NeuroImage 102 (2014) 92–117

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Review

A statistically motivated framework for simulation of stochastic data fusion models applied to multimodal neuroimaging

Rogers F. Silva [a,b,*], Sergey M. Plis [a], Tülay Adalı [d], Vince D. Calhoun [a,b,c]

[a] The Mind Research Network, 1101 Yale Blvd., Albuquerque, NM 87106, USA
[b] Department of ECE, MSC01 1100, 1 University of New Mexico, Albuquerque, NM 87131, USA
[c] Department of CS, MSC01 1130, 1 University of New Mexico, Albuquerque, NM 87131, USA
[d] Department of CSEE, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Accepted 9 April 2014
Available online 18 April 2014

Keywords:
Multimodal
Simulation
Fusion
ICA
Stochastic
Copula
Multidimensional

ABSTRACT

Multimodal fusion is becoming more common as it proves to be a powerful approach to identify complementary information from multimodal datasets. However, simulation of joint information is not straightforward. Published approaches mostly employ limited, provisional designs that often break the link between the model assumptions and the data for the sake of demonstrating properties of fusion techniques. This work introduces a new approach to synthetic data generation which allows full-compliance between data and model while still representing realistic spatiotemporal features in accordance with the current neuroimaging literature. The focus is on the simulation of joint information for the verification of stochastic linear models, particularly those used in multimodal data fusion of brain imaging data.

Our first goal is to obtain a benchmark ground-truth in which estimation errors due to model mismatch are minimal or none. Then we move on to assess how estimation is affected by gradually increasing model discrepancies toward a more realistic dataset. The key aspect of our approach is that it permits complete control over the type and level of model mismatch, allowing for more educated inferences about the limitations and caveats of select stochastic linear models. As a result, impartial comparison of models is possible based on their performance in multiple different scenarios.

Our proposed method uses the commonly overlooked theory of copulas to enable full control of the type and level of dependence/association between modalities, with no occurrence of spurious multimodal associations. Moreover, our approach allows for arbitrary single-modality marginal distributions for any fixed choice of dependence/association between multimodal features. Using our simulation framework, we can rigorously challenge the assumptions of several existing multimodal fusion approaches.

Our study brings a new perspective to the problem of simulating multimodal data that can be used for ground-truth verification of various stochastic multimodal models available in the literature, and reveals some important aspects that are not captured or are overlooked by ad hoc simulations that lack a firm statistical motivation.

Contents

Abbreviations: SA, simulated annealing; MH, Metropolis–Hastings; Pdf, probability distribution function; FIT, fusion ICA toolbox; ISI, inter-symbol interference; ICDF, inverse cumulative distribution function.

* Corresponding author at: The Mind Research Network, 1101 Yale Blvd., Albuquerque, New Mexico 87106, USA. Fax: +52 505 272 8002.
E-mail addresses: rsilva@mrn.org (R.F. Silva), splis@mrn.org (S.M. Plis), adali@umbc.edu (T. Adalı), vcalhoun@unm.edu (V.D. Calhoun).

http://dx.doi.org/10.1016/j.neuroimage.2014.04.035
1053-8119/

Introduction

Multimodal data fusion by stochastic linear models of the form shown in Eq. (1) is becoming increasingly popular in neuroimaging research. Put simply, these models assume separability of each modality's data $\mathbf{X}_m$ into latent, unobservable stochastic variables $\mathbf{S}_m$ which are linearly mixed through an unknown mixing matrix $\mathbf{A}_m$, where $m = 1, \ldots, M$ and M is the total number of modalities available. Typically, multimodal data consists of a collection of multimodal features pooled from a number of subjects. These features are often the result of single-modality first-level (i.e., subject-level) analyses. Thus, most data fusion approaches are characterized as second-level analyses (i.e., pooled group analyses). The joint decompositions of the multimodal $\mathbf{X}_m$ can then be made sensitive to multimodal associations among the mixing matrices $\mathbf{A}_m$ (Calhoun et al., 2006a; Correa et al., 2010; Sui et al., 2011) or among the component variables $\mathbf{S}_m$ (Groves et al., 2011; Sui et al., 2010). Here, we consider that $\mathbf{A}_m$ is a $(N \times C)$ matrix and $\mathbf{S}_m$ is a $(C \times V)$ matrix, where N is the number of subjects, C is the number of joint multimodal components, and V is the number of data points (e.g., voxels or timepoints).

$$\mathbf{X}_m = \mathbf{A}_m \mathbf{S}_m, m = 1, \ldots, M. \quad (1)$$

Our interest is in the question "How does the performance of a multimodal model change as the properties of $\mathbf{A}_m$ and $\mathbf{S}_m$ become less compliant with the underlying assumptions?" Historically, multimodal fusion studies have relied on the simulation of artificial datasets to showcase new models but seldom to address this question. Surely, well-designed simulations are useful to "debug" new algorithms, and make comparison and proof-of-concept demonstrations. Nevertheless, we have only recently observed a greater interest in carefully designed neuroimaging simulations that could venture after particular limitations and caveats of some popular models (Allen et al., 2012; Erhardt et al., 2012). Unfortunately, multimodal simulation studies have time and again neglected the virtues from the field of computerized simulation (Banks, 1998). This is a well-developed area that aims at making accurate predictions about the evolution and behavior of complex systems over time, typically modeling changes and state transitions via differential equations. Evidently, the class of stochastic *linear* models that we consider here is not nearly as elaborate. Still, we believe some of the same principles, especially those related with ground-truth verification (Oberkampf and Roy, 2010), apply when synthetic multimodal data is to be generated from these simpler models. We would expect this to improve our ability to rigorously evaluate any limitations of such models and, thus, better address the question above.

Translating this into practice involves making complete and gradual assessments of each multimodal model in different scenarios, both model-inspired and realistic ones. In realistic designs, multimodal associations are simulated based on the current literature, using prior knowledge about how associations are formed in real data. Model-based designs, on the other hand, simulate associations in agreement with the chosen model, complying with all its assumptions and limitations. Generally, when real data is well understood, a realistic design can help select the model that best recovers real data properties, whereas a model-based design allows ground-truth verification of a model implementation. Current multimodal simulation approaches are, for the most part, a blend between these two types of design (see examples in Appendix A). To our knowledge, however, no multimodal investigation has attempted to procedurally explore the spectrum of designs spanned between these extremes. Our premise throughout this work is that the optimal procedure to study the strengths and limitations of any model is by starting with a faithful model-based design, and gradually add discrepancies and violations to specific assumptions toward more realistic designs. In practice, however, we have observed that designing general multimodal associations in a manner that enables *controlled* model discrepancies is a great challenge. Lack of control leads to ad hoc solutions that often are rigid and only reflect the assumptions of a single stochastic model. We believe that addressing these issues would considerably increase our ability to recognize the key limitations of each model, especially in the anticipated cases where real data fails to comply with the model assumptions.

With that in mind, we introduce a new framework for synthetic data generation flexible enough to comply with many stochastic linear models of the form given in Eq. (1) while admitting gradual, controlled