections III and V.

The *goal of this paper* is to provide some ideas, perspectives, and guidelines as to how to approach data fusion. This paper is not a review, not a literature survey, not a tutorial nor a cookbook. As such, it does not propose or promote any specific solution or method. On the contrary, our message is that whatever specific method or approach is considered, it should be kept in mind that it is just one among a very large set, and should be critically judged as such. In the same vein, any example in this paper should only be regarded as a concretization of a much broader idea.

How to read this paper? In order to make this paper accessible for readers with various interests and backgrounds, it is organized in two types of cross-sections. The first part (Sections II–III) deals with the question "*why?*", i.e., why we need data fusion. The second part (Sections IV–V) deals with the question "*how?*", i.e, how we perform data fusion. Each question is treated on two levels: *data* (Sections II and IV), and *theory* (Sections III and V). More specifically, Section II presents the concepts of multimodality and data fusion, and motivates them using examples from various applications. In Section III we introduce the concept of diversity as a key to data fusion, and give it a concrete mathematical formulation. Section IV discusses complicating factors that should be addressed in the actual processing of heterogeneous data. Section V gives some guidelines as to how to actually approach a data fusion problem from a model design perspective. Section VI concludes our work.

II. What is Multimodality? Why do we need multimodality?

For living creatures, multimodality is a very natural concept. Living creatures use external and internal sensors, sometimes denoted as "senses", in order to detect and discriminate among signals, communicate, cross-validate, disambiguate, and add robustness to numerous life-and-death choices and responses that must be taken rapidly, in a dynamic and constantly changing internal and external environment.

The well-accepted paradigm that certain natural processes and phenomena can express themselves under completely different physical guises is the *raison d'être* of multimodal data fusion. Too often, however, very little is known about the underlying relationships among these modalities. Therefore, the most obvious and essential endeavour to be undertaken in any multimodal data analysis task is exploratory: to learn about relationships between modalities, their complementarity,

shared vs. modality-specific information content, and other mutual properties.

In this section, we try to provide, by numerous practical examples, a more concrete sense to what we mean when we speak of "diversity" and "multimodality". The examples below illustrate the complementary nature of multimodal data, and as a result, some of the prominent uses, benefits and insights that can be obtained from properly exploiting multimodal data, especially as opposed to the analysis of single-set and single-modal data. They also present various complicating factors, due to which multimodal data fusion is not an evident task. The purpose of this section is to show that multimodality is already present in almost every field of science and technology, and thus it is of potential interest to everyone.

A. Multisensory Systems

Example II-A.1: Audio-Visual Multimodality. Audio-visual multimodality is probably the most intuitive, since it uses two of our most informative senses. Most human verbal communication involves seeing the speaker [18]. Indeed, a large number of audio-visual applications involve human speech and vision. In such applications, it is usually the audio channel that conveys the information of interest. It is well-known that audio and video convey complementary information. Audio has the advantage over video that it does not require line of sight. On the other hand, the visual modality is resistant to various factors that make audio and speech processing difficult, such as ambient noise, reverberations, and other acoustic disturbances.

Perhaps the most striking evidence to the amount of caution that needs to be taken in the design and use of multimodal systems is the "McGurk effect" [18]. In their seminal paper, McGurk and McDonald [18] have shown that presenting contradictory, or discrepant, speech ["ba"] and visual lip movements ["ga"], can cause a human to perceive completely different syllables ["da"]. These unexpected results have since been the subject of ongoing exploratory research on human perception and cognition [22, Section VI.A.5]. The McGurk effect serves as an indication that in real-life scenarios, data fusion can take paths much more intricate than simple summation of information. Not less important, it serves as a lesson that fusing modalities can yield undesired results and severe degradation of performance if the underlying relationships between modalities are not properly understood.

Nowadays, audio-visual multimodality is used for a broad range of applications [10], [23]. Examples include: *speech processing*, including speech recognition, speech activity detection, speech enhancement, speaker extraction and separation; *scene analysis*, for example tracking a speaker within a group, biometrics and monitoring, for safety and security applications [24]; *human-machine interaction (HMI)* [10]; *calibration* [25] [10, Section V.C]; and more.

Example II-A.2: Human-Machine Interaction. A domain that is heavily inspired by natural multimodality is HMI. In HMI, an important task is to *design* modalities that will make HMI as natural, efficient and intuitive as possible [11]. The idea is to combine multiple interaction modes based on audio-vision, touch, smell, movement (e.g., gesture detection and user tracking), interpretation of human language commands, and other multisensory functions [10], [11]. The principal point that makes HMI stand out among other multimodal applications that we mention is that, in HMI, the modalities are often interactive (as their name implies). Unlike other multimodal applications that we mention, not one but two very different types of systems (human and machine) are "observed" by *each other*'s sensors, and the goal of data fusion is not only to *interpret* each system's output, but also to actively *convey* information *between* these two systems. An added challenge is that this task should usually be accomplished in real-time. An additional complicating factor that makes multimodal HMI stand out is due to the fact that the human user often plays an active part in the *choice* of modalities (from the available set) and in the way that they are used in practice. This implies that the design of the multimodal setup and data fusion procedure must rely not only on the theoretically and technologically optimal combination of data streams but also on the ability to predict and *adapt* to the subjective cognitive preferences of the individual user. We refer to [11] (and references therein) for further discussion of these aspects.

B. Biomedical, Health

Example II-B.1: Understanding Brain Functionality. Functional brain study deals with understanding how the different elements of the brain take part in various perceptual and cognitive activities. Functional brain study largely relies on non-invasive imaging techniques, whose purpose is to reconstruct a high-resolution spatio-temporal image of the neuronal activity within the brain. The neuronal activity within the brain generates ionic currents that are often modelled as dipoles. These dipoles induce electric and magnetic fields that can be directly recorded by electroencephalography (EEG) and magnetoencephalography (MEG), respectively. In addition, neuronal activity induces changes in magnetization between oxygen-rich and oxygen-poor blood, known as the haemodynamic response. This effect, also called blood-oxygen-level dependent (BOLD) changes, can be detected by functional magnetic resonance imaging (fMRI). Therefore, fMRI is an *indirect* measure of neuronal activity. These three modalities register data at regular time intervals and thus reflect temporal dynamics. However, these techniques vary greatly in their spatio-temporal resolutions: EEG and MEG data provide high temporal [millisecond] resolution, whereas fMRI images have low temporal [second] resolution. fMRI data are a set of high-resolution 3D images, taken at regular time intervals, representing the whole volume of the brain of a patient lying in an fMRI scanner. EEG and MEG data are a set of time-series signals reflecting voltage or neuromagnetic field changes recorded at each of the (usually a few dozen of) electrodes attached to the scalp (EEG) or fixed within an MEG scanner helmet. The sensitivity of EEG and MEG to deep-brain signals is limited. In addition, they have different selectivity to signals as a function of brain morphology. Therefore, they provide data at much poorer spatial resolution and do not have access to the full brain volume. Consequently, the spatio-temporal in-

formation provided by EEG, MEG and fMRI is highly complementary. Functional imaging techniques can be complemented by other modalities that convey structural information. For example, structural magnetic resonance imaging (sMRI) and diffusion tensor imaging (DTI) report on the structure of the brain in terms of gray matter, white matter and cerebrospinal fluid. sMRI is based on nuclear magnetic resonance of water protons. DTI measures the diffusion process of molecules, mainly water, and thus reports also on brain connectivity. Each of these methods is based on different physical principles and is thus sensitive to different types of properties within the brain. In addition, each method has different pros and cons in terms of safety, cost, accuracy, and other parameters. Recent technological advances allow recording data from several functional brain imaging techniques simultaneously [26], [27], thus further motivating advanced data fusion.

It is a well-accepted paradigm in neuroscience that EEG and fMRI carry complementary information about brain function [26], [28]. However, their very heterogeneous nature and the fact that brain processes are very complicated systems that depend on numerous latent phenomena imply that simultaneously extracting useful information from them is not an evident task. The fact that there is no ground truth is reflected in the very broad range of methods and approaches that are being proposed [12], [15], [17], [21], [28]–[31]. Works on biomedical brain imaging often emphasize the exploratory nature of this task. Despite decades of study, the underlying relationship between EEG and fMRI is far from being understood [17], [29], [30], [32].

A well-known challenge in brain imaging is the EEG inverse problem. A prevalent assumption is that the measured EEG signal is generated by numerous current dipoles within the brain, and the goal is to localise the origins of this neuronal activity. Often formulated as a linear inverse problem, it is ill-posed: many different spatial current patterns within the skull can give rise to identical measurements [33]. In order to make the problem well-conditioned, additional hypotheses are required. A large number of solutions are based on adding various priors to the EEG data [34]. Alternatively, an identifiable and unique solution can be obtained using spatial constraints from fMRI [12], [22], [30].

Example II-B.2: Medical Diagnosis. Various medical conditions such as potentially malignant tumours cannot be diagnosed by a single type of measurement due to many factors such as low sensitivity, low positive predictive values, low specificity (high false-positive), a limited number of spatial samples (as in biopsy), and other limitations of the various assessment techniques. In order to improve the performance of the diagnosis, risk assessment and therapy options, it is necessary to perform numerous medical assessments based on a broad range of medical diagnostic techniques [35], [36]. For example, one can augment physical examination, blood-tests, biopsies, static and functional magnetic resonance imaging, with other parameters such as genetic, environmental and personal risk factors. The question of how to analyse all these simultaneously available resources is largely open. Currently, this task relies mostly on human medical experts. One of the main challenges is the automation of such decision procedures, in order to improve correct interpretation, as well as save costs and time [35].

Example II-B.3: Developing Non-Invasive Medical Diagnosis Techniques. In some cases, the use of multimodal data fusion is only a first step in the design of a single-modal system. In [37], the challenge is understanding the link between surface and intra-cardiac electrodes measuring the same atrial fibrillation event and the goal is eventually extracting relevant atrial fibrillation activity using only the non-invasive modality. For this aim, the intra-cardiac modality is exploited as a reference to guide the extraction of an atrial electrical signal of interest from non-invasive electrocardiography (ECG) recordings. The difficulty lies in the fact that the intra-cardiac modality provides a rather pure signal whereas the ECG signal is a mixture of the desired signal with other sources, and the mixing model is unknown.

Example II-B.4: Smart Patient Monitoring. Health monitoring using multiple types of sensors is drawing increasing attention from modern health services. The goal is to provide a set of non-invasive, non-intrusive, reasonable-cost sensors that allow the patient to run a normal life while providing reliable warnings in real-time. Here, we focus on monitoring, predicting and warning epileptic patients from potentially dangerous seizures [38]. The gold standard in monitoring epileptic seizures is combining EEG and video, where EEG is manually analysed by experts and the whole diagnostic procedure requires a stay of up to several days in a hospital setting. This procedure is expensive, time consuming, and physically inconvenient for the patient. Obviously, it is not practical for daily life. While much effort has already been dedicated to the prediction of epileptic seizures from EEG, with no clear-cut results so far, a considerable proportion of potentially lethal seizures are hardly detectable by EEG at all. Therefore, a primary challenge is to understand the *link* between epileptic seizures and additional body parameters: movement, breathing, heart-rate, and others. Due to the fact that epileptic seizures vary within and across patients, and due to the complex relations between different body systems, it is likely that any such system should rely on more than one modality [38].

C. Environmental Studies

Example II-C.1: Remote Sensing and Earth Observations. Various sensor technologies can report on different aspects of objects on Earth. Passive optical hyperspectral (resp. multispectral) imaging technologies report on material content of the surface by reconstructing its spectral characteristics from hundreds of (resp. a few) narrow (resp. broad) adjacent spectral bands within the visible range and beyond. A third type of an optical sensor is panchromatic imaging, which generates a monochromatic image with a much broader band. Typical spatial resolutions of hyperspectral, multispectral and panchromatic images are tens of meters, a few meters and less than one meter, respectively. Hence, there exists a trade-off between spectral and spatial resolution [39], [40] [13, Chapter 9]. Topographic information can be acquired from active sensors such as light detection and ranging (LiDAR)

and synthetic aperture radar (SAR). LiDAR is based on a narrow pulsed laser beam and thus provides highly accurate information about distance to objects, i.e., altitude. SAR is based on radio waves that illuminate a rather wide area, and the backscattered components reaching the sensor are registered; interpreting the reflections from the surface requires some additional processing with respect to (w.r.t.) LiDAR. Both technologies can provide information about elevation, three-dimensional structure of the observed objects, and their surface properties. LiDAR, being based on a laser beam, generally reports on the structure of the surface, although it can partially penetrate through certain areas such as forest canopy, providing information on the internal structure of the trees, for example. This ability is a mixed blessing, however, since it generates reflections that have to be accounted for. SAR and LiDAR use different electromagnetic frequencies and thus interact differently with materials and surfaces. As an example, depending on the wavelength, SAR may see the canopy as a transparent object (waves reach the soil under the canopy), semi-transparent (they penetrate in the canopy and interact with it) or opaque (they are reflected by the top of the canopy). Optical techniques are passive, which implies that they rely on natural illumination. Active sensors such as LiDAR and SAR can operate at night and in shaded areas [41].

Beyond the strengths and weaknesses of each technology w.r.t. the others, the use of each is limited by a certain inherent ambiguity. For example, hyperspectral imaging cannot distinguish between objects made of the same material that are positioned at different elevations, such as concrete roofs and roads. LiDAR cannot distinguish between objects with the same elevation and surface roughness that are made of different materials such as natural and artificial grass [42]. SAR images may sometimes be difficult to interpret due to their complex dependence on the geometry of the surface [41].

In real-life conditions, interpretability of the observations of one modality may be difficult without additional information. For example, in hyperspectral imaging, on a flat surface, reflected light depends on the abundance (proportion of a material in a pixel) and on the endmember (pure material present in a pixel) reflectance. In a non-flat surface, the reflected light depends also on the topography, which may induce variations in scene illumination and scattering. Therefore, in non-flat conditions, one cannot accurately extract material content information from optical data alone. Adding a modality that reports on the topography, such as LiDAR, is necessary to resolve spectra accurately [43].

As an active initiative, we point out the yearly data fusion contest of the IEEE Geoscience and Remote Sensing Society (GRSS) (see dedicated paper in this issue [44]). Problems addressed include *multi-modal change detection*, in which the purpose is to detect changes in an area before and after an event (a flood, in this case), given SAR and multispectral imaging [45], using either all or part of the modalities; *multi-modal multi-temporal data fusion* of optical, SAR and LiDAR images taken at different years over the same urban area [41], where suggested applications include assessing urban density, change detection and overcoming adverse illumination conditions for optical sensors; and *proposing new methods for fusing* hyperspectral and LiDAR data of the same area, e.g., for improved classification of objects [42].

Example II-C.2: Meteorological Monitoring. Accurate measurements of atmospheric phenomena such as rain, water vapour, dew, fog and snow are required for meteorological analysis and forecasting, as well as for numerous applications in hydrology, agriculture and aeronautical services. Data can be acquired from various devices such as rain gauges, radars, satellite-borne remote sensing devices (see Example II-C.1), and recently also by exploiting existing commercial microwave links [46]. *Rain gauges*, as an example, are simply cups that collect the precipitation. Albeit the most direct and reliable technique, their small sampling area implies very localized representativeness and thus poor spatial resolution (e.g., [46], [47]). Rain gauges may be read automatically at intervals as short as seconds. *Satellites* observe Earth at different frequencies, including visible, microwave, infrared, and shortwave infrared to report on various atmospheric phenomena such as water vapour content and temperature. The accuracy of *radar* rainfall estimation may be affected by topography, beam effects, distance from the radar, and other complicating factors. Radars and satellite systems provide large spatial coverage; however, they are less accurate in measuring precipitation at ground level (e.g., [48]). *Microwave links* are deployed by cellular providers for backhaul communication between base stations. The signals transmitted by the base stations are influenced by various atmospheric phenomena (e.g., [49]), primarily attenuation due to rainfall [46], [47]. These changes in signal strength are recorded at predefined time intervals and kept in the cellular provider's logs. Hence, the precipitation data is in fact a "reverse engineering" of this information. The microwave links' measurements provide average precipitation on the entire link and close to ground level [46]. Altogether, these technologies are largely complementary in their ability to detect and distinguish between different meteorological phenomena, spatial coverage, temporal resolution, measurement error, and other properties. Therefore, meteorological data are often combined for better accuracy, coverage and resolution; see, e.g., [19], [47], [48] and references therein.

Example II-C.3: Cosmology. A major endeavour in astronomy and astrophysics is understanding the formation of our Universe. Recent results include robust support for the six-parameter standard model of cosmology, of a Universe dominated by Cold Dark Matter and a cosmological constant Λ, known as ΛCDM [50], [51]. The purpose of ongoing and planned sky surveys is to decrease the allowable uncertainty volume of the six-dimensional ΛCDM parameter space and to improve the constraints on the other cosmological parameters that depend on it [51]. The goal is to validate (or disprove) the standard model.

A major difficulty in astrophysics and cosmology is the absence of ground truth. This is because cosmological processes involve very high energies, masses, large space and time scales that make experimental study prohibitive. The lack of ground truth and experimental support implied that, from its very beginning, cosmological research had to rely on cross-validation of outcomes of different observations, numerical

simulations and theoretical analysis; in other words, *data fusion*. A complicating factor associated with this task is the fact that in many types of inferences, for all practical purposes, we have only one realization of the Universe. This means that even if we make statistical hypotheses about underlying processes, there is still only one sample. This fact induces an uncertainty called "cosmic variance" that cannot be accommodated by improving the measurement precision.

Despite its simplicity, the ΛCDM model has proved to be successful in describing a wide range of cosmological data [52]. In particular, it is predicted that its six parameters can fully explain the angular power spectra of the temperature and polarization fluctuations of the cosmic microwave background radiation (CMB). Therefore, since the first experimental discovery of the CMB in 1965 [53], there has been an ongoing effort to obtain better and more accurate measurements of these fluctuations.

A severe problem in validating the ΛCDM model from CMB observations is known as "parameter degeneracy". Although the CMB power spectrum can be fully explained by the standard model, this relationship is not unique in the sense that the same measured CMB power spectrum can be explained by other models, not only ΛCDM. These degeneracies can be broken by combining CMB observations with other cosmological data. While CMB corresponds to photons released about 300,000 years after the Big Bang, the same parameters that controlled the evolution of the early Universe continue to influence its matter distribution and expansion rate to our very days. Therefore, other measures, such as redshift from certain types of supernovae, angular and radial baryon acoustic oscillation scales that can be derived from galaxy surveys, galaxy clustering [54], [55], and stacked gravitational lensing, also serve as important cosmological probes [51], [52]. Since the cosmological parameters that determine the evolution of the early Universe are the same as those that control high-energy physics, cosmological observations are fused and cross-validated with experimental outcomes such as the Large Hadron Collider Higgs data [56].

III. Multimodality as a Form of Diversity

In this section, we discuss data fusion from a theoretical perspective. In order to contain our discussion, we focus on data-driven methods. Within these, we restrict our examples to a class of problems known as blind separation, and within these, to data and observations that can be represented by (multi-) linear relationships. Reasons are as follows. First, by definition, data-driven models may be useful to numerous applications, as will be explained in Section III-B. Second, there exist much established theory and numerous models that fit into this framework. Third, it is impossible to cover all types of models. Still, the ideas that these examples illustrate go far beyond these specific models.

A key property in any analytical model is uniqueness. Uniqueness is necessary in order to achieve interpretability, i.e., attach physical meaning to the output [2], [5]. In order to establish uniqueness, all blind separation problems invariably rely on one or more types of diversity [57]: concrete mathematical examples will be given in Section III-C1. In the sequel, we show how the concept of "diversity" plays part, under different guises, in data fusion. In particular, we show that multimodality can provide a new form of diversity that can achieve uniqueness even in cases that are otherwise non-unique.

The rest of this section is as follows. Section III-A presents some basic mathematical preliminaries that will serve us to provide a more concrete meaning to the ideas that we lay out in the rest of this work. Section III-B explains the model-driven vs. the data-driven approach, and motivates the latter. Section III-C discusses diversity and data fusion in datasets that are stacked in a single matrix or a higher-order array, also known as a tensor. In Section III-D, we go beyond single-array data analysis, and establish the idea of "a link between datasets as a new form of diversity" as the key to advanced data fusion. We conclude our claims and summarize these ideas in Section III-E.

A. Mathematical Preliminaries

In a large number of applications, one is interested in extracting knowledge from the data. In real-life scenarios, each observation or measurement often consists of contributions from multiple sources. These can be divided into sources of interest, which carry valuable information, and other sources, which do not carry any information of interest. The latter type of contribution is sometimes referred to as noise, or interference, depending on the scenario and context.

Consider one point x in the measurement space. We can approximate it as (we write equality but we mean that we attribute a certain model to it)

$$x = f(\mathbf{z}) \,, \tag{1}$$

where $\mathbf{z} = \{z_1, \ldots, z_V\}$ is the ensemble of points in the latent variable space. These could be signals, parameters or any other elements that contribute to the observation x, and f represents the corresponding transformation (e.g., channel effects). We are interested in scenarios where $\mathbf{z}$ is unknown, and in addition, cannot be observed directly without the intermediate transformation f. In certain scenarios, also f is unknown. We denote all the unknown elements of the model as "latent variables".

Perhaps the first and most obvious interpretation of (1) is an inverse problem, where the goal is to obtain an estimate as precise as possible of $\mathbf{z}$ and f given x. Recovering f and $\mathbf{z}$ can also be regarded as finding the *simplest* set of variables that *explains* the observations [5, Section I]. This interpretation particularly corresponds to exploratory research. In addition, and especially when the number of observations is large w.r.t. the size of $\mathbf{z}$, recovering the *smallest*-size $\mathbf{z}$ that best explains the observations can be regarded as a form of *compression*, which can be particularly useful in large-scale data scenarios. It is clear that in order to solve (1), one needs a sufficient number of constraints in order to (over-) determine the problem, i.e., constrain the number of degrees of freedom such that the problem is well-posed.

In the rest of this paper, we use standard mathematical notations. Scalars, vectors, matrices, and higher-order arrays

(tensors) are denoted as a, $\mathbf{a}$, $\mathbf{A}$, and $\mathcal{A}$, respectively. The dimensions of an Nth-order array (tensor) are $I_1 \times I_2 \times \cdots \times I_N$, where $N = 1, 2, 3, \ldots$ implies a vector, matrix or higher-order array (tensor), respectively. $(\cdot)^\top$ denotes transpose or conjugate transpose, where the exact interpretation should be understood from the context.

B. Data-Driven vs. Model-Driven Methods

Roughly, and for the sake of the discussion that follows, approaches to the problem in Section III-A can be divided into two groups: model driven, and data driven. Model-driven approaches rely on an explicit realistic model of the underlying processes [27, Section 3.3] [12], [58], and are generally successful if the assumptions are plausible and the model holds. However, model-driven methods may not always be the best choice, for example, when the underlying model of the signals or the medium in which they propagate is too complicated, varying rapidly, or simply unknown.

In the context of multimodal datasets that are generated by complex systems as those mentioned in Sections I–II, very little is known about the underlying relationships between modalities. The interactions between datasets and data types are not always known or sufficiently understood. Therefore, we focus on and advocate a data-driven approach. In practice, this means making the fewest assumptions and using the simplest models, both within and across modalities [5]. "Simple" means, for example, linear relationships between variables, avoiding model-dependent parameters, and/or use of model-independent priors such as sparsity, non-negativity, statistical independence, low-rank, and smoothness, to name a few. As its name implies, a data-driven approach is self-contained in the sense that it relies only on the observations and their assumed model: it avoids external input [5]. For this reason, and especially in the signal-processing community, data-driven methods are sometimes termed "blind". In the rest of this section, we give a more concrete meaning to these ideas.

Data-driven methods, both single-modal and multimodal, have already proven successful in a broad range of problems and applications. A non-comprehensive list includes astrophysics [59], biomedics [60], telecommunications [61], audio-vision [23], chemometrics [62], and more. For further examples see e.g., [63]–[65] and references therein, as well as the numerous models mentioned in the rest of this paper.

In the rest of this section, we discuss and explain the role of diversity in achieving uniqueness in data-driven models. In particular, we demonstrate how the presence of multiple data sets can be exploited as a new form of diversity.

C. Diversity in Single Matrix or Tensor Decomposition Models

Earlier in this section, we argued that diversity has a key role in achieving uniqueness of analytical models. We now give a concrete mathematical meaning to this statement, by way of examples from signal processing, linear and multilinear algebra. We begin by discussing diversity in datasets that can be stacked in a single array, be it a matrix or a higher-order array.

1) Diversity in Matrix Decomposition Models: Perhaps the most simple yet useful implementation of (1) is

$$x = \sum_{r=1}^{R} a_r b_r \ . \tag{2}$$

In many applications, model (2) is generalized as

$$x_{ij} = \sum_{r=1}^{R} a_{ir} b_{jr} \tag{3}$$

where $i = 1, \ldots, I$, $j = 1, \ldots, J$. An often-used interpretation is that x_{ij} is a linear combination of R signals $b_{j1}, \ldots, b_{jR}$ impinging on sensor i at sample index j, with weights $a_{i1}, \ldots, a_{iR}$. Eq. (3) can be rewritten in matrix form as

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{a}_r \mathbf{b}_r^\top = \mathbf{A}\mathbf{B}^\top \tag{4}$$

such that x_{ij} is the (i,j)th entry of $\mathbf{X} \in \mathbb{K}^{I \times J}$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, and similarly for $\mathbf{A} \in \mathbb{K}^{I \times R}$ and $\mathbf{B} \in \mathbb{K}^{J \times R}$. The rth column vectors of $\mathbf{A}$ and $\mathbf{B}$ are $\mathbf{a}_r = [a_{1r}, \ldots, a_{Ir}]^\top$ and $\mathbf{b}_r = [b_{1r}, \ldots, b_{Jr}]^\top$, respectively.

The model in (4) provides I linear combinations of the columns of $\mathbf{B}$ and J linear combinations of the columns of $\mathbf{A}$ [57]. In the terminology of [57], $\mathbf{X}$ provides I-fold *diversity* for $\mathbf{B}$ and J-fold *diversity* for $\mathbf{A}$. Unfortunately, these types of diversity are generally insufficient to retrieve the underlying factor matrices $\mathbf{A}$ and $\mathbf{B}$. For any $R \times R$ invertible matrix $\mathbf{T}$, it always holds that

$$\mathbf{X} = \mathbf{A}\mathbf{B}^\top = (\mathbf{A}\mathbf{T}^{-1})(\mathbf{T}\mathbf{B}^\top) \, . \tag{5}$$

Hence, the pairs $(\mathbf{A}\mathbf{T}^{-1}, \mathbf{T}\mathbf{B}^\top)$ and $(\mathbf{A}, \mathbf{B}^\top)$ have the same contribution to the observations $\mathbf{X}$ and thus cannot be distinguished. Consequently, one cannot uniquely identify the rank-1 terms $\mathbf{a}_r \mathbf{b}_r^\top$ unless $R \leq 1$ [66, Lemma 4i]. We refer to this matter as the *indeterminacy problem*. A prevalent approach is to reduce $\mathbf{T}$ to a unitary matrix using a simplifying assumption that the columns of $\mathbf{B}$ are decorrelated. In such cases, the indeterminacy (5) is referred to as the *rotation problem* [67, Section 4] [2], [5]. Conversely, even if the rank-1 terms are known, it is clear from (4) that they can be uniquely characterized, at most, up to $(\alpha_r \mathbf{a}_r)(\beta_r \mathbf{b}_r)^\top = \mathbf{a}_r \mathbf{b}_r^\top$, $\alpha_r \beta_r = 1$, if $R \leq \min(I, J)$. The latter amounts to $\mathbf{T} = \mathbf{P}\mathbf{\Lambda}$, where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ is diagonal and invertible. The presence of $\mathbf{P}$ implies that the indexing $1, \ldots, R$ is arbitrary. This indeterminacy is inherent to the problem and thus *inevitable*. If all decompositions yield the same rank-1 terms then we say that the model is *unique*. The fact that the factorization of a matrix into a product of several matrices is generally *not* unique for $R > 1$ unless additional constraints are imposed is well-known [66, Section 3].

We now discuss approaches to fix the indeterminacy in (5). In a general algebraic context, matrix factorizations such as singular value decomposition (SVD) and eigenvalue decomposition (EVD) are made unique by imposing orthogonality on the underlying matrices and inequality on the singular or eigenvalues [66, Section 3] [68]. Such constraints are convenient mathematically but usually physically implausible

since they yield non-interpretable results [69]. It is thus desirable to find other types of constraints that allow for better representation of the natural properties of the data.

Depending on the application, the matrix factorization model in (4) may be interpreted in different ways that give rise to different types of constraints. When the model in (4) is used to analyse data, it is sometimes termed factor analysis (FA) [70]. In the signal processing community, when the columns of $\mathbf{B}$ represent signal samples and the goal is to recover these signals given only the observations $\mathbf{X}$, model (4) is commonly associated with the blind source separation (BSS) problem [63], [71]. The goal of FA and BSS is to represent $\mathbf{X}$ as a sum of low-rank terms with interpretable factors [65], where the difference lies in the type of assumptions being used.

In FA, one approach to fixing the indeterminacy (5) is by imposing external constraints [5, Section I]. This is not a data-driven approach and is thus excluded from our discussion. A data-driven approach to FA is to use physically-meaningful constraints on the factor matrices that reduce the number of degrees of freedom. For example, a specific arrangement of a receive antenna array or other properties of a communication system may be imposed via a Vandermonde [72]–[75] or Toeplitz [76] structure. Alternatively, a factor may reflect a specific signal type such as constant modulus or finite alphabet [57], [61]. Another approach is to use sparsity [77]–[79].

Probably the most well-known BSS approach to fix the indeterminacy in (5) is independent component analysis (ICA). ICA is more commonly formulated as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)\ , \quad t = 1, \dots, T \tag{6}$$

where $\mathbf{s}(t) = [s_1(t), \dots, s_R(t)]^\top \in \mathbb{K}^{R\times 1}$ is a vector of R statistically independent random processes known as "sources", and $\mathbf{x}(t) \in \mathbb{K}^{I\times 1}$ their observations. $\mathbf{A}$ is full column rank. The link with (4) is established via $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(T)]$, $J = T$, and $\mathbf{B}^\top = [\mathbf{s}(1), \dots, \mathbf{s}(T)]$ such that the R columns of $\mathbf{B}$ represent samples from the R statistically independent random processes. ICA uses the "spatial diversity" provided by an array of sensors, which amounts to the I-fold diversity for $\mathbf{B}$ mentioned before, together with an assumption of statistical independence on the sources, in order to obtain estimates of $\mathbf{s}(t)$ whose entries are as statistically independent as possible. This amounts to fixing the indeterminacy (5). Under these assumptions, separation can be achieved if the statistically independent sources are non-stationary, non-white, or non-Gaussian [71], [80]–[82]. The first two can be interpreted as *diversity* across time or *diversity* in the spectral domain: the sources must have different nonstationarity profiles or power spectra [81, Section 6]. Non-Gaussianity is associated with diversity in higher-order statistics (HOS). A plethora of methods has been devised to exploit this diversity [63], [80], [83]–[86], and the matter is far from being exhausted.

Both FA and ICA have been used for decades and with much success to analyse a very broad range of data, their success being much due to the simplicity of their basic idea and the fact that very robust algorithms exist that yield satisfying results. Therefore, they are at the focus of our discussion. It should be kept in mind, however, that in practice, many observations can be *better* explained by other types of underlying models that are not limited to decomposition into a sum of rank-1 terms, statistical independence, linear relationships, or even matrix factorizations. Other properties that are often used to achieve uniqueness, improve numerical robustness and enhance interpretability are, for example, non-negativity, sparsity, and smoothness [63]. Proving uniqueness for these types of factorizations is a matter of ongoing research.

> Any type of constraint or assumption on the underlying variables that helps achieve essential uniqueness can be regarded as a "diversity".

2) Going up to Higher-Order Arrays:: In Section III-C1, we have seen that the two linear types of diversity that are present in the rows and columns of $\mathbf{X}$ are not sufficient in order to obtain a unique matrix factorization. We saw that uniqueness *can* be established by imposing sufficiently strong constraints on the factor matrices $\mathbf{A}$ and $\mathbf{B}$ in (4). An alternative approach is to enrich the observational domain, *without* constraining the factor matrices. For example, if the two linear diversities given by the two-dimensional array $\mathbf{X}$ are interpreted as spatial and temporal, it is possible to obtain uniqueness by adding a third diversity in the frequency domain, without imposing constraints on the factor matrices. We now explain how this can be done.

The two-way model (4) can be generalized by extending (3) to

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \tag{7}$$

with $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, K$. These observations can be collected into a three-way array (third-order tensor) with dimensions $I \times J \times K$,

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{8}$$

whose (i,j,k)th entry is x_{ijk}. $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_R] \in \mathbb{K}^{I\times R}$, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_R] \in \mathbb{K}^{J\times R}$ and $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_R] \in \mathbb{K}^{K\times R}$ are matrices whose column vectors are $\mathbf{a}_r$, $\mathbf{b}_r$ and $\mathbf{c}_r = [c_{1r}, \dots, c_{Kr}]^\top$, respectively. Here, $\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \in \mathbb{K}^{I\times J\times K}$ is an outer product of three vectors and thus is a rank-1 term. Its (i,j,k)th entry is $a_{ir}b_{jr}c_{kr}$. When (8) holds and is irreducible in the sense that R is minimal, it is sometimes referred to as the canonical polyadic decomposition (CPD) [4], [87]. Note that (4) can be rewritten as $\mathbf{X} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r$.

In striking difference to (5), the pair $\{(\mathbf{A}, \mathbf{B}, \mathbf{C}), (\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\mathbf{C}})\}$ has the same triple product (8) if and only if there exists an $R \times R$ permutation matrix $\mathbf{P}$ and three diagonal matrices $\mathbf{\Lambda}_A, \mathbf{\Lambda}_B, \mathbf{\Lambda}_C$ such that

$$\overline{\mathbf{A}} = \mathbf{A}\mathbf{P}\mathbf{\Lambda}_A\ ,\ \overline{\mathbf{B}} = \mathbf{B}\mathbf{P}\mathbf{\Lambda}_B\ ,\ \overline{\mathbf{C}} = \mathbf{C}\mathbf{P}\mathbf{\Lambda}_C \quad \text{and} \quad \mathbf{\Lambda}_A\mathbf{\Lambda}_B\mathbf{\Lambda}_C = \mathbf{I}_R \tag{9}$$

even for $R > 1$, under very mild constraints on $\mathbf{A}, \mathbf{B}, \mathbf{C}$ [66], [67], [88]. Eq. (9) can be reformulated as $\mathcal{X} = \sum_{r=1}^{R} (\alpha_r \mathbf{a}_r) \circ$

$(\beta_r \mathbf{b}_r) \circ (\gamma_r \mathbf{c}_r)$ $\forall \alpha_r \beta_r \gamma_r = 1$. If a three-way array is subject only to these trivial indeterminacies (alternatively: if all CPDs yield the same rank-1 terms) then we say that it is (essentially) unique.

> The key difference between matrix and tensor factorizations is that CPD is inherently "essentially unique" up to a scaled *permutation* matrix, whereas in the bilinear case the indeterminacy is an arbitrary non-singular matrix.

The uniqueness of the CPD becomes even more pronounced when it is joined with the fact that it holds also for $R > \max(I, J, K)$ [67], [89]. This is in contrast to FA, where it holds only for $R \leq \min(I, J)$. The immediate outcome is that underdetermined cases of "more sources than sensors" can be handled straightforwardly. In addition, the factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ need not be full column rank [67], [89] [90, Theorem 2.2], see Example III-D.2. *Upper* bounds on R have first been derived by [88], [89]. These results have later been extended to higher-order arrays, where "order" indicates the number N of indices $x_{ijk\ldots}$ and $N \geq 3$ [57], [72], [91]. Recently, more relaxed bounds that guarantee uniqueness for larger R have been derived; see e.g., [92]–[97] and references therein.

In analogy to (4), the three-way array $\mathcal{X}$ provides three modes of linear diversity. It contains JK linear combinations of the columns of $\mathbf{A}$, IK of $\mathbf{B}$ and IJ of $\mathbf{C}$ [57]. The fact that there exist multiple linear relationships within the model gives it the name "multilinear". As argued by [57], in many real-life scenarios, often there exist $N \geq 3$ linear types of diversity that admit the multilinear decomposition (8) and thus guarantee uniqueness without any further assumptions. For example, in direct-sequence code-division multiple access (DS-CDMA) communication systems, one may exploit (spatial $\times$ temporal $\times$ spreading code) [57] or (sensor $\times$ polarization $\times$ source signal) types of diversity; in psychometrics, (occasions $\times$ persons $\times$ tests) [70] or (observations $\times$ scores $\times$ variables) [98]; in chemometrics and metabolomics, (sample $\times$ frequency $\times$ emission profile $\times$ excitation profile) [8], [62], [99]; in polarized Raman spectroscopy, (polarization $\times$ spatial diversity $\times$ wavenumber) [100]; in EEG, (time $\times$ frequency $\times$ electrode) [101]–[103]; and in fMRI, (voxels $\times$ scans $\times$ subjects) [104].

> Each type of constraint, structural (i.e., on the factor matrices) or observational (i.e., any of the non-degenerate modes of a matrix or a higher-order array), that contributes to the unique decomposition and thus to the identifiability of the model, and *cannot be deduced from the other constraints*, i.e., is "disjoint" [16], can be regarded as a "diversity". In particular, each observational mode in the Nth order tensor (8) is a "diversity". Hence, a tensor order corresponds to the number of types of (observational) diversity [57], [61].

The explicit link between tensor order as a diversity and data fusion has been made in [16]. The fact that we can now associate "diversity" with well-defined mathematical properties of an analytical model implies that we can now link results on uniqueness, identifiability, and performance with the number of types of diversity that this model involves. Hence, the contribution of each "diversity" to the model can now be characterized and quantified [57], [82].

An application of this idea is the question raised in [57] as to how the number of types of observational diversity, i.e., tensor order $N \geq 3$, contributes to the identifiability. To answer this question, it is shown that as N increases, indeed the bound on the number of rank-1 terms that can be uniquely identified becomes more relaxed. In other words, more observational modes allow to identify more sources in the same setup. Hence, this is *a proof that increasing observational diversity improves identifiability*. This is an example how questions regarding multimodality and diversity are a stimulus for new mathematical and theoretical insights.

Until now, we have looked at N-way arrays as a way to represent simultaneously N (multi-) linear types of diversity. An interesting link with the matrix factorization problem in Section III-C1 is achieved if we look at an N-way array as a structure that stores $(N-1)$-way arrays by stacking them along the Nth dimension. As noted, e.g., by [4], [5], [70], [89], [105], the CPD can be thought of as a generalization of FA, as follows. Let

$$\mathbf{X}_k = \mathbf{A}\mathbf{\Lambda}_k\mathbf{B}^\top, \quad k = 1, \ldots, K \tag{10}$$

denote K instances of the FA problem (4) where the diagonal $R \times R$ matrix $\mathbf{\Lambda}_k = \text{diag}\{c_{k1}, \ldots, c_{kR}\}$ can be regarded as a scaling of the rows of $\mathbf{B}$. It can be readily verified that stacking the K matrices $\mathbf{X}_k$ in parallel along the third dimension results in (8). As we already know from (9), the rotation problem is eliminated [89]. It is thus no surprise that the tensor decomposition (8) is also known as parallel factor analysis (PARAFAC) [5]. Combining this observation with the perspective of data fusion, it has been noted that a tensor decomposition can be regarded as a way to fuse and jointly analyse data of multiple observations when all the datasets have the same size and share the same type of decomposition [16]. Note that this notion applies also to two-way arrays. For example, if we associate a BSS interpretation to the model in (4), the ith row can be regarded as the contribution of the ith sensor, and stacking all I observations yields the $I \times J$ observation matrix $\mathbf{X}$ [16].

Model (10) can be linked not only to FA but also to BSS, as follows. In Section III-C1, we mentioned that uniqueness of BSS can be achieved if the sources are non-Gaussian, non-stationary, or non-spectrally-flat (i.e., coloured). These properties can be reformulated algebraically as a symmetric joint diagonalization (JD) of several matrices [81], [83], i.e., a special case of (10) when $\mathbf{A} = \mathbf{B}$. As we have just explained, JD can be interpreted as a simple data fusion problem in which several datasets share the same mixing matrix. A key point is that diagonalization of a single matrix has an infinite number of solutions, and each of these "non-properties" [81] provides a set of at least two matrices that can be jointly diagonalized, thus fixing the indeterminacies.

The discussion in this section implies that if we can represent our observations in terms of $N \geq 3$ linear types of diversity or stack multiple datasets in an Nth-order tensor then we may benefit from the following powerful properties:

Why are tensor decompositions useful for data fusion?
(1) The model for $R \geq 1$ rank-1 terms is identifiable: The exact maximal number of identifiable rank-1 terms is generally unknown, though bounds that depend on various properties of the factor matrices exist.
(2) Under-determined mixtures are identifiable: identification of $R \geq 1$ rank-1 terms even for "more sources than sensors" cases.
(3) Factor matrices need not be full rank: identifiability of $R \geq 1$ rank-1 terms even if no factor matrix $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}, \ldots$, is of full rank.
(4) Rank-1 terms are identifiable up to permutation: when a tensor decomposition is interpreted as joint analysis of lower-order tensors, the arbitrary individual permutation that arises if each decomposition is done separately becomes common to all decompositions.
(5) Increasing N allows uniqueness for higher R: more types of observational diversity allow to resolve more latent sources.
(6) There is no need for structural constraints or assumptions such as statistical independence, non-negativity, sparsity, or smoothness in order to achieve a unique decomposition. And yet, multilinear structures readily admit such additional types of diversity that can further contribute to interpretability, robustness, uniqueness, and other desired properties; see end of Section III-D for examples.

More properties of tensor decompositions and their uses in various engineering applications can be found for example in [64], [65], [106] and references therein.

Concluding Section III-C, Sections III-C1 and III-C2 presented two ways to look at matrices or tensors as data fusion structures. We have shown that matrix or tensor decompositions provide a natural framework to incorporate multiple types of observational diversity [16] on top of structural ones. We have shown that matrices and higher-order tensors can be regarded as ways to jointly analyse multiple observations of the same data, when datasets share the same underlying structure [16]. It is thus no surprise that many multimodal data fusion models use matrix or tensor decompositions as their underlying analytical engine.

Until now, we focused on decompositions in sum of rank-1 factors and statistical independence. In fact, these constraints can be regarded as too strong. Indeed, there exist other factorizations that may represent more flexible underlying relationships; see end of Section III-D for examples. It is only for the sake of simplicity and limited space that we restrict our discussion to one type of decomposition.

D. A Link Between Datasets as a New Form of Diversity

As explained in Section III-C, if all datasets share the same underlying factorization model, and in addition, admit a (multi-) linear relationship, then it may be possible to use a single matrix or tensor decomposition in order to perform data fusion. This assumption may be challenged in various scenarios. An obvious conflict arises when datasets are given in different types of physical units. A technical difficulty is when datasets are stacked in arrays of different orders, such as matrices vs. higher-order arrays. Further examples are datasets with different latent models, different types of uncertainty, or when not all factors or latent variables are shared by all datasets. In such cases, we say that datasets are *heterogeneous* [8]. While each of these complicating factors may be accommodated by preprocessing the datasets such that they all comply, e.g., by normalizing, realigning, interpolating, up- or down-sampling, using features, or reducing dimensions, these procedures have the risk of being lossy in various respects (for further discussion on complicating factors in data fusion, see Section IV). For these reasons, more elaborate models that allow heterogeneous datasets to remain in their most explanatory form and still perform true data fusion, i.e., in the sense of Definition I.2 and Section V-A, have been devised.

In the following, we discuss data fusion approaches that go beyond single matrix or tensor factorization. Our emphasis is on demonstrating how the concepts of true data fusion allow pushing even further the limits of extracting knowledge from data that were summarized in Section III-C2. We show how these properties are carried over to more elaborate data fusion models and how they can be reinforced into stronger properties that cannot be achieved using single-set single-modal data. In particular, (i) allowing *more relaxed uniqueness conditions* that admit more challenging scenarios: for example, more relaxed assumptions on the underlying factors, and the ability to resolve more latent variables (low-rank terms) in each dataset, and (ii) terms that are shared across datasets enjoy the *same permutation at all datasets*. This obviates the need for an additional step of identifying the arbitrarily-ordered outputs of each individual decomposition and matching them, a task that generally cannot be accomplished without additional information, in a blind or data-driven context. Fixing the permutation reduces the number of degrees of freedom and thus enhances performance and interpretability. The following examples illustrate these points.

Example III-D.1: Coupled Independent Component Analysis. Consider the ICA problem (6). It is well-known that statistically independent real-valued Gaussian processes with independent and identically distributed (i.i.d.) samples, mixed by an invertible $\mathbf{A}$, cannot be blindly separated based on their observations $\mathbf{x}(t)$ alone [71], [107]. If *several* such datasets are considered simultaneously, however, without changing the model *within* each mixture, but allowing statistical dependence *across* datasets, then a unique and identifiable solution to all these mixtures, up to unavoidable scale and permutation ambiguities, exists [82]. This model, when not restricted to Gaussian i.i.d. samples, is known as independent vector

analysis (IVA) [82], [108], [109] and can be solved using second-order statistics (SOS) alone [110], [111].

IVA was originally proposed to separate convolutive mixtures of audio signals [108], [109]. In the frequency domain, this amounts (approximately) to resolving M ICA mixtures (6),

$$\mathbf{x}^{(m)}(t) = \mathbf{A}^{(m)}\mathbf{s}^{(m)}(t)\ ,\ t = 1, \ldots, T\ , \tag{11}$$

where the M matrices $\mathbf{A}^{(m)}$, $m = 1, \ldots, M$, are generally different (in this context, t denotes samples in the frequency domain and m are the frequency bins). For simplicity, we assume that both $\mathbf{x}^{(m)}(t)$ and $\mathbf{s}^{(m)}(t)$ are $I \times 1$. When each mixture (11) is solved separately, it is associated with an individual permutation matrix $\mathbf{P}^{(m)}$. It is clear that proper separation and reconstruction of the I audio signals cannot be achieved if the elements of the same source at different frequency bins are not properly matched. The key point in IVA w.r.t. a collection of ICA is that it exploits statistical dependence among latent sources that belong to different mixtures, as illustrated in Figure 1. Under certain conditions, the IVA framework provides a single $R \times R$ permutation matrix $\mathbf{P}^{(m)} = \mathbf{P}$ that applies to all the involved mixtures [82], [109].

The ability of IVA to obviate the need to match the outputs of M separate ICA soon turned out useful far beyond convolutive mixtures: it has since been applied to fMRI group data analysis [112], [113], multimodal fusion of several brain-imaging modalities [114], and the analysis of temporal dynamic changes [115]. IVA extends CCA [1] and its multiset extension (MCCA) [3], which have both been widely used for fusion [31], [36], [58], [116]–[118], to the case where not only second-order statistics but all-order statistics are taken into account [82]. Recently, a generalization of IVA that allows decomposition into terms of rank larger than one has been proposed [119]–[121]. In addition, since IVA is a generalization of ICA, it readily accommodates additional types of diversity such as coloured (i.e., non-spectrally-flat) or non-stationary sources [111], [122] (recall Section III-C1). Identifiability analysis of the multiple types of diversity in IVA is given in [82], [123]. It should be noted that the uniqueness results for coupled CPD [90] (Example III-D.2) require at least one tensor of order larger than two in the coupled set and thus they cannot be applied to IVA.

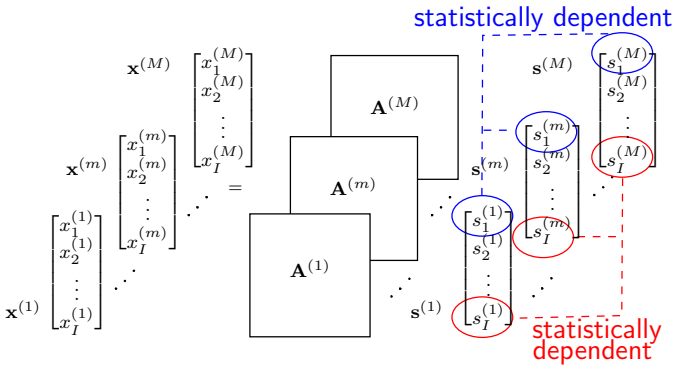

Fig. 1: Diagram of the IVA model. Figure reproduced from [116, Figure 1].

Example III-D.2: Coupled Tensor Decompositions. In multilinear algebra, an ongoing endeavour is to obtain uniqueness conditions on a tensor decomposition [67], [88], [92]–[97]. The goal is to derive bounds that are as relaxed as possible on the largest R that still satisfies essential uniqueness (9). As an example, two *necessary* conditions for the essential uniqueness of the CPD of a third-order tensor (8) are that

$$\begin{aligned}&(\mathbf{A} \odot \mathbf{B})\ ,\ (\mathbf{C} \odot \mathbf{A}) \text{ and } (\mathbf{B} \odot \mathbf{C}) \text{ have full column rank,}\\ &\text{and} \quad \min(k_{\mathbf{A}}, k_{\mathbf{B}}, k_{\mathbf{C}}) \geq 2\end{aligned} \tag{12}$$

(e.g., [72], [92]) where $\odot$ denotes the column-wise Khatri-Rao product and $k_{\mathbf{A}}$ is the Kruskal-rank of matrix $\mathbf{A}$, equal to the largest integer $k_{\mathbf{A}}$ such that every subset of $k_{\mathbf{A}}$ columns of $\mathbf{A}$ is linearly independent [67].

Consider now M third-order tensors $\mathcal{X}^{(m)} \in \mathbb{C}^{I_m \times J_m \times K}$, $m = 1, \ldots, M$, with the same factorization as (8), that are coupled by sharing one factor,

$$\mathcal{X}^{(m)} = \sum_{r=1}^{R} \mathbf{a}_r^{(m)} \circ \mathbf{b}_r^{(m)} \circ \mathbf{c}_r \tag{13}$$

where the factor matrices of the mth tensor are $\mathbf{A}^{(m)} = [\mathbf{a}_1^{(m)}, \ldots, \mathbf{a}_R^{(m)}] \in \mathbb{C}^{I_m \times R}$, $\mathbf{B}^{(m)} = [\mathbf{b}_1^{(m)}, \ldots, \mathbf{b}_R^{(m)}] \in \mathbb{C}^{J_m \times R}$, $\mathbf{C} = [\mathbf{c}, \ldots, \mathbf{c}_R] \in \mathbb{C}^{K \times R}$. The *coupled rank* of the set $\{\mathcal{X}^{(m)}\}$ is defined as the minimal number of rank-1 terms $\mathbf{a}_r^{(m)} \circ \mathbf{b}_r^{(m)} \circ \mathbf{c}_r$ that yield $\{\mathcal{X}^{(m)}\}$ in a linear combination [90]. If the coupled rank of $\{\mathcal{X}^{(m)}\}$ is R, then (13) is called the *coupled CPD* of $\{\mathcal{X}^{(m)}\}$. It has recently been shown that the coupled CPD may be unique even if conditions (12) are violated such that none of the individual CPDs in (13) is unique [90].

This fundamental result extends to more elaborate scenarios. Uniqueness can be further improved if the order of (at least one of) the involved tensors increases [90]. This is analogous to the previously-mentioned result (Section III-C2) for a single tensor, that increasing its order N relaxes the bound on R [57], [91]. Adding assumptions such as individual uniqueness of one of the involved CPDs, full column rank of the shared factor $\mathbf{C}$, or a specific structure such as a Vandermonde matrix, also reinforces the uniqueness of the whole decomposition [90], [124]. Finally, all these results can be extended to more elaborate tensor decompositions that are not limited to rank-1 terms [90] .

Another benefit from coupling is that it helps relax the permutation ambiguity. Coupled tensor decompositions have a unique arbitrary permutation matrix in a manner that extends single-tensor results (9) [125, Section III.A] [90]. Consequently, the low-rank terms that are shared by all the coupled tensors automatically have the same ordering at the output of the algorithm.

Linked-mode PARAFAC in which two or more third-order tensors share a mode has first been suggested in [126, p. 281]. The idea was extended to the case of arrays of different orders (one of them must be three-way or higher) in [69, Section 5.1.1]. Coupled tensor decompositions have already proven useful in telecommunications [125], multidimensional harmonic retrieval [124], chemometrics and psychometrics [8], [99], and more. See Figure 2a for an example in metabolomics.

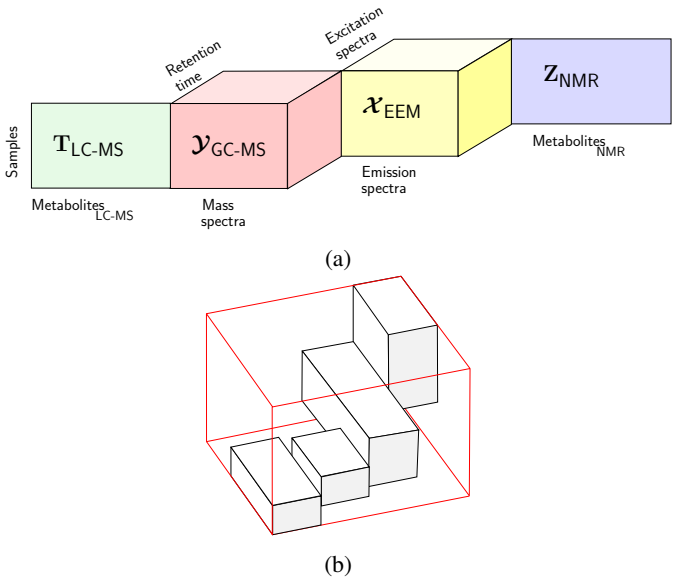

Fig. 2: Illustration of different types of coupling between matrices and third-order tensors. (a) Linked-mode matrices and tensors in metabolomics. Datasets represent four different acquisition methods. All datasets share the same "samples" mode. Figure reproduced from [127]. (b) Arrays (in this case, third-order tensors) may be coupled in different modes and also via only part of a mode. In addition, linked arrays may be regarded as elements in a larger volume (the red cube), in which certain data points are missing. Figure reproduced from [69, Figure 3].

Linked-mode analysis has also been proposed as a means to represent *missing values*: each tensor is a dataset that by itself is complete, but as a whole, each dataset has only partial information w.r.t. a larger array in which all these datasets are enclosed. This idea is accompanied by a more flexible coupling design where more than one mode may be shared between two tensors and the coupling may even involve only parts of modes, i.e., shared (sub) factors [69, Section 5.1.2]. Figure 2b illustrates this idea. Missing values are further discussed in Challenge IV-B.4.

We now summarize Examples III-D.1 and III-D.2. In Section III-C, we have shown that both ICA and PARAFAC can provide sufficient diversity to overcome the indeterminacy problem inherent to FA. We then extended our discussion to jointly analysing M such problems: $M\times$ICA $\xrightarrow{\text{joint pdf}}$ IVA (Example III-D.1) and $M\times$PARAFAC $\xrightarrow{\text{shared factor}}$ coupled CPD (Example III-D.2). We have shown that by properly defining a link between datasets, we can extend and reinforce uniqueness and identifiability beyond those obtained by individual analysis, up to the point of establishing uniqueness of otherwise non-unique scenarios. In PARAFAC, mixtures share certain factors, whereas in IVA, each mixture has its own individual parameters and the link is via statistical dependence between certain variables. Next, we have shown that all these models are flexible in the sense that they can easily be fine-tuned and modified in multiple ways, in order to better fit various real-life data. More specifically, they readily admit various types of diversity. A first generalization of these basic models is by relaxing the assumptions within each decomposition: allowing statistical dependence between latent sources of the *same* mixture in ICA (resp. IVA) leads to independent subspace analysis (ISA) [128]–[132] (resp. joint independent subspace analysis (JISA) [119]–[121]) as well as other BSS models [133], [134]. Relaxing the sum-of-rank-1-terms constraint in PARAFAC leads to more flexible tensor decompositions such as Tucker [6], [7], block term decomposition (BTD) [135], three-way decomposition into directional components (DEDICOM) [136], and others [137]. A second generalization is by combining several types of constraints and assumptions: for example, PARAFAC may be combined with statistical independence [104], [138], non-negativity, sparsity, as well as structure of the latent factors: Vandermonde [72]–[75], Toeplitz [76], among others [16], [65], [106], [139]. A third generalization is increasing the number of types of observational diversity by increasing the tensor order [57], [91]. A fourth is by linking datasets, leading to various coupled models, as explained in this section. When all these types of generalizations are taken into account, one obtains very general data fusion frameworks such as structured data fusion (SDF) [16], coupled matrix and tensor factorization (CMTF) [99], linked multiway component analysis (LMWCA) [65], and others [140]–[142]. These generalizations, and many more, are further discussed in Section V. In all cases, the link between underlying factors at different modalities helps not only to enhance uniqueness but also to enable the same ordering for all decompositions, thus further enhancing performance, identifiability, and interpretability.

E. Conclusion: A Link Between Datasets is Indeed a New Form of Diversity

The strength of IVA and coupled CPD over a set of unlinked factorizations lies in their ability to exploit commonalities among datasets. In IVA, it is the statistical dependence of sources across mixtures; in coupled CPD, it is the shared factors. In both scenarios, the links themselves are new types of information: the fact that datasets are linked, that elements in different datasets are related (or not), and the nature of these interactions, bring new types of constraints into the system that allow to reduce the number of degrees of freedom and thus enhance uniqueness, performance, interpretability, and robustness, among others. On top of that, the links among the datasets allow desired properties within one dataset to propagate to the ensemble and enhance the properties of the whole decomposition [16]. This is a concrete mathematical manifestation of the *raison d'être* of data fusion that we have mentioned in Section II, implying that [11, Section 9] [82], [143]

> An ensemble of datasets is "more than the sum of its parts" in the sense that it contains precious information that is *lost if these relations are ignored*.

The models that we have just presented allow multiple datasets to inform each other and interact, as formulated in Defini-

tion I.2 and further elaborated in Section V-A. Therefore, in the same vein of the preceding discussion and Definition I.1, we conclude that [16], [82], [143]

> Properly linking datasets can be regarded as introducing a new form of diversity, and this diversity is the basis and driving force of data fusion.

IV. Challenges at the Data Level

Thanks to recent advances, the *availability* of multimodal data is now a fact of life. The *acquisition* of multimodal data, however, is only a first step. In this section, and Section V that follows, we discuss some of the issues that should be addressed in the actual processing of multimodal data. In this section, we focus on challenges *imposed* by the data. These can be partitioned into challenges at the *acquisition and observation level* and challenges due to various types of *uncertainty in the data*. A number of approaches, to both types of challenges, are briefly mentioned in this section. Section V complements this section with a more comprehensive discussion of *how* to approach, in practice, some of these challenges, from a model design perspective.

A. Challenges at the Acquisition and Observation Level

Challenge IV-A.1: Non-commensurability. As explained in Sections I–II, a key motivation for multimodality is that different instruments are sensitive to different physical phenomena, and consequently, report on different aspects of the underlying processes. A natural outcome is that the raw measurements may be represented by different types of physical units that do not commute. This situation is known as *non-commensurability*. Numerous examples of non-commensurable data fusion scenarios were given in Section II. Allowing non-commensurable datasets to inform each other and interact is probably the first and foremost task that one encounters in a large number of multimodal data fusion scenarios [8].

Challenge IV-A.2: Different resolutions. It is most natural that different types of acquisition methods and observation setups provide data at different sampling points, and often at very disparate resolutions. The specific type of challenge that is associated with this property varies according to the task. Consequently, solutions are diverse. In some cases, different resolutions may be associated with various types of uncertainty, as explained in Section IV-B. Below, we list some scenarios in which challenges related to different resolutions occur. Data with different resolutions is a prevalent challenge in multimodal image fusion [13], as well as many other imaging techniques. For example, EEG has an excellent temporal but low spatial resolution, whereas fMRI has a fine spatial resolution but a very large integration time (Example II-B.1). In remote sensing (Example II-C.1), a common task is "pan-sharpening" [13, Chapter 9] [40]: merging a high-spatial, low-spectral (single band) resolution panchromatic image with a lower-spatial, higher-spectral (several bands) resolution multispectral image, in order to generate a new synthetic image that has both the higher spectral and spatial resolution of the two. In audio-visual applications, the temporal resolution of the signals differs by orders of magnitude. An audio signal is usually sampled at several kHz whereas the video signal is typically sampled at 15–60 Hz [144] (Example II-A.1). In meteorological monitoring (Example II-C.2), each modality has very distinct spatial and temporal resolutions; this is probably the reason why solutions based on data integration [47] (see Section V-A) are preferred. Different sampling schemes in coupled matrix and tensor decompositions are discussed, e.g., in [145], [146].

Challenge IV-A.3: Incompatible size. In practical situations, it is quite rare that different datasets contain exactly the same number of data samples. As explained in Section IV-B, this incompatibility may be associated with various types of uncertainty. Data size incompatibility may be due to a different number of samples at each observational mode, as explained in Challenge IV-A.2. Among possible causes are different acquisition techniques and experimental setups. The difference in size becomes even more acute if datasets are arrays of different orders [8], [147], as is often encountered in chemometrics, metabolomics (e.g., Figure 2a), and psychometrics, among others.

Challenge IV-A.4: Alignment and registration. Registration is the task of aligning several datasets, images, on the same coordinate system. Registration is particularly challenging when 3D biomedical imaging techniques are involved (Example II-B.1). In a *first* scenario, images of the same subject are taken at different times using the same imaging technique. The difficulty arises from the fact that each image has some bias and spatial distortion w.r.t. the others since the patient is never precisely in the same position. In this case, image registration usually relies on the basic assumption that image intensities are linearly correlated [148]. This assumption, however, is much less likely in the *second* scenario, for multimodal images. Consider, for example, registration of modalities that convey anatomical information with others that report on functional and metabolic activity. Naturally, the information conveyed by each modality is inherently of different physical nature. Other complicating factors include different types of noise, spatial distortions, varying contrasts, and different positions of the imaging instruments. One approach uses information theory and maximizes mutual information [148], [149]. In remote sensing (Example II-C.1), images of the same area are taken by different types of instruments, e.g., airborne SAR and satellite-borne LiDAR, and possibly at different times and conditions, e.g., before and after landscape-changing events such as natural disasters. In principle, one can use global positioning system (GPS) for aligning the images. However, even the GPS signal has a finite spatial precision. In biomedical imaging, the BOLD signal, to which fMRI is sensitive, has a large integration time and thus a delay w.r.t. EEG. This leads to non-instantaneous coupling, even if the measurements themselves are perfectly synchronized.

Calibration can be interpreted as a special case of alignment and registration using two sets of measurements, and thus it can be considered as a form of data fusion. Calibration is

a major task in chemometrics, where it is often achieved via regression methods. Frequently-used models such as multiway partial least squares (PLS) [150] and PARAFAC [62] are less adequate when the underlying profiles change shape from sample to sample. A regression method that can accommodate such variability in multiway arrays is proposed in [151], and a multimodal audio-visual calibration technique is given in [25]. The advantage of the proposed solution is that it is based on direction of arrival estimation, an easier task than microphone-based time-difference of arrival estimation, which requires strict synchronization between microphones. The challenge of automatic calibration of audio-visual sensors (and others) in the context of HMI (Example II-A.2) is discussed in [10, Section V.C].

B. Challenges due to Various Types of Uncertainty

We now turn to discussing uncertainty in the data. Any real-world set of observations is prone to various types of uncertainty. The presence of heterogeneous multiple datasets creates *new types of uncertainty* that may also be heterogeneous. We argue that in such cases, it is the complementarity and diversity (Definition I.1) of the datasets that should be exploited to *resolve* these challenges [9].

Challenge IV-B.1: Noise. Thermal noise, calibration errors, finite precision, quantization or any other quality degradation in the measurements is unavoidable. For simplicity, we denote all these unavoidable phenomena as "noise". Naturally, each acquisition method produces not only heterogeneous types of desired data, but also heterogeneous types of errors [8]. The question of how to *jointly* weigh or balance *different* sources of error is brought up in a number of data fusion scenarios, although most data fusion work currently ignores noise. Naturally, in the presence of noise, an appropriate model yields a more precise inference.

Several authors [152]–[154] use an additive noise model with a distribution whose parameters may vary within and across datasets. A Bayesian or maximum likelihood (ML) framework is then applied to estimate the noise parameters. In some cases, the noise estimates are interpreted as weights that balance the contribution of each element [153] (note the link with Challenge IV-B.2). Beal et al. [24] propose a graphical model for audio-visual object tracking in which they attribute different parameters to audio and video noise, and estimate both in a Bayesian inference framework. All these methods assume independence among sources of noise across modalities. However, ignoring possible links (correlations) between noise across datasets may lead to bias [9].

Challenge IV-B.2: Balancing information from different origins. In practice, for various reasons, not all observations or data entries have the same level of confidence, reliability or information quality [11], [21], [155]. Below, we list scenarios in which this occurs, as well as some approaches to resolve these problems.

In real-life scenarios, certain sensors may be provide information that has more value than others, or certain measurements might be taken at better-controlled scenarios. For example, in a medical questionnaire about patients $\times$ symptoms, certain symptoms may be more obvious and some harder to define [155]. In the same vein, heterogeneity of acquisition methods implies heterogeneity in their level of importance or usefulness. For example, we may use questionnaires filled by specialists (experts) and others by patients (non-experts). Alternatively, if we consider two medical questionnaires, patients $\times$ symptoms and patients $\times$ diagnosis, the first one may be more reliable since symptoms are observed directly whereas diagnosis relies on interpretation [155]. In order to address this issue, Wilderjans et al. [153], [155] propose to associate the level of reliability with noise, and use appropriate weights, obtained via an ML variant of simultaneous component analysis (SCA). Şimsekli et al. [156] propose individual weights for datasets with different divergence measures based on their relative 'importance'. Similar to [153], these weights are interpreted as noise variances. Finally, in [47], Liberman et al. process several meteorological monitoring modalities separately and then make a soft decision, using an optimal weighted average based on location, number of links, rainfall intensity and other parameters.

Another source of potential unbalance is datasets of different size (recall Challenge IV-A.3). In the absence of additional assumptions, a simulation study favours equal weight to each data entry regardless of its dataset of origin, over the alternative approach of weighting datasets by the number of their respective entries [147] . This approach is generalized to the case of missing values, where the weights should be proportional to the number of non-missing entries in each dataset [16].

Challenge IV-B.3: Conflicting, contradicting or inconsistent data. Whenever more than one origin of information is available, be it a single sensor or an ensemble of observations, conflicts, contradictions, and inconsistencies may occur. If data is fused at the decision level, then a decision or voting [8] rule may be applied, for example, in the fusion of different classification maps in remote sensing (see Example II-C.1). Other approaches, related to multisensor data fusion, are discussed in [9], [157]. When only two datasets are confronted, more elaborate solutions may be required. An obvious challenge is to devise a suitable compromise. A more fundamental challenge, however, is *identifying* these inconsistencies.

In [158], Tmazirte et al. consider the problem of detecting faults in multimodal sensors in a distributed data fusion framework, and dynamically reconfigure the system using information theoretical concepts. Their approach is based on detecting inconsistency in the mutual information contribution of each sensor w.r.t. its history. Kumar et al. [159] deal with the problem of multimodal sensors that occasionally produce spurious data, possibly due to sensor failure or environmental issues, and thus may bias estimation. The challenge arises from the fact that spurious events are difficult to predict and to model. Kumar et al. [159] propose a Bayesian approach that can identify and eliminate spurious data from a sensor. The procedure attributes less weight to the measurement from a suspected sensor when fused with measurements from other sensors. In the inference of cosmological parameters (Example II-C.3), detecting and explaining (in)consistencies of observations from different experiments is of utmost im-

portance. A methodology for validation is comparing various error measures on several types of analytical products (Example V-A.1): cosmological parameters, CMB power spectra, and full sky maps, with and without the inclusion of datasets from both space-borne satellite missions and Earth-bound telescopes [50]. These experiments vary in the spectral bands at which they observe the sky, angular (spatial) resolution, sensitivity to different types of polarization, sky coverage, sky-scanning strategies [50, Section 4.1], types of noise, and other parameters. Therefore, they carry complementary information.

Challenge IV-B.4: Missing values. The challenge of missing values is not new and not unique to data fusion. The problem of matrix and tensor completion is long-standing in linear and multilinear algebra. However, its prevalence in data fusion draws special attention to it. In Section III, we have seen that low-rank tensors decompositions provide redundancy that results in strong uniqueness that is further improved in the presence of coupling or additional constraints. It turns out that the same applies also in the case of missing values, see [16], [99], [160]–[162] and references therein. Approaches to missing values that are motivated by various aspects of data fusion can be found, e.g., in [16], [160], [161].

Missing values may occur in various scenarios. While the first case that we mention below is not unique to data fusion, the remaining ones are. More specifically, the first case deals with samples that are locally missing within an individual dataset, whereas the other cases arise due to interaction among datasets. *First*, certain data entries may be unreliable, discarded, or unavailable due to faulty detectors, occlusions, partial coverage, or any other unavoidable effects. *Second*, sometimes a modality can report only on part of the system w.r.t. the other modalities, as is the case with EEG vs. MEG [12], nuclear magnetic resonance (NMR) vs. liquid chromatography—mass spectrometry (LC-MS) [99], occlusions or partial spatial coverage in remote sensing (Example II-C.1), audio-video (Example II-A.1), meteorological monitoring (Example II-C.2), and HMI (Example II-A.2). A *third* scenario is illustrated in Figure 2b. In this case, there exist several datasets, depicted as complete third-order tensors. However, when linked together, they can be regarded as elements in a larger third-order tensor in which they are all contained, and whose volume is only partially filled. *Fourth*, data may be regarded as structurally missing if samples at different modalities are not taken at comparable sampling points [8] (recall Challenge IV-A.2), and we would like to construct a more complete picture from the entire sample set. In this case, each modality is properly sampled on its own, but there exist points on the common sampling grid that do not contain data from all modalities; these points can be regarded as missing values. A *fifth* scenario is link prediction. This is a common issue in recommender systems and social network analysis. In social network analysis, the challenge is predicting social links or activities based on an existing database of connections or activities, where only a few entries are known. As an example for a recommender system we mention the "Netflix Prize", where the challenge is defined as improving the accuracy of predictions about how much a person is going to enjoy a future movie based on past preferences. The data can be regarded as an incomplete user $\times$ movie matrix, whose entries are user ratings in an ordinal 1–5 scale, and the challenge is to fill in the missing entries (initially set to zero). Among the many and diverse methods that have been proposed we mention that some are based on (coupled) matrix or tensor factorizations, possibly by augmenting these data with further types of diversity; see, e.g., [16], [161], [163] and references therein.

V. Challenges at the Model Design Level

In this section, we confront the unavoidable "how" question, presenting some *guidelines* that might be helpful in the actual design of data fusion *solutions*, from a model design perspective. This question has already been raised by numerous authors, e.g., by [8]–[14], [16], [152], [163], [164], among others, and the following discussion builds upon these foundations. In a sense, this section concludes our paper. It complements Section III by proposing theoretical model design principles that allow diversity to manifest itself. It complements Section II and Section IV by presenting model design principles that can accommodate the practical data-level challenges presented in Section IV and the numerous tasks given in Section II. It provides examples of approaches that allow datasets to interact and inform each other, in the sense of Definition I.2. As in previous sections, due to the vastness of the field, the discussion in this section is far from being exhaustive: we only touch at certain topics, and leave others, such as computation, algorithms and fusion of large-scale data, outside the scope of this overview. The rest of this section is organised as follows. In Section V-A, we discuss different strategies to data fusion, and address, in particular, at which level of abstraction, reduction and simplification the data should be fused. Section V-B discusses mathematical models for links between datasets that maximally exploit diversity, enhance interpretability and performance. Section V-C discusses some theoretical approaches to the analysis of the ensemble of linked datasets. Section V-D brings together the numerous model design steps and considerations in a unified framework of "structured data fusion". We conclude our discussion with validation issues in Section V-E.

A. Level of Data Fusion

At first thought, it may seem that fusing multiple datasets at the raw-data level should always yield the best inference, since there would be no loss of information. In practice, however, due to the complex and largely unknown nature of the underlying phenomena (Section II), various complicating factors (Section IV), as well as the specific research question (Sections I–II), it may turn out to be more useful to fuse the datasets at a higher level of abstraction [9], and after certain simplification and reduction steps. The procedures listed below precede the actual fusion of the data. Therefore, they are related to the preprocessing stage. Naturally, the choice of analytical model is influenced by decisions taken at this point.

The first strategy that we mention is *data integration*. It implies parallel processing pipelines for each modality,

followed by a decision-making step. Integration is a common approach to deal with heterogeneous data. When modalities are completely non-commensurable (Challenge IV-A.1), as with remote sensing techniques that report on material content vs. others that report on three-dimensional structures (Example II-C.1), integration becomes a natural choice, and is often related to classification tasks. Integration can be done via soft decision, using optimal weights, as in the fusion of data from wireless microwave sensor networks and radar for rainfall measurement and mapping [47] (Example II-C.2). Bullmore and Sporns [165] study brain networks by first constructing separate models of structural and functional networks based on several brain imaging modalities and fuse them using a graph-theoretical framework. Data integration may be preferred when modality-specific information carries more weight compared with the shared information, as argued for the joint analysis of EEG–fMRI in [32] (Example II-B.1). A framework to choose between alternative soft decision strategies in the presence of multiple sensor outputs, given various assumptions on uncertainty or partial knowledge, confidence levels, reliability, and conflicts, in a data fusion context, is given in [157]. Due to its simplicity, and relative stability since it allows to rely on well-established methods from single-modal data analysis, a large number of existing data fusion approaches are still based on decision-level fusion. Pros and cons to data integration are further discussed in [21].

A second type of data fusion strategy is *processing modalities sequentially*, where one (or more) modality(ies) is used to constrain another. Mathematically, this amounts to using one modality to restrict the number of degrees of freedom, and thus the set of possible solutions, in another. A sequential approach makes sense when one modality has better quality in terms of the information that it conveys than the others in a certain respect, as in certain audio-visual scenarios [10], [23] (Example II-A.1), as well as in the fMRI-constrained solution for the otherwise-underdetermined, ill-posed EEG inverse problem [12], [26] (Example II-B.1).

In this paper, we focus on a third strategy, *true fusion*, that lets modalities fully interact and inform each other as claimed in Section I. True fusion is also characterized by assigning a symmetric role to all modalities, i.e., not sequential. The data fusion models mentioned in Section III fall into this category, as well as most of the models that we mention in the rest of this section. Within "true fusion", there are varying degrees:

True fusion using high-level features. In this approach, the dimensionality is significantly reduced by associating each modality with a small number of variables. High-level features are often univariate. Examples include standard variation, skewness, ratio of active voxels, other variables which concisely summarize statistics, or geometrical and other properties. In this case, inference is typically of classification type. Examples include multi-sensor [9], HMI [10] and remote sensing [42] applications.

True fusion using multivariate features. Unlike high-level features, this approach leaves the data sufficiently multivariate within each modality (which now is in feature form) such that data in each modality can fully interact [21], [58]. In neuroimaging, common features are task-related spatial maps from fMRI, gray matter images from sMRI, and event-related potential (ERP) from EEG, extracted for each subject [21], [58], [60]. In audio-visual applications, features often correspond to speech spectral coefficients and visual cues such as lip contours or speaker's presence in the scene [23].

True fusion using the data as is, or with minimal reduction. In fact, working with features implies a two-step approach: in the first step, features are computed using a certain criterion; in the second step, features are fused using a different, second criterion. An approach that merges the two, and thus expected to better exploit the whole raw data, is proposed in [166] for the fusion of fMRI and EEG. A remote sensing application in which it is natural to work with raw data is pan-sharpening (explained in Challenge. IV-A.2). Here, acquisition conditions are favourable since the two sensors (multispectral and pan) acquire data over the same area, with same angle of view and simultaneously, and the modalities are commensurable.

Features, at different levels, may accommodate heterogeneities across modalities, such as different types of uncertainty and non-commensurability (Section IV). Features may significantly reduce the number of samples involved, i.e., allow *compression*. Example V-A.1 illustrates this point. It also serves as a conclusion to the discussion on the strategy for data fusion by showing how different levels of features can be used for varying data fusion purposes. For further discussion on features and choosing the right level of data fusion, see, e.g., [21], [27], [58] (biomedical imaging) and [10], [11] (HMI).

Example V-A.1: Use of Features in Cosmological Inference from CMB Observations. In the inference of cosmological parameters from CMB observations (Example II-C.3), the raw data consist of detector readouts as well as other auxiliary information that amounts to several Tera-bytes, or $O(10^{12})$, of observations [167]. The scientific products are usually provided in several levels of multivariate "features", as follows: (i) full-sky maps, of CMB and non-CMB emissions, amounting to roughly $O(10^8)$ pixels, (ii) CMB power spectrum computed from the CMB spatial map, at $O(10^3)$ spectral multipoles, and (iii) six cosmological parameters that represent the best-fit of the CMB power spectrum to the ΛCDM model. It is clear that each level represents a strong compression of the data w.r.t. the preceding one. Each level of features is useful for a different type of inference. High-resolution component maps are the first useful outcome from the component separation procedure [59]. Apart from providing valuable information about the sky, they are useful for instance for consistency checks between instruments, experiments and methods [50], [59]. Power spectra are useful to compare outcomes of different experiments that measure the CMB, e.g., Planck and BICEP2/Keck [168], whereas cosmological parameters form the link, via the ΛCDM model, with datasets that do not involve astrophysical observations, e.g., high-energy physics at CERN [56].

Order selection and dimension reduction. Related to the open issue of choosing the most appropriate strategy of data fusion is *order selection*. As in non-multimodal analysis, a dimension reduction step may be required in order to avoid

over-fitting the data, as well as a form of compression [9]. In a data fusion framework, this step must take into consideration the possibly different representations of the latent variables across datasets. As an example, a solution that maximally retains the joint information while also ensuring that the decomposed sources are independent from each other, in the context of a "joint ICA"-based approach, is proposed in [117]. Dimension reduction may be performed locally, at each sensor or modality, or at a central processing unit [9].

B. Link Between Datasets

Data fusion is all about enabling modalities to fully interact and inform each other. Hence, a key point is choosing an analytical model that faithfully represents the relationship between modalities and yields a meaningful combination thereof, without imposing phantom connections or suppressing existing ones. The underlying idea of data fusion is that an ensemble of datasets is "more than the sum of its parts" in the sense that it contains precious information that is *lost if these relations are ignored*. The purpose of properly-defined links is to support this goal, as motivated by the discussion in Section III. In order to maximize diversity, we would like links to be able to exploit the heterogeneity among datasets. Properly-defined links provide a clear picture of the underlying structure of the ensemble of the related datasets [147]. Consider, for example, two datasets, patients $\times$ symptoms and patients $\times$ diagnosis; we would like the data fusion model to allow us to uncover the medical conditions that underlie both symptoms and diagnoses [155]. Properly-defined links explain similarities and differences among datasets and allow better interpretability. As explained in Section III, one of the first motivations for linking datasets in joint matrix decomposition scenarios is resolving the arbitrary ordering of the latent components in each individual dataset. It is interesting to note that all types of links eventually alleviate this problem since they provide a single frame of reference.

Since data fusion generally deals with heterogeneous datasets, we would like links to be flexible enough to allow each dataset to remain in its most explanatory form, as further discussed in Section V-B1. In various scenarios, certain elements may be present only in a specific dataset whereas others are shared by two or more. We would like the model not only to properly express these elaborate interactions but also to have the capacity to *inform* us about (non-) existence of links when this information is not available in advance, a topic further elaborated in Section V-B2.

As stated in Section II, the *raison d'être* of multimodal data fusion is the paradigm that certain natural processes and phenomena express themselves under completely different physical guises. Due to the often complex nature of the driving phenomena, it is likely that datasets will be related via more than one type of diversity; e.g., time, space, and frequency. Therefore, links should be designed such that they support a relationship via several types of diversity simultaneously, whenever applicable. Models based on multilinear relationships, as well as those that admit multiple types of links simultaneously, seem to better support this aim.

1) "Soft" and "Hard" Links Between Datasets: One type of decision that has to be made is whether each dataset will have its own set of individual parameters, disjoint of the others, or not. In the first case, *none* of the parameters that define each dataset's model are shared by any other dataset. As a result, additional information is required to define the link. In such cases, the link is often defined as some correspondence between datasets that can be interpreted as similarity, smoothness or continuity [169]. Therefore, we call such links "soft". In the second case, datasets explicitly share certain factor matrices or latent variables. For the sake of our discussion, we call such links "hard" [145].

"Hard" links between datasets. We have already seen shared factor matrices in numerous examples in Section III. Naturally, data fusion methods that are based on stacking data in a single tensor fall within this category. Such are PARAFAC (Section III-C2), generalized singular value decomposition (GSVD) [170] and its higher-order generalization [171], the higher-order SVD (HOSVD) [172], and more. In joint ICA [173] and group ICA [174], [175], several ICA problems share a mixing matrix or source subspace, respectively, by concatenating the observation matrices in rows or columns. Simultaneous factor analysis (FA) [152] and simultaneous component analysis (SCA)-based methods [98], [176]–[179] deal with multiway data that have at least one shared mode, but do not stack it in a single tensor (Section III-C2) due to various complicating factors, such as those mentioned in Section IV. Linked tensor ICA [154] has one factor matrix shared by all decompositions. In Bayesian group FA [180] and its tensor generalization [142], [181], as well as in collective matrix factorization (CMF) [164], several matrices or tensors share all but one factor matrix. In fusion of hyperspectral and multispectral images (Example II-C.1), the joint factor is a matrix that reflects the (desired, unknown) high-resolution image before spatial and spectral degradation [182]–[185]. In group non-negative matrix factorization (NMF), shared columns of the feature matrix reflect task-related variations [186]. The generalized linked-mode framework for multiway data [8] allows flexible links across datasets by shared (sub-) factors, as do other flexible tensor-based data fusion models such as coupled matrix and tensor factorization (CMTF) [99] and its probabilistic extension generalized coupled tensor factorization (GCTF) [161], linked multiway component analysis (LMWCA) [65] and structured data fusion (SDF) [16]. In the fusion of astrophysical observations of the CMB from different experiments (Example II-C.3), the link may be established by a joint distribution of the ensemble of samples from all datasets. In this case, the fusion is based on the assumption that the random processes from which all samples are generated are controlled by the same underlying cosmological parameters [50]. A shared random variable is used also in [187] to extract a common source of variability from measurements in multiple sensors using diffusion operators [187].

"Soft" links between datasets. Prevalent types of "soft" links are *statistical dependence*, as in IVA (Example III-D.1); *covariations*, as in CCA [1] and its extension to more than two matrices, multiset CCA (MCCA) [3], [58], [110], [118], and parallel ICA [188]; and "*similarity*", in the sense of

minimizing some distance measure between corresponding elements, as in soft non-negative matrix co-factorization [189] and joint matrix and tensor decompositions with flexible coupling [145]. For audio-visual data fusion, a dictionary learning model where each atom consists of an audio and video component has been proposed in [144]. A graphical model in which audio and video shifts are linearly related in far-field conditions is proposed in [24]. Generalized linked-mode for multiway data [8] and LMWCA [65] mention explicitly that they can be defined both with "soft" or "hard" links.

Although the partition into "soft" and "hard" links is conceptually appealing and simplifies our presentation, the following reservation is in order. In practice, when it comes to writing the optimization problem, models with "soft" links are often reformulated using shared variables. In the models that we have just mentioned, shared variables are, for example, cross-correlation or cross-cumulants when statistical (in)dependence and co-variation are concerned, or a shared latent variable in regression models. The reformulated models often take the form of (approximate) (coupled) matrix or tensor factorizations. We mention [111], [190], [191] as just a few examples; further discussion is beyond the scope of this paper. The bottom line is that the distinction between "soft" and "hard" links is often immaterial. The implication is that models with "soft" links can sometimes neatly fit within optimization frameworks that assume shared variables, for example SDF [16].

2) Shared vs. Unshared Elements: The idea that datasets have both shared (common) and unshared (individual, modality-specific) elements w.r.t. the others can be found in numerous models. It can be formulated mathematically by defining certain columns of a factor matrix or sub-elements of a latent variable as shared while others are unshared. Models that admit this formulation include incomplete mode PARAFAC [69, Section 5.1.2] (Figure 2b), group NMF [186], LMWCA [192] and SDF [16]. In the extraction of a common source of variability from heterogeneous sensors [187], it is hidden random variables that are either shared or unshared. Another example is Bayesian group FA [180] and its tensor extensions [142], [181], where a dedicated factor matrix determines which of the factors in a common pool are active within each dataset.

The more fundamental challenge, however, is to *identify* the shared and unshared elements from the data itself, without a priori assignment of individual and shared variables. Bayesian linked tensor ICA [154] holds a modality-specific factor matrix of optimally-determined weights that can eliminate a source from some modalities while keeping it in others. In [170], Alter et al. propose GSVD to infer, from two genome-scale expression datasets, shared and individual processes. Ponnapalli et al. [171] extend this GSVD-based approach to more than two datasets. Shared and unshared processes in genomic and metabolomic data may also be revealed by a proper rotation of the components resulting from SCA. The proposed approach, called distinctive and common components with simultaneous-component analysis (DISCO-SCA) [177], [193], [194], may outperform GSVD in certain scenarios and can be straightforwardly generalized to more than two datasets. A comparative study of GSVD, DISCO-SCA and other methods that can identify shared and unshared processes underlying multiset data can be found in [179]. HOSVD [195] can differentiate between shared and unshared phenomena in DNA analysis from multiple experiments [172]. As a last example, in CMTF [196], model constraints may be defined in the form of sparse weights such that unshared components have norms equal or close to zero in one of the datasets.

C. Analytical Framework

Certain data fusion approaches rely on existing theoretical analytical frameworks that have originally been devised for non-fusion applications, at least not explicitly. Such are ICA and algebraic-based methods such as PARAFAC, generalized eigenvalue decomposition (GEVD), GSVD and HOSVD, as will be elaborated below. These methods have been around for a while and there is a large body of works that has been dedicated to their computation. Data fusion approaches that rely on these well-established, widely-known methods are often more easily accepted and integrated within the research communities. However, these approaches may not be able to exploit the full range of diversity in the data, and thus, more advanced data fusion methods may be preferred. Below, we briefly review some of the analytical approaches that have been proposed for data fusion.

Well-known matrix and tensor factorizations can be used for data fusion. In [170], Alter et al. use GSVD for the comparison of genetic data from two different organisms. In [172], HOSVD is proposed for the analysis of data from different studies. In the presence of two datasets, or in a noise-free scenario, many matrix- and tensor-based methods can be reformulated as GEVD [197]. This holds for various BSS closed-form solutions [198], CCA [199, Chapter 12] and its multi-set extension [3], [110], joint BSS [111], and coupled tensor decompositions [200]. As explained in Section V-B1, algebraic (possibly approximate) solutions to models with "soft" links often exist.

Certain data fusion methods concatenate or re-organize data such that it can be analysed by a classical ICA algorithm. Such is the case in joint ICA [173] and group ICA [174], [175]. These models can thus be solved using any existing ICA approach [63].

Guo et al. [201] propose a tensor extension to group ICA [174], [175] and to tensor ICA [104] that can accommodate different group structures. Parallel ICA [188] and IVA [108]–[111] (Example III-D.1) jointly solve several separate ICA problems by exploiting co-variations or statistical dependence, respectively. CCA [1], its extension to multiple datasets [3], as well as one of the approaches to LMWCA [65], search for maximal correlation, or other second-order-based relationships, between variables.

Certain methods minimize the Euclidean distance (Frobenius norm) between model and data. In the presence of additive white Gaussian noise, this amounts to maximum likelihood (ML). Further considerations associated with this choice of norm are given in [16, Section II]. This type of optimization is used in group NMF [186], coupled NMF [182], certain

SCA-based methods [98], [176]–[178], and numerous coupled tensor decompositions, see e.g., [16], [99], [124] and references therein. In some cases, it may be better to tailor loss functions individually to each dataset, and use norms other than Frobenius [8], [153]. Such is the case in the GCTF framework [156], [161], for example. An ML framework can accommodate datasets with different noise patterns. Such are flexible simultaneous FA [152] and ML-based SCA [153], [155]. ML underlies certain noiseless stochastic models, e.g., SOS-IVA [202] and JISA [119]–[121].

Regression provides another solution to data fusion. Regression searches for latent factors that best explain the covariance between two sets of observations. We mention PLS [62], [203] and its multilinear extensions N-way PLS [62], [150] and higher-order PLS (HOPLS) [141].

Bayesian group FA [180], its tensor extension [142], [181], coupled matrix and tensor decompositions with flexible coupling [145], and certain methods for the fusion of hyperspectral and multispectral images [183], [184], rely on a Bayesian framework for the decomposition. Certain tensor extensions of ICA rely on a probabilistic Bayesian framework [154]. In [185], fusion of hyperspectral and multispectral images is achieved via dictionary learning and sparse coding. This is also the underlying technique of [144] for learning bimodal structure in audio-visual data. Beal et al. [24], [204] use probabilistic generative models, also termed graphical models, in order to fuse audio and video models into a single probabilistic graphical model. Lederman and Talmon [187] use an alternating-diffusion method for manifold learning that extracts a common source of variability from measurements in multiple sensors, where all sensors observe the same physical phenomenon but have different sensor-specific effects. Combining labelled and unlabelled data via co-training is described in [205]. A survey of techniques for multi-view machine learning can be found in [206]. A multimodal deep-learning method for information retrieval from bi-modal data consisting of images and text is described in [207].

D. Structured Data Fusion: A General Mathematical Framework

In the preceding sections, we mentioned a large number of data fusion models. However, it is clear that no list of existing solutions, comprehensive as it might be, can cover the practically endless number of current, future and potential datasets, problems and tasks. Indeed, the purpose of this paper is not in promoting specific models or methods. Instead, and building upon [8]–[14], [16], [152], [163], [164] and others, we wish to provide a deeper and broader understanding of the concepts and ideas that underlie data fusion. As such, in the model design front, our goal is providing guidelines and insights that may apply also to datasets, problems and tasks that do not necessarily conform to any of the specific examples, solutions, and mathematical frameworks that we mention. The concept of diversity, presented in Section III, is one such example. In the same vein, we now present a general mathematical framework that will allow us to give a more concrete meaning to some of the model design concepts that have been discussed. Although this formulation is given in terms of matrices and higher-order arrays, also known as tensors, the underlying ideas behind "structured data fusion" [16], such as flexibility and modularity, are not limited to these. The mathematical formulation that we use is only a concretization of a more general idea, applied to datasets that admit certain types of decompositions.

We now present a formulation proposed by Sorber et al. [16], followed by a few examples for motivation and clarification.

Model of an individual dataset: Consider an ensemble of M datasets, collected in M arrays (tensors) $\mathcal{T}^{(m)} \in \mathbb{C}^{I_1 \times \cdots \times I_{N_m}}$, $m = 1, \ldots, M$, where $N_m = 1, 2, 3, \ldots$ implies a vector, a matrix or a higher-order tensor, respectively. In order to allow maximal flexibility in the model associated with each of these datasets, [16] define several layers of underlying structures. A *first* layer is an ordered set of V variables $\mathbf{z} = \{z_1, \ldots, z_V\}$, where each variable may be anything from a scalar to a higher-order tensor, real or complex (recall (1)). A *second* layer is an ordered set of F factors $\mathcal{X}(\mathbf{z}) = \{x_1(z_{i_1}), \ldots, x_F(z_{i_F})\}$ that are driven by the V variables $\mathbf{z}$. Each factor $x_f(z_{i_f})$ is a mapping of the i_fth variable to a tensor. In a *third* layer, each dataset $\mathcal{T}^{(m)}$ is associated with a decomposition model $\mathcal{M}^{(m)}$ that approximates it. Function $\mathcal{M}^{(m)}(\mathcal{X}(\mathbf{z}))$ maps a subgroup of the factors $\mathcal{X}(\mathbf{z})$ to a tensor. Figure 3 illustrates these layers. For simplicity, each dataset $\mathcal{T}^{(m)}$ is associated with a tensor model $\mathcal{M}^{(m)}$ of the same order and size. This is not evident: the order N_m and dimensions $I_1 \times \cdots \times I_{N_m}$ of the model tensor, as used in the analysis, may *differ* from those that most naturally represent the acquired data, as well as from the natural way to visualize the samples. As a first example, raw EEG data is a time series in several electrodes, i.e., electrode $\times$ time. However, it has been proposed to augment this data using a third type of diversity, electrode $\times$ time $\times$ frequency [101]–[103]. In this case, the EEG data will be stacked in a third-order tensor. On the other hand, data that is naturally visualized in 2D or 3D arrays such as images does not necessarily admit any useful (multi-) linear relationships among its pixels or voxels. Hence, in the analysis, image data are often vectorized into 1D arrays. Further discussion of how to choose the right array structure for data analysis can be found, e.g., in [8], [64], [65], [106] and references therein.

Link between datasets: In the structured data fusion (SDF) formulation, a link between datasets can be established if their models share at least one factor or variable. This corresponds to the "hard" links, mentioned in Section V-B1. However, this does not exclude other types of interaction between datasets: "soft" links may be established by reformulating "soft" links using shared parameters, as explained in Section V-B1, and possibly via regularization terms.

The following examples provide a more concrete meaning to the mathematical formulation that we have just laid out. Consider two matrix datasets, patients $\times$ diagnosis and patients $\times$ symptoms. A latent variable may be the syndrome that underlies both diagnoses and symptoms factors. The link is established via the shared "patients" mode [155]. As a second example, certain properties of a communication system in a coupled CPD framework may be expressed using a factor with a Vandermonde structure [124]. This can be implemented

in SDF with z_f a vector of p scalars and $x_f(z_f)$ a $p \times q$ Vandermonde matrix constructed thereof. Further examples for factor structures that can be reformulated in SDF include orthogonality, Toeplitz structure, non-negativity, as well as fixed and known entries. Constraints such as sparsity and smoothness within factors may be implemented using regularization terms. Tensor decompositions that can be reformulated as $\mathcal{M}^{(m)}(\mathcal{X}(\mathbf{z}))$ include rank-1 decompositions (CPD, PARAFAC), decompositions into rank$\geq$1 terms, Tucker, BTD, PARAFAC2 [70], and many others. These options, as well as other alternatives for latent variables, factors and models for each dataset, are mentioned in Section III. For further explanations about implementation, see [16], [208].

Loss/objective function, regularization and penalty terms: The next step is to fit the model to the data. Depending on the analytical framework that we choose (Section V-C), each dataset or the whole ensemble is attributed with a loss/objective function $D^{(m)}(\cdot,\cdot)$, between observed and modelled data [153], [156], [161]. An individual loss/objective function allows flexibility both in the analytical framework applied to each dataset, and in the individual types of uncertainty (Challenge IV-B.1). The loss/objective function may be complemented by various penalty or regularization terms, in order to impose constraints that are not expressed by the other optimization functionals. Regularization terms may impose certain types of sparsity, non-negativity [196], similarity [145], [189], or coherence [209], to name a few.

Missing values: In order not to take account of unknown data entries in the optimization procedure, these values are masked. This is done via an entry-wise (Hadamard) product (denoted as $\circledast$) of the data tensor $\mathcal{T}^{(m)}$ with a binary tensor $\mathcal{B}^{(m)}$ of the same size; see e.g., [16], [160] and references therein. Missing values are discussed in Challenge IV-B.4.

The whole optimization problem: Given these elements, SDF may be written as the optimization problem

$$\min_{\mathbf{z}} \quad \sum_{m=1}^{M} \frac{\omega_m}{2} D^{(m)}\left(\mathcal{M}^{(m)}(\mathcal{X}(\mathbf{z})), \mathcal{T}^{(m)}\right)_{\mathcal{B}^{(m)}} + \text{regularization terms}, \quad (14)$$

where $D^{(m)}(\cdot,\cdot)_{\mathcal{B}^{(m)}}$ implies $D^{(m)}(\mathcal{B}^{(m)}\circledast\cdot, \mathcal{B}^{(m)}\circledast\cdot)$. Scalars ω_m denote weights, reflecting the relative importance of the loss/objective functions in the ensemble. Scenarios in which weights are useful are discussed in Section IV-B. The optimization problem (14) implements the overall analytical framework associated with the model. Eq. (14) is a slight generalization of the original SDF formulation [16, Eq. (1)], in which the loss/objective function is a weighted Frobenius norm, $D^{(m)}(\mathcal{M}^{(m)}, \mathcal{T}^{(m)})_{\mathcal{B}^{(m)}} = \|\mathcal{B}^{(m)}\circledast(\mathcal{M}^{(m)} - \mathcal{T}^{(m)})\|_F^2$. Numerical and computational advantages associated with the Frobenius norm in the context of SDF are discussed in [16, Section II]. An illustration of SDF is given in Figure 3.

As noted by [16], a large number of the existing data fusion models can be reformulated in terms of SDF, or some variation thereof. However, an even more interesting insight is that each step in the design of (14) is independent of the others: to a large extent, the choice of constraints, assumptions, types of links, loss/objective functions, and other parameters, can be done disjointly. In other words, SDF is a *modular approach* to data fusion. These insights have led to the key observation that in fact, *a large number of existing data fusion models can be regarded as composed of a rather small number of building blocks* [16]. In other words, if all admissible combinations are considered, then the number of potential analytical data fusion models is significantly larger than what is currently available in the literature [8].

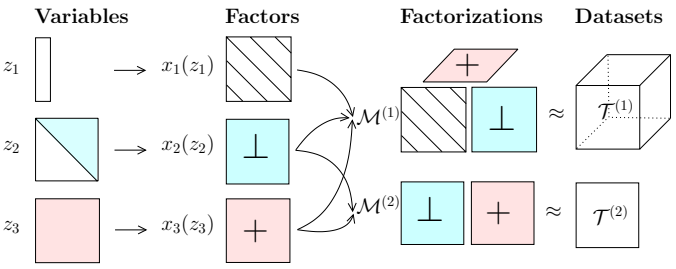

Fig. 3: Schematic illustration of structured data fusion. For example, vector z_1, upper triangular matrix z_2, and full matrix z_3 are transformed, using mappings x_1, x_2 and x_3, into a Toeplitz, orthogonal, and nonnegative matrix, respectively. The resulting factors are then used to jointly factorize two coupled datasets. Figure and caption reproduced from [16, Figure 1].

The modular perspective on data fusion offers several benefits. First, a major challenge in data fusion is its augmented complexity due to the increased number of degrees of freedom. The modular approach to data fusion *answers this challenge* by reformulating the problem in a small set of disjoint simpler components that can be separately analysed, optimized and coded. Second, the modular approach, in which the problem is factorized into smaller stand-alone elements, allows a broader view that makes it is easier to come up with new combinations of the basic building blocks, thus leading to new mathematical models, algorithms and concepts [8], [16], [163], [164]. Third, modularity of the formulation makes it easier to adapt it to computational challenges such as large-scale data [16], [65], [162], [210], [211]. Fourth, in Section I–II, we have emphasized the importance of exploratory research in data fusion. The modularity of the design is particularly helpful in that, making it straightforward to come up with new exploratory variations, to test and compare alternatives with minimal effort [16]. Modularity allows to easily diagnose which elements in the model are particularly useful, need to be modified, replaced or fine-tuned, without having to undo the whole derivation, coding or analysis. The latter also facilitates the *validation* stage, see Section V-E for further discussion.

E. Validation

Despite accumulating empirical evidence of the benefits of data fusion, there is still very little theoretical validation and quantitative measure of its gain [11], [99]. Choosing an appropriate model is a widely open question, and approximate and highly simplified models are often preferred. Therefore, a validation step is indispensable. The following points are of particular interest: (i) Lower bounds on the best achievable error: how far are we from the best possible result (for a given

dataset, task, goal, and model)? (ii) Theoretical results on the reliability and practical usefulness of the method: can we prove that the model is identifiable? Is the solution unique? Is the output physically meaningful? Are the results sufficiently interpretable? IVA (Example III-D.1) and coupled tensor decompositions (Example III-D.2) are two of the models for which there now exists a comprehensive theoretical analysis that answers this type of questions.

Although these questions are not specific to multimodal data fusion, they take special interpretation in the presence of multiple datasets. Some of the new questions that arise are, for instance: (i) What is the mathematical formulation of "success", "optimality" and "error", when heterogeneous modalities and types of uncertainty are involved? What is the most appropriate target function and criterion of success? (ii) How to evaluate performance of exploratory tasks? (iii) How to design a figure of merit that can inform us how to exploit the advantages of each modality without suffering from its drawbacks w.r.t. the other modalities? (iv) How to identify and process information that is shared by several modalities, and how to identify and exploit modality-specific information? (v) How to compare alternative design choices such as of level of data fusion, order selection, and analytical model within and across modalities? As an example, theoretical figures of merit such as the Cramér-Rao lower bound may help answer some of these questions. However, calculating theoretical error bounds for all possible alternatives (especially in view of the modular approach of Section V-D) is a prohibitive task, both due to the very large number of options, and also since many models are not mathematically tractable. Rescue may come from the computational front. As an example, Tensorlab [208], a MATLAB toolbox that follows the modular principles of SDF [16] (Section V-D), enables the user to switch between the numerous combinations arising from multiple choices in the model design. As such, it allows the user to rapidly iterate towards a plausible solution for the problem at hand. Therefore, Tensorlab [208] (or any other computational tool following the modularity principles) may serve as a verification and validation tool, at least in the preliminary stages of the design.

A class on their own are questions regarding the choice of modalities and the added value from using multiple modalities in general. (i) Should all available modalities be used, and/or given equal importance? (ii) How much (information, diversity, redundancy) does each modality bring in to the total equations? How to quantify this "extra contribution"? Some of these questions (and examples of possible answers) have been brought up within the challenges in Section IV; others are related to the design of a data fusion model (Section V). Information theory seems like a natural framework to evaluate the contribution of various types of diversity, as discussed, e.g., in [12]. Uniqueness analysis of (coupled) tensor decompositions, as well as other forms of error analysis, such as those mentioned in Section III, quantify the added value of diversity in terms of the admissible number of uniquely identifiable components or factors. Attention should be paid, for example, when modalities are too close to each other: in this case, they may not really convey new information; in addition, they may be exposed to similar noise, and thus bias results [9]. Due to the heterogeneous characteristics of the data, and particularly in exploratory tasks, the interpretability of the output should be given special care. Questions related to the representation of the output of multimodal data analysis are discussed, e.g., in [11, Section 8].

VI. Conclusion

We enter an era where the abundance of diverse sources of information makes it practically impossible to ignore the presence of multiple datasets that are possibly related. It is very likely that an ensemble of related datasets is "more than the sum of its parts", in the sense that it contains precious information that is lost if these relations are ignored. The information of interest that is hidden in these datasets is usually not easily accessible, however. We argue that the road to this added value must go through first understanding and identifying the particularities of multimodal and multiset data, as opposed to other types of aggregated datasets. At the same time, the joint analysis of multiple datasets "stands on the shoulders of" single-set analysis. Hence, the development of methods and techniques for single-set analysis is a cornerstone for advanced data fusion. In this paper, we have shown that methods that properly account for the links among datasets indeed have the potential to achieve gains and benefits that go far beyond those possible when each dataset is processed individually. As argued in this paper, the potential impact of these gains is high, and spans the whole spectrum from solving theoretical problems that cannot be solved in single-set scenarios, to opening up new opportunities in numerous medical, environmental, psychological, social and technological domains, among others. By adopting a data-driven approach, we have shown that the encountered challenges are ubiquitous, whence the incentive that both challenges and solutions be discussed at a level that brings together all involved communities.

Acknowledgment

The authors would like to thank Lieven De Lathauwer, Jean-François Cardoso, Jocelyn Chanussot, Mauro Dalla Mura, Noam David, Inbar Fijalkow, Hagit Messer, Gadi Miller, and Sabine Van Huffel, whose expertise, insightful remarks and feedback have greatly helped extend the scope of this paper; and the anonymous reviewers, for their careful reading and valuable remarks.

References

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, Dec. 1936.
[2] R. B. Cattell, ""Parallel proportional profiles" and other principles for determining the choice of factors by rotation," *Psychometrika*, vol. 9, no. 4, pp. 267–283, Dec. 1944.
[3] J. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
[4] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sep. 1970.
[5] R. A. Harshman, "Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, Dec. 1970.

[6] L. R. Tucker, *Contributions to Mathematical Psychology*. New York: Holt, Rinehardt & Winston, 1964, ch. The extension of factor analysis to three-dimensional matrices, pp. 109–127.

[7] ——, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep. 1966.

[8] I. Van Mechelen and A. K. Smilde, "A generic linked-mode decomposition model for data fusion," *Chemom. Intell. Lab. Syst.*, vol. 104, no. 1, pp. 83–94, Nov. 2010.

[9] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, Jan. 2013.

[10] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct. 2010.

[11] M. Turk, "Multimodal interaction: A review," *Pattern Recognition Letters*, vol. 36, pp. 189–195, Jan. 2014.

[12] F. Bießmann, S. Plis, F. C. Meinecke, T. Eichele, and K. Müller, "Analysis of multimodal neuroimaging data," *IEEE Rev. Biomed. Eng.*, vol. 4, pp. 26–58, 2011.

[13] T. Stathaki, *Image fusion: algorithms and applications*. Elsevier, 2008.

[14] H. B. Mitchell, *Data fusion: concepts and ideas*, 2nd ed. Springer, 2012.

[15] T. Adalı, Z. J. Wang, V. D. Calhoun, T. Eichele, M. J. McKeown, and D. Van de Ville, Eds., *Special Section on Multimodal Biomedical Imaging: Algorithms and Applications*, vol. 15, no. 5. IEEE Trans. Multimedia, Aug. 2013.

[16] L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 586–600, Jun. 2015.

[17] A. R. McIntosh, F. L. Bookstein, J. V. Haxby, and C. L. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *NeuroImage*, vol. 3, no. 3, pp. 143–157, Jun. 1996.

[18] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976.

[19] D. McLaughlin, "An integrated approach to hydrologic data assimilation: interpolation, smoothing, and filtering," *Adv. Water Resour.*, vol. 25, no. 8–12, pp. 1275–1286, Aug.-Dec. 2002.

[20] H. Boström, S. F. Andler, M. Brohede, R. Johansson, A. Karlsson, J. van Laere, L. Niklasson, M. Nilsson *et al.*, "On the definition of information fusion as a field of research," University of Skövde, School of Humanities and Informatics, Tech. Rep. HS- IKI -TR-07-006, 2007.

[21] V. D. Calhoun and T. Adalı, "Feature-based fusion of medical imaging data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 711–720, Sep. 2009.

[22] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography – theory, instrumentation, and applications to noninvasive studies of the working human brain," *Rev. Mod. Phys.*, vol. 65, pp. 413–497, Apr. 1993.

[23] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.

[24] M. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 828–836, Jul. 2003.

[25] A. Plinge and G. A. Fink, "Geometry calibration of distributed microphone arrays exploiting audio-visual correspondences," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 116–120.

[26] P. L. Nunez and R. B. Silberstein, "On the relationship of synaptic activity to macroscopic measurements: Does co-registration of EEG with fMRI make sense?" *Brain Topography*, vol. 13, no. 2, pp. 79–96, Dec. 2000.

[27] X. Lei, P. A. Valdes-Sosa, and D. Yao, "EEG/fMRI fusion based on independent component analysis: Integration of data-driven and model-driven methods," *J. Integr. Neurosci.*, vol. 11, no. 03, pp. 313–337, 2012, pMID: 22985350.

[28] B. Horwitz and D. Poeppel, "How can EEG/MEG and fMRI/PET data be combined?" *Human brain mapping*, vol. 17, no. 1, pp. 1–3, 2002.

[29] S. Makeig, T.-P. Jung, and T. J. Sejnowski, *Exploratory Analysis and Data Modeling in Functional Neuroimaging*. MIT Press, 2003, ch. Having your voxels and timing them too?, p. 195.

[30] E. Martínez-Montes, P. A. Valdés-Sosa, F. Miwakeichi, R. I. Goldman, and M. S. Cohen, "Concurrent EEG/fMRI analysis by multiway partial least squares," *NeuroImage*, vol. 22, no. 3, pp. 1023–1034, Jul. 2004.

[31] J. Sui, T. Adalı, Y.-O. Li, H. Yang, and V. D. Calhoun, "A review of multivariate methods in brain imaging data fusion," in *SPIE Medical Imaging*, R. C. Molthen and J. B. Weaver, Eds., vol. 7626. International Society for Optics and Photonics, 2010, pp. 76 260D–1–76 260D–11.

[32] M. De Vos, R. Zink, B. Hunyadi, B. Mijovic, S. Van Huffel, and S. Debener, "The quest for single trial correlations in multimodal EEG–fMRI data," in *Proc. EMBC'13*, Osaka, Japan, Jul. 2013, pp. 6027–6030.

[33] J. C. Mosher, P. S. Lewis, and R. M. Leahy, "Multiple dipole modeling and localization from spatio-temporal MEG data," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 6, pp. 541–557, Jun. 1992.

[34] H. Becker, L. Albera, P. Comon, R. Gribonval, F. Wendling, and I. Merlet, "A performance study of various brain source imaging approaches," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 5910–5914.

[35] C. M. A. Hoeks, J. O. Barentsz, T. Hambrock, D. Yakar, D. M. Somford, S. W. T. P. J. Heijmink, T. W. J. Scheenen, P. C. Vos *et al.*, "Prostate cancer: Multiparametric MR imaging for detection, localization, and staging," *Radiology*, vol. 261, no. 1, pp. 46–66, Oct. 2011.

[36] A. R. Croitor-Sava, M. C. Martinez-Bisbal, T. Laudadio, J. Piquer, B. Celda, A. Heerschap, D. M. Sima, and S. Van Huffel, "Fusing in vivo and ex vivo NMR sources of information for brain tumor classification," *Meas. Sci. Technol.*, vol. 22, no. 11, p. 114012, 2011.

[37] M. Garibaldi and V. Zarzoso, "Exploiting intracardiac and surface recording modalities for atrial signal extraction in atrial fibrillation," in *Proc. EMBC'13*, Osaka, Japan, Jul. 2013, pp. 6015–6018.

[38] A. Van de Vel, K. Cuppens, B. Bonroy, M. Milosevic, K. Jansen, S. Van Huffel, B. Vanrumste, L. Lagae *et al.*, "Non-EEG seizure-detection systems and potential SUDEP prevention: State of the art," *Seizure-Eur. J. Epilep.*, vol. 22, no. Issue 5, pp. 345–355, Jun. 2013.

[39] N. Yokoya, T. Yairi, and A. Iwasaki, "Hyperspectral, multispectral, and panchromatic data fusion based on coupled non-negative matrix factorization," in *Proc. WHISPERS*, Lisbon, Portugal, Jun. 2011.

[40] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[41] C. Berger, M. Voltersen, R. Eckardt, J. Eberle, T. Heyer, N. Salepci, S. Hese, C. Schmullius *et al.*, "Multi-modal and multi-temporal data fusion: Outcome of the 2012 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 6, no. 3, pp. 1324–1340, Jun. 2013.

[42] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens *et al.*, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[43] R. B. Smith, "Introduction to hyperspectral imaging," TNTmips®, Mar. 2013. [Online]. Available: http://www.microimages.com/documentation/Tutorials/hyprspec.pdf

[44] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proc. IEEE*, 2015, submitted.

[45] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel *et al.*, "Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009–2010 data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 5, no. 1, pp. 331–342, Feb. 2012.

[46] H. Messer, A. Zinevich, and P. Alpert, "Environmental monitoring by wireless communication networks," *Science*, vol. 312, no. 5774, p. 713, May 2006.

[47] Y. Liberman, R. Samuels, P. Alpert, and H. Messer, "New algorithm for integration between wireless microwave sensor network and radar for improved rainfall measurement and mapping," *Atmos. Meas. Tech.*, vol. 7, pp. 3549–3563, 2014.

[48] H. Seyyedi, "Comparing satellite derived rainfall with ground based radar for North-western Europe," Master's thesis, ITC, Enschede, The Netherlands, Jan. 2010.

[49] N. David, P. Alpert, and H. Messer, "Technical note: Novel method for water vapour monitoring using wireless communication networks measurements," *Atmos. Chem. Phys.*, vol. 9, no. 7, pp. 2413–2418, 2009.

[50] Planck Collaboration, "Planck 2013 results. I. Overview of products and scientific results," *A&A*, vol. 571, no. A1, pp. 1–48, Nov. 2014.

[51] G. Hinshaw, D. Larson, E. Komatsu, D. N. Spergel, C. L. Bennett, J. Dunkley, M. R. Nolta, M. Halpern *et al.*, "Nine-year Wilkinson microwave anisotropy probe (WMAP) observations: Cosmological parameter results," *ApJS*, vol. 208, no. 2, pp. 19 (1–25), 2013.

[52] Planck Collaboration, "Planck 2013 results. XVI. Cosmological parameters," *A&A*, vol. 571, no. A16, pp. 1–66, Nov. 2014.

[53] A. A. Penzias and R. W. Wilson, "A measurement of excess antenna temperature at 4080 Mc/s," *ApJ*, vol. 142, pp. 419–421, Jul. 1965.

[54] M. Betoule, R. Kessler, J. Guy, J. Mosher, D. Hardin, R. Biswas, P. Astier, P. El-Hage *et al.*, "Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples," *A&A*, vol. 568, no. A22, Aug. 2014.

[55] N. Palanque-Delabrouille, C. Yèche, J. Lesgourgues, G. Rossi, A. Borde, M. Viel, E. Aubourg, D. Kirkby *et al.*, "Constraint on neutrino masses from SDSS-III/BOSS Lyα forest and other cosmological probes," *JCAP*, vol. 2015, no. 02, p. 045, Feb. 2015.

[56] M. Krawczyk, D. Sokołowska, P. Swaczyna, and B. Świeżewska, "Constraining inert dark matter by $R_{\gamma\gamma}$ and WMAP data," *J. High Energy Phys.*, vol. 55, no. 9, Sep. 2013.

[57] N. D. Sidiropoulos and R. Bro, "On communication diversity for blind identifiability and the uniqueness of low-rank decomposition of N-way arrays," in *Proc. ICASSP*, vol. 5, Istanbul, Turkey, Jun. 2000, pp. 2449–2452.

[58] N. M. Correa, T. Adalı, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *IEEE Signal Process. Mag.*, vol. 27, no. 4, pp. 39–50, Jul. 2010.

[59] Planck Collaboration, "Planck 2013 results. XII. Diffuse component separation," *A&A*, vol. 571, p. A12, Nov. 2014.

[60] J. Sui, T. Adalı, Q. Yu, J. Chen, and V. Calhoun, "A review of multivariate methods for multimodal fusion of brain imaging data," *J. Neurosci. Methods*, vol. 204, no. 1, pp. 68–81, 2012.

[61] N. D. Sidiropoulos, G. B. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 810–823, Mar 2000.

[62] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons, Aug. 2004.

[63] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Academic Press, Feb. 2010.

[64] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIREV*, vol. 51, no. 3, pp. 455–500, Sep. 2009.

[65] A. Cichocki, D. Mandic, A. H. Phan, C. Caiafa, G. Zhou, Q. Zhao, and L. De Lathauwer, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 145–163, Mar. 2015.

[66] J. B. Kruskal, *Multiway data analysis*. Amsterdam, The Netherlands: Elsevier, 1989, ch. Rank, decomposition, and uniqueness for 3-way and N-way arrays, pp. 7–18.

[67] ——, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra and its Applications*, vol. 18, no. 2, pp. 95–138, 1977.

[68] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[69] R. A. Harshman and M. E. Lundy, "PARAFAC: Parallel factor analysis," *Comput. Statist. Data Anal.*, vol. 18, no. 1, pp. 39–72, Aug. 1994.

[70] R. A. Harshman, "PARAFAC2: Mathematical and technical notes," *UCLA Working Papers in Phonetics*, vol. 22, pp. 30–47, 1972, university Microfilms, Ann Arbor, No. 10,085.

[71] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[72] X. Liu and N. D. Sidiropoulos, "Cramér-Rao lower bounds for low-rank decomposition of multidimensional arrays," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 2074–2086, Sep. 2001.

[73] N. D. Sidiropoulos, "Generalizing Carathéodory's uniqueness of harmonic parameterization to N dimensions," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1687–1690, May 2001.

[74] N. D. Sidiropoulos and X. Liu, "Identifiability results for blind beamforming in incoherent multipath with small delay spread," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 228–236, Jan. 2001.

[75] M. Sørensen and L. De Lathauwer, "Blind signal separation via tensor decomposition with Vandermonde factor: Canonical polyadic decomposition," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5507–5519, Nov. 2013.

[76] L. De Lathauwer and A. de Baynast, "Blind deconvolution of DS-CDMA signals by means of decomposition in rank-$(1, L, L)$ terms," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1562–1571, Apr. 2008.

[77] J. Paisley and L. Carin, "Nonparametric factor analysis with Beta process priors," in *Proc. ICML*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 777–784.

[78] P. Rai and H. Daumé III, "The infinite hierarchical factor regression model," in *Proc. NIPS*, Dec. 2008, pp. 1321–1328.

[79] D. Knowles and Z. Ghahramani, "Nonparametric Bayesian sparse factor models with application to gene expression modeling," *Ann. Appl. Stat.*, vol. 5, no. 2B, pp. 1534–1552, Jun. 2011.

[80] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[81] J.-F. Cardoso, "The three easy routes to independent component analysis; contrasts and geometry," in *Proc. ICA*, San Diego, CA, USA, Dec. 2001, pp. 1–6.

[82] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Process. Mag.*, pp. 18–33, May 2014.

[83] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Radar and Signal Process., IEE Proceedings F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[84] L. De Lathauwer, "Signal processing based on multilinear algebra," Ph.D. dissertation, KU Leuven, Leuven, Belgium, Sep. 1997.

[85] A. L. F. de Almeida, X. Luciani, A. Stegeman, and P. Comon, "CONFAC decomposition approach to blind identification of underdetermined mixtures based on generating function derivatives," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5698–5713, Nov 2012.

[86] E. Moreau and T. Adalı, *Blind Identification and Separation of Complex-Valued Signals*. Hoboken, NJ USA: John Wiley & Sons, Inc., 2013.

[87] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *J. Math. Phys.*, vol. , 6, pp. 164–189, 1927.

[88] R. A. Harshman, "Determination and proof of minimum uniqueness conditions for PARAFAC1," *UCLA Working Papers in Phonetics*, vol. 22, pp. 111–117, 1972, university Microfilms, Ann Arbor, No. 10,085.

[89] J. B. Kruskal, "More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling," *Psychometrika*, vol. 41, no. 3, pp. 281–293, Sep. 1976.

[90] M. Sørensen and L. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms—part I: Uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 2, pp. 496–522, May 2015.

[91] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemometrics*, vol. 14, no. 3, pp. 229–239, May–Jun. 2000.

[92] A. Stegeman and N. D. Sidiropoulos, "On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition," *Linear Algebra and its Applications*, vol. 420, no. 2, pp. 540–552, 2007.

[93] I. Domanov and L. De Lathauwer, "On the uniqueness of the canonical polyadic decomposition of third-order tensors—part I: Basic results and uniqueness of one factor matrix," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 855–875, 2013.

[94] ——, "On the uniqueness of the canonical polyadic decomposition of third-order tensors—part II: Uniqueness of the overall decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 876–903, 2013.

[95] ——, "Generic uniqueness conditions for the canonical polyadic decomposition and INDSCAL," *arXiv:1405.6238 [math.AG]*, 2014.

[96] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven, "An algorithm for generic and low-rank specific identifiability of complex tensors," *SIAM J. Matrix Anal. Appl.*, vol. 35, no. 4, pp. 1265–1287, 2014.

[97] I. Domanov and L. De Lathauwer, "Canonical polyadic decomposition of third-order tensors: relaxed uniqueness conditions and algebraic algorithm," ESAT-STADIUS, KU Leuven, Leuven, Belgium, Tech. Rep. 14-152, 2015.

[98] K. De Roover, E. Ceulemans, M. E. Timmerman, J. B. Nezlek, and P. Onghena, "Modeling differences in the dimensionality of multiblock data by means of clusterwise simultaneous component analysis," *Psychometrika*, vol. 78, no. 4, pp. 648–668, Oct. 2013.

[99] E. Acar, M. A. Rasmussen, F. Savorani, T. Næs, and R. Bro, "Understanding data fusion within the framework of coupled matrix and tensor factorizations," *Chemom. Intell. Lab. Syst.*, vol. 129, pp. 53–63, 2013.

[100] S. Miron, M. Dossot, C. Carteret, S. Margueron, and D. Brie, "Joint processing of the parallel and crossed polarized Raman spectra and uniqueness in blind nonnegative source separation," *Chemom. Intell. Lab. Syst.*, vol. 105, no. 1, pp. 7–18, Jan. 2011.

[101] E. Acar, C. Aykut-Bingol, H. Bingol, R. Bro, and B. Yener, "Multiway analysis of epilepsy tensors," *Bioinformatics*, vol. 23, no. 13, pp. i10–i18, 2007.

[102] M. De Vos, L. De Lathauwer, B. Vanrumste, S. Van Huffel, and W. Van Paesschen, "Canonical decomposition of ictal scalp EEG and accurate source localisation: Principles and simulation study," *Computational Intelligence and Neuroscience*, vol. 2007, pp. 1–10, 2007.

[103] F. Miwakeichi, E. Martínez-Montes, P. A. Valdés-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, "Decomposing EEG data into space–time–frequency components using parallel factor analysis," *NeuroImage*, vol. 22, no. 3, pp. 1035–1045, Jul. 2004.

[104] C. F. Beckmann and S. M. Smith, "Tensorial extensions of independent component analysis for multisubject fMRI analysis," *Neuroimage*, vol. 25, no. 1, pp. 294–311, Mar. 2005.

[105] J. Levin, "Simultaneous factor analysis of several Gramian matrices," *Psychometrika*, vol. 31, no. 3, pp. 413–419, 1966.

[106] P. Comon, "Tensors : A brief introduction," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 44–53, May 2014.

[107] J. J. Lacoume and P. Ruiz, "Sources indentification: a solution based on the cumulants," in *4th Annual ASSP Workshop on Spectrum Estimation and Modeling*, Minneapolis, MN, USA, Aug 1988, pp. 199–203.

[108] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: Definition and algorithms," in *Proc. ACSSC*, Pacific Grove, CA, Nov. 2006, pp. 1393–1396.

[109] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Independent Component Analysis and Blind Signal Separation*, ser. LNCS, vol. 3889. Springer Berlin Heidelberg, 2006, pp. 165–172.

[110] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3918–3929, Oct. 2009.

[111] X.-L. Li, T. Adalı, and M. Anderson, "Joint blind source separation by generalized joint diagonalization of cumulant matrices," *Signal Process.*, vol. 91, no. 10, pp. 2314–2322, Oct. 2011.

[112] J.-H. Lee, T.-W. Lee, F. A. Jolesz, and S.-S. Yoo, "Independent vector analysis (IVA): multivariate approach for fMRI group study." *Neuroimage*, vol. 40, no. 1, pp. 86–109, Mar. 2008.

[113] A. M. Michael, M. Anderson, R. L. Miller, T. Adalı, and V. D. Calhoun, "Preserving subject variability in group fMRI analysis: performance evaluation of GICA vs. IVA," *Frontiers in Systems Neuroscience*, vol. 8, no. 106, Jun. 2014.

[114] Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, "Data-driven fusion of EEG, functional and structural MRI: A comparison of two models," in *Proc. CISS*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[115] S. Ma, V. D. Calhoun, R. Phlypo, and T. Adalı, "Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis," *NeuroImage*, vol. 90, pp. 196–206, Apr. 2014.

[116] Y.-O. Li, T. Eichele, V. D. Calhoun, and T. Adalı, "Group study of simulated driving fMRI data by multiset canonical correlation analysis," *J. Sign. Process. Syst.*, vol. 68, no. 1, pp. 31–48, Jul. 2012.

[117] J. Sui, H. He, G. D. Pearlson, T. Adalı, K. A. Kiehl, Q. Yu, V. P. Clark, E. Castro *et al.*, "Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia," *NeuroImage*, vol. 66, pp. 119–132, Feb. 2013.

[118] A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.

[119] D. Lahat and C. Jutten, "Joint blind source separation of multidimensional components: Model and algorithm," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 1417–1421.

[120] R. F. Silva, S. Plis, T. Adalı, and V. D. Calhoun, "Multidataset independent subspace analysis extends independent vector analysis," in *Proc. ICIP*, Paris, France, Oct. 2014, pp. 2864–2868.

[121] D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," GIPSA-Lab, Grenoble, France, Technical report hal-01132297, Mar. 2015.

[122] J. Chatel-Goldman, M. Congedo, and R. Phlypo, "Joint BSS as a natural analysis framework for EEG-hyperscanning." in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 1212–1216.

[123] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, "Independent vector analysis: Identification conditions and performance bounds," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4399–4410, Sep. 2014.

[124] M. Sørensen and L. De Lathauwer, "Multidimensional harmonic retrieval via coupled canonical polyadic decomposition'," ESAT-SISTA, KU Leuven, Leuven, Belgium, Internal Report 13-240, 2013.

[125] ——, "Coupled tensor decompositions for applications in array signal processing," in *Proc. CAMSAP*. IEEE, 2013, pp. 228–231.

[126] R. A. Harshman and M. E. Lundy, *Research methods for multimode data analysis*. New York: Praeger, 1984, ch. Data preprocessing and the extended PARAFAC model, pp. 216–284.

[127] Joint data analysis for enhanced knowledge discovery in metabolomics. [Online]. Available: http://www.models.life.ku.dk/joda

[128] P. Comon, "Supervised classification, a probabilistic approach," in *Proc. ESANN*, Brussels, Belgium, Apr. 1995, pp. 111–128.

[129] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Fetal electrocardiogram extraction by source subspace separation," in *Proc. IEEE SP/ATHOS Workshop on HOS*, Girona, Spain, Jun. 1995, pp. 134–138.

[130] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP*, vol. 4, Seattle, WA, May 1998, pp. 1941–1944.

[131] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces," *Neural Comput.*, vol. 12, no. 7, pp. 1705–1720, Jul. 2000.

[132] D. Lahat, J.-F. Cardoso, and H. Messer, "Second-order multidimensional ICA: Performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4598–4610, Sep. 2012.

[133] M. Castella and P. Comon, "Blind separation of instantaneous mixtures of dependent sources," in *Independent Component Analysis and Signal Separation*, ser. LNCS, M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Eds., vol. 4666. Springer Berlin Heidelberg, 2007, pp. 9–16.

[134] A. Boudjellal, K. Abed-Meraim, A. Belouchrani, and P. Ravier, "Informed separation of dependent sources using joint matrix decomposition," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 1945–1949.

[135] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms. Part II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, 2008.

[136] R. A. Harshman, "Models for analysis of asymmetrical relationships among N objects or stimuli," in *Proc. First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology*, McMaster University, Hamilton, Ontario, Canada, Aug. 1978.

[137] G. Favier and A. L. F. de Almeida, "Overview of constrained PARAFAC models," *EURASIP JASP*, vol. 2014, no. 1, 2014.

[138] M. De Vos, D. Nion, S. Van Huffel, and L. De Lathauwer, "A combination of parallel factor and independent component analysis," *Signal Process.*, vol. 92, no. 12, pp. 2990–2999, 2012.

[139] T. Yokota, A. Cichocki, and Y. Yamashita, "Linked PARAFAC/CP tensor decomposition and its fast implementation for multi-block tensor analysis," in *Neural Information Processing*, ser. LNCS, T. Huang, Z. Zeng, C. Li, and C. Leung, Eds. Springer Berlin Heidelberg, 2012, vol. 7665, pp. 84–91, 19th International Conference, ICONIP 2012, Doha, Qatar, Nov. 12–15, 2012, Proceedings, Part III.

[140] A. H. Phan and A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional datasets," *Nonlinear Theory and Its Applications, IEICE*, vol. 1, no. 1, pp. 37–68, 2010.

[141] Q. Zhao, C. F. Caiafa, D. P. Mandic, Z. C. Chao, Y. Nagasaka, N. Fujii, L. Zhang, and A. Cichocki, "Higher order partial least squares (HOPLS): A generalized multilinear regression method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1660–1673, Jul. 2013.

[142] S. A. Khan, E. Leppäaho, and S. Kaski, "Multi-tensor factorization," *arXiv:1412.4679 [stat.ML]*, 2014.

[143] D. Lahat, T. Adalı, and C. Jutten, "Challenges in multimodal data fusion," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 101–105.

[144] G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audio–visual data," *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1898–1910, Dec 2009.

[145] R. Cabral Farias, J. E. Cohen, C. Jutten, and P. Comon, "Joint decompositions with flexible couplings," GIPSA-Lab, Grenoble, France, Tech. Rep. hal-01135920, Apr. 2015.

[146] A. Liutkus, U. Şimşekli, and T. Cemgil, "Extraction of temporal patterns in multi-rate and multi-modal datasets," in *Proc. LVA/ICA*, ser. LNCS. Liberec, Czech Republic: Springer-Verlag, Aug. 2015.

[147] T. Wilderjans, E. Ceulemans, and I. Van Mechelen, "Simultaneous analysis of coupled data blocks differing in size: A comparison of two weighting schemes," *Comput. Statist. Data Anal.*, vol. 53, no. 4, pp. 1086–1098, Feb. 2009.

[148] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[149] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, "Nonrigid image registration using conditional mutual information," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 19–29, Jan. 2010.

[150] R. Bro, "Multiway calibration. Multilinear PLS," *J. Chemometrics*, vol. 10, no. 1, pp. 47–61, Jan.–Feb. 1996.

[151] F. Marini and R. Bro, "SCREAM: A novel method for multi-way regression problems with shifts and shape changes in one mode," *Chemom. Intell. Lab. Syst.*, vol. 129, pp. 64–75, 2013, special Issue: Multiway and Multiset Methods.

[152] K. G. Jöreskog, "Simultaneous factor analysis in several populations," *Psychometrika*, vol. 36, no. 4, pp. 409–426, Dec. 1971.

[153] T. F. Wilderjans, E. Ceulemans, I. Van Mechelen, and R. A. van den Berg, "Simultaneous analysis of coupled data matrices subject to different amounts of noise," *Br. J. Math. Stat. Psychol.*, vol. 64, no. 2, pp. 277–290, May 2011.

[154] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, "Linked independent component analysis for multimodal data fusion," *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, Feb. 2011.

[155] T. F. Wilderjans, E. Ceulemans, and I. Van Mechelen, "The SIMCLAS model: Simultaneous analysis of coupled binary data matrices with noise heterogeneity between and within data blocks," *Psychometrika*, vol. 77, no. 4, pp. 724–740, Oct. 2012.

[156] U. Şimşekli, B. Ermiş, A. T. Cemgil, and E. Acar, "Optimal weight learning for coupled tensor factorization with mixed divergences," in *Proc. EUSIPCO*, Marrakech, Morocco, Sep. 2013.

[157] I. Bloch, "Information combination operators for data fusion: a comparative review with classification," *IEEE Trans. Syst., Man, Cybern. A*, vol. 26, no. 1, pp. 52–67, Jan. 1996.

[158] N. A. Tmazirte, M. E. El Najjar, C. Smaili, and D. Pomorski, "Dynamical reconfiguration strategy of a multi sensor data fusion algorithm based on information theory," in *IEEE Intelligent Vehicles Symposium (IV)*, Gold Coast, QLD, Australia, Jun. 2013, pp. 896–901.

[159] M. Kumar, D. P. Garg, and R. A. Zachery, "A method for judicious fusion of inconsistent multiple sensor data," *IEEE Sensors J.*, vol. 7, no. 5, pp. 723–733, May 2007.

[160] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemom. Intell. Lab. Syst.*, vol. 106, no. 1, pp. 41–56, Mar. 2011.

[161] B. Ermiş, E. Acar, and A. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 203–236, 2015.

[162] N. Vervliet, O. Debals, L. Sorber, and L. De Lathauwer, "Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in big data analysis," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 71–79, Sep. 2014.

[163] A. P. Singh and G. J. Gordon, "A unified view of matrix factorization models," in *Machine Learning and Knowledge Discovery in Databases*, ser. LNCS, W. Daelemans, B. Goethals, and K. Morik, Eds. Springer Berlin Heidelberg, 2008, vol. 5212, pp. 358–373, European Conference ECML PKDD, Antwerp, Belgium, September 15–19, 2008, Proceedings, Part II.

[164] ——, "Relational learning via collective matrix factorization," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Las Vegas, NV, USA: ACM, Aug. 2008, pp. 650–658.

[165] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, Mar. 2009.

[166] N. M. Correa, T. Eichele, T. Adalı, Y.-O. Li, and V. D. Calhoun, "Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," *NeuroImage*, vol. 50, no. 4, pp. 1438–1445, May 2010.

[167] R. Stompor, "Data analysis of massive data sets a Planck example," Presentation at LOFAR workshop, Meudon, France, Mar. 2006. [Online]. Available: www.lesia.obspm.fr/plasma/LOFAR2006/Stompor.pdf

[168] BICEP2/Keck and Planck Collaborations, "A joint analysis of BICEP2/Keck Array and Planck data," *Phys. Rev. Lett.*, vol. 114, no. 10, p. 101301, Mar. 2015.

[169] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[170] O. Alter, P. O. Brown, and D. Botstein, "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms," *PNAS*, vol. 100, no. 6, pp. 3351–3356, 2003.

[171] S. P. Ponnapalli, M. A. Saunders, C. F. Van Loan, and O. Alter, "A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms," *PLoS One*, vol. 6, no. 12, Dec. 2011, article e28072.

[172] L. Omberg, G. H. Golub, and O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies," *PNAS USA*, vol. 104, no. 47, pp. 18 371–18 376, Nov. 2007.

[173] V. D. Calhoun, T. Adalı, G. D. Pearlson, and K. A. Kiehl, "Neuronal chronometry of target detection: Fusion of hemodynamic and event-related potential data," *NeuroImage*, vol. 30, no. 2, pp. 544–553, Apr. 2006.

[174] V. D. Calhoun, T. Adalı, G. Pearlson, and J. Pekar, "Group ICA of functional MRI data: Separability, stationarity, and inference," in *Proc. ICA*, San Diego, CA, USA, Dec. 2001, pp. 155–160.

[175] V. D. Calhoun, T. Adalı, G. D. Pearlson, and J. J. Pekar, "A method for making group inferences from functional MRI data using independent component analysis," *Human Brain Mapping*, vol. 14, no. 3, pp. 140–151, Nov. 2001.

[176] H. A. L. Kiers and J. M. F. ten Berge, "Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure," *Br. J. Math. Stat. Psychol.*, vol. 47, no. 1, pp. 109–126, May 1994.

[177] K. Van Deun, I. Van Mechelen, L. Thorrez, M. Schouteden, B. De Moor, M. J. van der Werf, L. De Lathauwer, A. K. Smilde *et al.*, "DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes," *PloS one*, vol. 7, no. 5, p. e37840, May 2012.

[178] K. De Roover, M. E. Timmerman, I. Van Mechelen, and E. Ceulemans, "On the added value of multiset methods for three-way data analysis," *Chemom. Intell. Lab. Syst.*, vol. 129, pp. 98–107, Nov. 2013, multiway and Multiset Methods.

[179] K. Van Deun, A. K. Smilde, L. Thorrez, H. A. L. Kiers, and I. Van Mechelen, "Identifying common and distinctive processes underlying multiset data," *Chemom. Intell. Lab. Syst.*, vol. 129, no. 0, pp. 40–51, Nov. 2013.

[180] S. Virtanen, A. Klami, S. Khan, and S. Kaski, "Bayesian group factor analysis," in *Proc. AISTATS*, vol. 22:. La Palma, Canary Islands: JMLR, Apr. 2012, pp. 1269–1277.

[181] S. A. Khan and S. Kaski, "Bayesian multi-view tensor factorization," in *Machine Learning and Knowledge Discovery in Databases*, ser. LNCS, T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Eds. Springer Berlin Heidelberg, 2014, vol. 8724, pp. 656–671.

[182] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb 2012.

[183] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of hyperspectral and multispectral images," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3176–3180.

[184] ——, "Bayesian fusion of multispectral and hyperspectral images with unknown sensor spectral response," in *Proc. ICIP*, Paris, France, Oct 2014, pp. 698–702.

[185] Q. Wei, J. M. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Fusion of multispectral and hyperspectral images based on sparse representation," in *Proc. EUSIPCO*, Lisbon, Portugal, Sep. 2014, pp. 1577–1581.

[186] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proc. AISTATS*, vol. 5, Clearwater Beach, Florida, USA, 2009, pp. 320–327.

[187] R. R. Lederman and R. Talmon, "Common manifold learning using alternating-diffusion," Yale University, New Haven, CT, USA, Tech. Rep. YALEU/DCS/TR-1497, Mar. 2015.

[188] J. Liu, G. Pearlson, A. Windemuth, G. Ruaño, N. I. Perrone-Bizzozero, and V. D. Calhoun, "Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA," *Human Brain Mapping*, vol. 30, no. 1, pp. 241–255, Jan. 2009.

[189] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5940–5949, Nov. 2014.

[190] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Independent component analysis and (simultaneous) third-order tensor diagonalization," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2262–2271, 2001.

[191] G. Chabriel, M. Kleinsteuber, E. Moreau, H. Shen, P. Tichavský, and A. Yeredor, "Joint matrices decompositions and blind source separation: A survey of methods, identification, and applications," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 34–43, 2014.

[192] A. Cichocki, "Tensor decompositions: A new concept in brain data analysis?" *arXiv:1305.0395 [cs.NA]*, 2013.

[193] M. Schouteden, K. Van Deun, T. F. Wilderjans, and I. Van Mechelen, "Performing DISCO-SCA to search for distinctive and common infor-

mation in linked data," *Behavior Research Methods*, vol. 46, no. 2, pp. 576–587, Jun. 2013.
[194] M. Schouteden, K. Van Deun, S. Pattyn, and I. an Mechelen, "SCA with rotation to distinguish common and distinctive information in linked data," *Behavior research methods*, vol. 45, no. 3, pp. 822–833, 2013.
[195] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.
[196] E. Acar, A. J. Lawaetz, M. A. Rasmussen, and R. Bro, "Structure-revealing data fusion model with applications in metabolomics," in *Proc. EMBC'13*, Osaka, Japan, Jul. 2013, pp. 6023–6026.
[197] I. Domanov and L. De Lathauwer, "Canonical polyadic decomposition of third-order tensors: reduction to generalized eigenvalue decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 35, no. 2, pp. 636–660, Apr.–May 2014.
[198] A. Yeredor, "Performance analysis of GEVD-based source separation with second-order statistics," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 5077–5082, Oct. 2011.
[199] T. W. Anderson, *An introduction to multivariate statistical analysis*. John Wiley & Sons, 1958.
[200] M. Sørensen, I. Domanov, and L. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms—part II: Algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 3, pp. 1015–1045, 2015.
[201] Y. Guo and G. Pagnoni, "A unified framework for group independent component analysis for multi-subject fMRI data," *NeuroImage*, vol. 42, no. 3, pp. 1078–1093, Sep. 2008.
[202] M. Anderson, T. Adalı, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1672–1683, Apr. 2012.
[203] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, no. 0, pp. 1–17, 1986.
[204] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. ECCV*, ser. LNCS, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Copenhagen, Denmark: Springer Berlin Heidelberg, May 28–31, 2002, vol. 2350, pp. 736–750.
[205] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. COLT*, ser. COLT' 98. New York, NY, USA: ACM, 1998, pp. 92–100.
[206] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7–8, pp. 2031–2038, Dec. 2013.
[207] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. NIPS*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Lake Tahoe, NV, USA: Curran Associates, Inc., 2012, pp. 2222–2230.
[208] L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab v2.0," Jan. 2014. [Online]. Available: http://www.tensorlab.net/
[209] S. Sahnoun and P. Comon, "Joint source estimation and localization," *IEEE Trans. Signal Process.*, vol. 63, no. 10, pp. 2485–2495, May 2015.
[210] N. Sidiropoulos, E. Papalexakis, and C. Faloutsos, "Parallel randomly compressed cubes : A scalable distributed architecture for big tensor decomposition," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 57–70, Sep. 2014.
[211] A. Cichocki, "Era of Big Data processing: A new approach via tensor networks and tensor decompositions," *arXiv:1403.2048 [cs.ET]*, 2014.

Dana Lahat Dana Lahat received the Ph.D. in Electrical Engineering from Tel Aviv University, Israel, in 2013. She is currently a post-doctoral researcher at the Grenoble Images, Speech, Signals and Control Lab (GIPSA-lab), Grenoble, France. She has been awarded the Chateaubriand Fellowship of the French Government for the academic year 2007–2008. Her research interests include statistical and deterministic methods for signal and data processing, blind source separation, linear and multilinear algebra.

Tülay Adalı Tülay Adalı (S'89–M'93–SM'98–F'09) received the Ph.D. degree in Electrical Engineering from North Carolina State University, Raleigh, NC, USA, in 1992 and joined the faculty at the University of Maryland Baltimore County (UMBC), Baltimore, MD, USA, the same year. She is currently a Distinguished University Professor in the Department of Computer Science and Electrical Engineering at UMBC. Prof. Adalı assisted in the organization of a number of international conferences and workshops including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), the IEEE International Workshop on Neural Networks for Signal Processing (NNSP), and the IEEE International Workshop on Machine Learning for Signal Processing (MLSP). She was the General Co-Chair, NNSP (2001–2003); Technical Chair, MLSP (2004–2008); Program Co-Chair, MLSP (2008, 2009, and 2014), 2009 International Conference on Independent Component Analysis and Source Separation; Publicity Chair, ICASSP (2000 and 2005); and Publications Co-Chair, ICASSP 2008. Prof. Adalı chaired the IEEE Signal Processing Society (SPS) MLSP Technical Committee (2003–2005, 2011–2013), served on the SPS Conference Board (1998–2006), and the Bio Imaging and Signal Processing Technical Committee (2004–2007). She was an Associate Editor for IEEE Transactions on Signal Processing (2003–2006), IEEE Transactions on Biomedical Engineering (2007–2013), IEEE Journal of Selected Areas in Signal Processing (2010-2013), and Elsevier Signal Processing Journal (2007–2010). She is currently serving on the Editorial Boards of the Proceedings of the IEEE and Journal of Signal Processing Systems for Signal, Image, and Video Technology, and is a member of the IEEE Signal Processing Theory and Methods Technical Committee.

Prof. Adalı is a Fellow of the IEEE and the AIMBE, recipient of a 2010 IEEE Signal Processing Society Best Paper Award, 2013 University System of Maryland Regents' Award for Research, and an NSF CAREER Award. She was an IEEE Signal Processing Society Distinguished Lecturer for 2012 and 2013. Her research interests are in the areas of statistical signal processing, machine learning for signal processing, and biomedical data analysis.

Christian Jutten Christian Jutten (AM'92–M'03–SM'06–F'08) received the Ph.D. and Doctor ès Sciences degrees in signal processing from Grenoble Institute of Technology (GIT), France, in 1981 and 1987, respectively. From 1982, he was an Associate Professor at GIT), before being Full Professor at University Joseph Fourier of Grenoble, in 1989. For 30 years, his research interests have been machine learning and source separation, including theory (separability, source separation in nonlinear mixtures, sparsity, multimodality) and applications (brain and hyperspectral imaging, chemical sensor array, speech). He is author or co-author of more than 85 papers in international journals, 4 books, 25 keynote plenary talks, and 190 communications in international conferences.

He has been visiting professor at Swiss Federal Polytechnic Institute (Lausanne, Switzerland, 1989), at Riken labs (Japan, 1996) and at Campinas University (Brazil, 2010). He was director or deputy director of his lab from 1993 to 2010, especially head of the signal processing department (120 people) and deputy director of GIPSA-lab (300 people) from 2007 to 2010). He was a scientific advisor for signal and images processing at the French Ministry of Research (1996–1998) and for the French National Research Center (2003–2006). Since May 2012, he is deputy director at the Institute for Information Sciences at French National Center of Research (CNRS) in charge of signal and image processing.

Christian Jutten was organizer or program chair of many international conferences, especially of the 1st International Conference on Blind Signal Separation and Independent Component Analysis in 1999. He has been a member of a few IEEE Technical Committees, and currently in "SP Theory and Methods" of the IEEE Signal Processing society. He received best paper awards of EURASIP (1992) and of IEEE GRSS (2012), and Medal Blondel (1997) from the French Electrical Engineering society for his contributions in source separation and independent component analysis. He is IEEE fellow (2008) and EURASIP fellow (2013). He is a Senior Member of the Institut Universitaire de France since 2008, with renewal in 2013. He is the recipient of a 2012 ERC Advanced Grant for a project on challenges in extraction and separation of sources (CHESS).