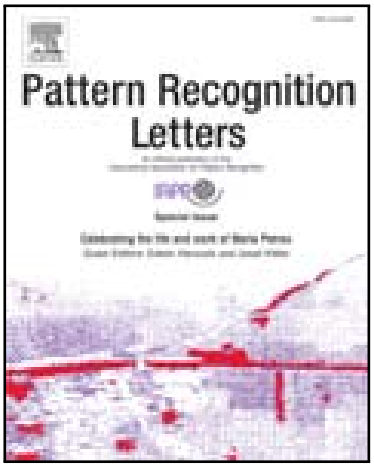
Accepted Manuscript



Structural network inference from time-series data using a generative model and transfer entropy

Zhihong Zhang, Genzhou Zhang, Zhonghao Zhang, Guo Chen, Yangbin Zeng, Beizhan Wang, Edwin R. Hancock

PII: S0167-8655(19)30168-0
DOI: https://doi.org/10.1016/j.patrec.2019.05.019
Reference: PATREC 7530

To appear in: *Pattern Recognition Letters*

Received date: 9 February 2019
Revised date: 6 May 2019
Accepted date: 25 May 2019

Please cite this article as: Zhihong Zhang, Genzhou Zhang, Zhonghao Zhang, Guo Chen, Yangbin Zeng, Beizhan Wang, Edwin R. Hancock, Structural network inference from time-series data using a generative model and transfer entropy, *Pattern Recognition Letters* (2019), doi: https://doi.org/10.1016/j.patrec.2019.05.019

Highlights

- We concentrate on the problem of describing the directed flow of information between nodes based on transfer entropy.
- We have developed a weighted directed supergraph based on the von Neumann entropy of a directed graph.
- Our model can improve the classification performance on fMRI brain connectivity data when the training data are limited.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Structural network inference from time-series data using a generative model and transfer entropy

Zhihong Zhang[a,2], Genzhou Zhang[b], Zhonghao Zhang[c], Guo Chen[c], Yangbin Zeng[a,2], Beizhan Wang[a,**], Edwin R. Hancock[d]

[a]*Xiamen University, Xiamen, China*
[b]*State Grid Shaanxi Electric Power Company, Xian, China*
[c]*State Grid Shaanxi Information and Telecommunication Company, LTD, Xian, China*
[d]*University of York, York, UK*

ABSTRACT

In this paper we develop a novel framework for inferring a generative model of network structure representing the causal relations between data for a set of objects characterized in terms of time series. To do this we make use of transfer entropy as a means of inferring directed information transfer between the time-series data. Transfer entropy allows us to infer directed edges representing the causal relations between pairs of time series, and has thus been used to infer directed graph representations of causal networks for time-series data. We use the expectation maximization algorithm to learn a generative model which captures variations in the causal network over time. We conduct experiments on fMRI brain connectivity data for subjects in different stages of the development of Alzheimer's disease (AD). Here we use the technique to learn class exemplars for different stages in the development of the disease, together with a normal control class, and demonstrate its utility in both graph multi-class and binary classifications. These experiments are showing the effectiveness of our proposed framework when the amounts of training data are relatively small.

Keywords: transfer entropy, supergraph, time series, network inference, expectation maximization algorithm

1. Introduction

A key goal of multivariate time-series data analysis is to infer a network which underpins the observed interactions between individual variables. This line of inquiry has permeated miscellaneous communities, including computational neuroscience, financial market modelling and social media analysis. Recently, transfer entropy (TE) has been recognised as a natural tool for inferring causal or directed relationships between pairs of variables. It has been widely used for example in the analysis of magnetoencephalography (MEG) (Vicente et al., 2011; Sokolova and Lapalme, 2009), electroencephalography (EEG) (Staniek and Lehnertz, 2008) and functional magnetic resonance imaging (fMRI) data (Hinrichs et al., 2006; Wibral et al., 2011). By contrast, mutual information (MI) (Kraskov et al., 2004) is the amount of shared information between individual variables while Pearson's correlation coefficient (PCC) (Lawrence and Lin, 1989) is a measure of the degree to which two random variables diverge from independence. Such measures reflect the symmetric connectivity of a functional network and lack the ability to capture asymmetric connectivity and describe the directional transfer of information flow between nodes. When compared with the closely related Granger causality (Granger, 1969), transfer entropy is characterized as model-free and capable of capturing non-linear relationships.

Not surprisingly the directed relationships between variables gauged by transfer entropy can be considered as directional edge connections in a causal network or directed graph. A considerable amount of literature has been published on the issue of representing data using graph structure. However little attention has been paid to the problem of how to capture structural variations based on edge connectivity in such representations. Existing methods for learning edge connectivity can be roughly categorized into two different classes: 1) spectral graph-based

**Corresponding author
e-mail: wangbz@xmu.edu.cn (Beizhan Wang)
[2]Co-first author

methods which are simple powerful yet lack of stability under slight perturbations in network structure (Luo et al., 2006); 2) probabilistic-based methods which posses the property of being underpinned by a well-knit probability theory. Considerable effort has been expended at describing the variability of edge connectivity pattern using such methods. For instance, Torsello and Hancock (Torsello and Hancock, 2006) have reconstructed trees using a Bernoulli distribution for node occurrences in samples of trees with unknown node correspondences. They adopt a minimum description length framework. This encodes the complexity for both a) of a set of tree-unions used to impose correspondences and infer connectivity for different classes of tree data and b) the number of mixture components needed to capture the class or cluster structure of the tree data. Wilson et al. (Wilson et al., 2015) have extended these ideas from trees to graphs. They have proposed a method for constructing a generative model represented by a supergraph from which a set of smaller sample graphs can be obtained by edit operations. Their method estimates a probability distribution for the occurrence of nodes and edges over the supergraph. This work is restricted to unweighted undirected networks.

Functional MRI is generally characterized as time series, and recently much of literature pays particular attention to capture underlying relationships between this kind of series, aimed at classifying subjects at different stages of AD. Existing methods may be roughly divided into two main categories, namely undirected and directed graph-based methods. The method of calculating Pearson's correlation coefficients, represented the connectivity between different brain regions, based on a sliding window approach has been proposed in (Chen et al., 2017, 2016). This kind of approach falls into the first categories. Khazaee et al. (Khazaee et al., 2017) proposed a directed graph model for identifying the changes in brain networks using multivariate Granger causality analysis. In our previous work (Wu et al., 2018), we employed histogram statistics and transfer entropy to measure causality relationships between time-series variables. Together these studies provide significant insights into the modeling of relationships between time-series variables in brain functional connectivity networks.

In this paper, we concentrate on the problem of describing the directed transfer or flow of information between nodes based on transfer entropy. Using transfer entropy, we extend the work of Wilson et al. from unweighted undirected graphs to a weighted directed supergraph model, and then propose a novel framework that combines the supergraph with transfer entropy. This framework is capable of not only effectively inferring fMRI brain connectivity structure, but also achieve significant improvements in classification accuracy for the publicly available Alzheimers Disease Neuroimaging Initiative (ADNI) fMRI dataset [3] .

2. Material

In this section we present the terminology and notation which underpin our study.

[3] http://adni.loni.usc.edu/

2.1. Transfer entropy

Entropy as a well-known information theoretic concept which measures of the average uncertainty or equivocation in a system. Specifically, if we take the expectation of the information according to the probability distribution $p(x)$, we end up with the Shannon entropy (Shannon, 1948):

$$H(X) = -\sum_{x} p(x) \log p(x). \tag{1}$$

The mutual information (Shannon, 1948) of two discrete random variables X and Y with the joint probability distribution $p(x, y)$ is a measure of their statistical dependence. In terms of probabilities, we take the form:

$$I(X;Y) = \sum_{x}\sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2}$$

Note that the mutual information is symmetric, i.e., $I(X;Y) = I(Y;X)$. In contrast, transfer entropy is a causally asymmetric measure of information transfer between two random processes. To frame this mathematically, we introduce the notation $X_n^{(k)} = \{X_{n-k+1}, ..., X_{n-1}, X_n\}$ and $Y_n^{(l)} = \{Y_{n-l+1}, ..., Y_{n-1}, Y_n\}$ to denote the k- and l-length history of the variables X and Y, up to and including time step n, which have realizations $x_n^{(k)} = \{x_{n-k+1}, ..., x_{n-1}, x_n\}$ and $y_n^{(l)} = \{y_{n-l+1}, ..., y_{n-1}, y_n\}$, respectively. In (Schreiber, 2000), Schreiber et al. define the transfer entropy as the reduction of uncertainty in a destination process that results from knowing the source process in the context of the causal past of the destination. This yields the following definition of transfer entropy:

$$T_{Y\to X} = \sum p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log \frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})}. \tag{3}$$

This is the central concept in measuring a directed edge connectivity. While the mathematical formulation above of the transfer entropy is relatively straightforward, in practice accurately estimating its value from time-series data is very challenging. The main reason is that it is highly sensitive to the type and quality of the available data. We thus discuss the various types of estimators available.

Gaussian estimator. The simplest estimator uses a multivariate Gaussian model for the random variables X of d dimensions, and the corresponding average entropy can be defined as (Cover and Thomas, 2012):

$$H(X) = \frac{1}{2} \ln((2\pi e)^d |\Omega|), \quad \Omega = \overline{XX^T}, \tag{4}$$

where the overbar denotes an average over the statistical ensemble. Since Eq. 4 sidesteps the computation of probability density functions (PDFs), the local entropy can be obtained by reconstructing the probability of a given observation sequence x in a multivariate process using the covariance matrix Ω (Lizier, 2014):

$$p(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Omega|}} \exp\left(-\frac{1}{2}(x-\mu)\Omega^{-1}(x-\mu)^T\right). \tag{5}$$

As above, where the observations used for the corresponding PDFs are from the whole time series of the processes then the transfer entropy, $T_{Y\to X}$, is the expectation of the local transfer values:

$$T_{Y\to X} = E\left(\log \frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})}\right). \tag{6}$$

Note that the Gaussian estimator is fast ($O(Nd^2)$) and parameter-free, but suffers from the limitation of assuming linear interactions between variables.

Kernel estimator. Schreiber et al. (Schreiber, 2000) proposed a approximate solution of the Eq. 3, and the joint PDF $\hat{p}_r(x_{n+1}, x_n^{(k)}, y_n^{(l)})$ is estimated by a kernel function Θ,

$$\begin{aligned}\hat{p}_r(x_{n+1}, x_n, y_n) &= \frac{1}{N}\sum_{n'=1}^{N}\Theta\left(\left\|\begin{pmatrix} x_{n+1} - x_{n'+1} \\ x_n - x_{n'} \\ y_n - y_{n'}\end{pmatrix}\right\| - r\right), \\ \Theta(x) &= \begin{cases}1, & x > 0 \\ 0, & x \le 0\end{cases},\end{aligned} \tag{7}$$

where the norm $|\cdot|$ is the maximum distance. Unlike Gaussian estimators, kernel estimation is model-free and capable of capturing non-linear relationships, although it requires a greater computational complexity $O(N^2)$.

Kraskov-Stögbauer-Grassberger (KSG) estimator. Initially, Kraskov (Kraskov, 2004) suggested that the transfer entropy (Eq. 3) is equal to the difference of two mutual information quantities:

$$T_{Y\to X} = I(X_{n+1}, X_n^{(k)}; Y_n^{(l)}) - I(X_n^{(k)}; Y_n^{(l)}). \tag{8}$$

Here the above expression leads to an over estimation for transfer entropy. However, this limitation has been addressed by extending the KSG estimation (algorithm 1) to conditional mutual information in (Kraskov et al., 2004; Frenzel and Pompe, 2007). Hence, the transfer entropy estimator can be rewritten as:

$$T_{Y\to X} = \psi(k) - E\left[\psi(\eta_{x_n^{(k)}} + 1) - \psi(\eta_{x_{n+1}x_n^{(k)}} + 1) - \psi(\eta_{y_n^{(l)}x_n^{(k)}})\right]. \tag{9}$$

Here ψ denotes the digamma function, ε is the max norm to the k-th nearest neighbor in the full $\{x_{n+1}, y_n^{(l)}, x_n^{(k)}\}$ space and $\{\eta_{x_n^{(k)}}, \eta_{x_{n+1}x_n^{(k)}}, \eta_{y_n^{(l)}x_n^{(k)}}\}$ are the neighbour counts strictly within max norms of ε in the $\{x_n^{(k)}\}, \{x_{n+1}, x_n^{(k)}\}$ and $\{y_n^{(l)}, x_n^{(k)}\}$ spaces, respectively.

KSG estimation inherits the non-linear and model-free characteristics of kernel estimation. Being effectively parameter-free it benefits from the stability to the choice of k. Despite its relatively expensive computation which requires time $O(kN^2)$, KSG estimation represents the seminal solution to estimating transfer entropy and measuring directed connectivity between time-series variables.

2.2. Generative model

We consider a problem of learning a generative model from a set of sample graphs by matching them to a so-called supergraph that characterizes the high-level structural information contained within the graphs. To frame this formally, we now commence to defining some notation. We use the notation $\mathcal{G} = \{G_1, ..., G_i, ..., G_N\}$ to denote the set of sample graphs from which we aim to learn the supergraph, where $G_i = (V_i, E_i)$ is the i-th graph with the set of nodes, V_i, and the set of edges, E_i. Similarly, the supergraph is represented by $\mathcal{F} = (V_\mathcal{F}, E_\mathcal{F})$. Further, we represent the structural information of the i-th graph, G_i, using a $|V_i| \times |V_i|$ weighted adjacency matrix D^i and that of supergraph, $\mathcal{F}$, using a $|V_\mathcal{F}| \times |V_\mathcal{F}|$ weighted adjacency matrix M. Clearly, we have $D^i_{ab} \in (0, 1]$, $(a, b) \in E_i$ and $M_{\alpha\beta} \in (0, 1]$, $(\alpha, \beta) \in E_\mathcal{F}$. We also define a set of assignment matrix $\mathcal{S} = \{S^1, ..., S^i, ..., S^N\}$, where S^i is of size $|V_i| \times |V_\mathcal{F}|$ and its elements indicate the corresponding structure matching between the graph G_i and supergraph $\mathcal{F}$ as follows:

$$S^i_{a\alpha} = \begin{cases} 1 & \text{if } f(a) = \alpha, \\ 0 & \text{otherwise,}\end{cases} \tag{10}$$

where the mapping function $f(a) = \alpha$ implies that the node $a \in V_i$ is assigned to the node $\alpha \in V_\mathcal{F}$.

Having established the necessary notation, we now proceed to develop the generative model. The idea underpinning the probabilistic framework of the generative model is that one maximizes a posteriori probability of the observed graph G_i given the supergraph $\mathcal{F}$ and assigned matrix S^i. According to (Luo and Hancock, 2001; Wilson et al., 2015), the posterior probability can be represented by

$$\begin{aligned}P(G_i|\mathcal{F}, S^i) &= \prod_{a\in V_i}\sum_{\alpha\in V_\mathcal{F}} K^i_a \exp\left(\mu \sum_{b\in V_i}\sum_{\beta\in V_\mathcal{F}} D^i_{ab} M_{\alpha\beta} S^i_{b\beta}\right), \\ \mu &= \ln\frac{1-P_e}{P_e}, \quad K^i_a = P_e^{|V_i|\cdot|V_\mathcal{F}|} B^i_a.\end{aligned} \tag{11}$$

Here P_e is the error of relation matching between the nodes of an observed graph and those of the sueprgraph, and B^i_a is a probability of observing a node a in graph G_i, its value depends only on the identity of the node a. The conditional likelihood above is appropriate for both undirected and directed graphs, and also gauges the difference between the two graphs.

Under the assumption that the graphs in $\mathcal{G}$ are independent from each other, the conditional likelihood over the set of observed graphs has realizations:

$$\begin{aligned}P(\mathcal{G}|\mathcal{F}, S) &= \prod_{G_i\in\mathcal{G}} P(G_i|\mathcal{F}, S^i) \\ &= \prod_{G_i\in\mathcal{G}}\prod_{a\in V_i}\sum_{\alpha\in V_\mathcal{F}} K^i_a \exp\left(\mu \sum_{b\in V_i}\sum_{\beta\in V_\mathcal{F}} D^i_{ab} M_{\alpha\beta} S^i_{b\beta}\right).\end{aligned} \tag{12}$$

3. Weighted directed supergraph learning

Bearing in mind that in this work we focus only on directed graphs, the main objective of this section is therefore to demonstrate how to construct a weighted directed supergraph learning framework.

3.1. Minimum Description Length Coding

The minimum description length (MDL) principle is of paramount importance for learning the model that best codes the

observed data (Grünwald et al., 2005; Jorma, 1998). Motivated by (Wilson et al., 2015), we adopt a two-part MDL criterion to seek the optimal supergraph structure, resulting in a total coding length

$$\mathcal{L}(\mathcal{G},\mathcal{F}) = \mathcal{L}(\mathcal{G}|\mathcal{F}) + \mathcal{L}(\mathcal{F}), \tag{13}$$

where $\mathcal{L}(\mathcal{G}|\mathcal{F})$ is the code-length over the observed graphs given the supergraph and $\mathcal{L}(\mathcal{F})$ is the code-length of measuring the complexity of the supergraph. The optimal supergraph can hence be obtained by weighing the goodness-of-fit of the observed graphs against the complexity of the supergraph.

For the use of the two-part MDL principle, an original idea of computing the code-length of the observed graphs given the supergraph is to adopt an average of the negative logarithm of the likelihood function given in Eq. 12. As a result, we have

$$\begin{aligned}\mathcal{L}(\mathcal{G}|\mathcal{F}) &= -\frac{1}{|\mathcal{G}|}\ln P(\mathcal{G}|\mathcal{F},S)\\ &= -\frac{1}{|\mathcal{G}|}\ln\left\{\sum_{\alpha\in V_{\mathcal{F}}} K_a^i \exp\left(\mu\sum_{b\in V_i}\sum_{\beta\in V_{\mathcal{F}}} D_{ab}^i M_{\alpha\beta} S^i_{b\beta}\right)\right\}.\end{aligned} \tag{14}$$

Having explained how the first term in the MDL criterion is computed, we now proceed to measure the complexity of the supergraph. Empirically, counting the number of parameters in the model can be considered as a simple solution to measure the complexity of a model. However, some estimators such as the numbers of nodes or edges in a graph, do not work well for as measure of true graph complexity. To overcome this bottleneck, Han et al. (Wilson et al., 2015) have proposed an interesting measure of graph-model complexity, namely the von Neumann entropy, and developed an approximation to compute the complexity of the unweighted undirected supergraph which depends on the node degree combinations of constituent edges. Unfortunately, for weighted directed graphs this is not a viable proposition since it neither distinguishes between the in-degree and out-degree of nodes, not assigns weights to the nodes or edges.

Motivated by the well documented capabilities of the von Neumann entropy in characterizing structural properties of networks (Han et al., 2012; Anand et al., 2011), Ye et al. (Ye et al., 2014) have extended its computation to weighted directed graphs by distinguishing between the in- and out-degree of nodes, leading to the following expression for the directed graph entropy

$$\begin{aligned}H &= 1 - \frac{1}{|V_{\mathcal{F}}|} - \frac{1}{2|V_{\mathcal{F}}|^2}\left\{\sum_{(\alpha,\beta)\in E_{\mathcal{F}}}\frac{d_\alpha^{in}}{d_\beta^{in}d_\alpha^{out^2}} + \sum_{(\alpha,\beta)\in E_{\mathcal{F}1}}\frac{1}{d_\beta^{out}d_\alpha^{out}}\right\},\\ d_\alpha^{in} &= \sum_{\gamma\in V_{\mathcal{F}}} M_{\gamma\alpha},\ \ d_\alpha^{out} = \sum_{\gamma\in V_{\mathcal{F}}} M_{\alpha\gamma},\\ d_\beta^{in} &= \sum_{\gamma\in V_{\mathcal{F}}} M_{\gamma\beta},\ \ d_\beta^{out} = \sum_{\gamma\in V_{\mathcal{F}}} M_{\beta\gamma},\end{aligned} \tag{15}$$

where E is the set of all the edges and E_1 is the set of bidirectional edges. Hence, by adding together the two contributions to the code-length, the overall code-length (Eq. 13) can be rewritten as

$$\begin{aligned}\mathcal{L}(\mathcal{G},\mathcal{F}) &= \mathcal{L}(\mathcal{G}|\mathcal{F}) + \mathcal{L}(\mathcal{F})\\ &= -\frac{1}{|\mathcal{G}|}\ln\left\{\sum_{\alpha\in V_{\mathcal{F}}} K_a^i \exp\left(\mu\sum_{b\in V_i}\sum_{\beta\in V_{\mathcal{F}}} D_{ab}^i M_{\alpha\beta} S^i_{b\beta}\right)\right\} + 1\\ &\quad - \frac{1}{|V_{\mathcal{F}}|} - \frac{1}{2|V_{\mathcal{F}}|^2}\left\{\sum_{(\alpha,\beta)\in E_{\mathcal{F}}}\frac{d_\alpha^{in}}{d_\beta^{in}d_\alpha^{out^2}} + \sum_{(\alpha,\beta)\in E_{\mathcal{F}1}}\frac{1}{d_\beta^{out}d_\alpha^{out}}\right\}.\end{aligned} \tag{16}$$

Unfortunately, work aimed at directly estimating the code-length is intractable due to the mixture structure, and this motivates us to resort to the iterative expectation maximization (EM) algorithm.

3.2. Optimization with EM algorithm

Having posed the problem of learning a weighted directed supergraph as that of code length optimization, we now proceed to use the EM algorithm to locate the structural characteristics of the supergraph. Noting the equivalence of the minimization of the overall code-length (Eq. 16) and the maximization of its negative, we follow the MDL setting of the EM algorithm in (Figueiredo and Jain, 2002) and the weighted log-likelihood function in (Wilson et al., 2015; Luo and Hancock, 2001), leading to the following expression

$$\begin{aligned}&\mathcal{A}^{(n+1)}(\mathcal{G}|\mathcal{F},\mathcal{S}^{(n+1)})\\ &= \frac{1}{|\mathcal{G}|}\sum_{G_i\in\mathcal{G}}\sum_{a\in V_i}\sum_{\alpha\in V_{\mathcal{F}}} Q_{a\alpha}^{i,(n)}\left\{\ln K_a^i + \mu\sum_{b\in V_i}\sum_{\beta\in V_{\mathcal{F}}} D_{ab}^i M_{\alpha\beta}^{(n)} S^{i,(n+1)}_{b\beta}\right\}\\ &\quad - 1 + \frac{1}{|V_{\mathcal{F}}|} + \frac{1}{2|V_{\mathcal{F}}|^2}\left\{\sum_{(\alpha,\beta)\in E_{\mathcal{F}}}\frac{d_\alpha^{in}}{d_\beta^{in}d_\alpha^{out^2}} + \sum_{(\alpha,\beta)\in E_{\mathcal{F}1}}\frac{1}{d_\beta^{out}d_\alpha^{out}}\right\}.\end{aligned} \tag{17}$$

For the expression above, we observe that $\sum_{G_i\in\mathcal{G}}\sum_{a\in V_i}\sum_{\alpha\in V_{\mathcal{F}}} Q_{a\alpha}^{i,(n)}\ln K_a^i = \sum_{G_i\in\mathcal{G}}\sum_{a\in V_i}\ln K_a^i$, which contributes a constant amount. As a result the weighted log-likelihood function can be rewritten as

$$\begin{aligned}&\hat{\mathcal{A}}^{(n+1)}(\mathcal{G}|\mathcal{F},\mathcal{S}^{(n+1)})\\ &= \frac{1}{|\mathcal{G}|}\sum_{G_i\in\mathcal{G}}\sum_{a\in V_i}\sum_{\alpha\in V_{\mathcal{F}}}\sum_{b\in V_i}\sum_{\beta\in V_{\mathcal{F}}} Q_{a\alpha}^{i,(n)} D_{ab}^i M_{\alpha\beta}^{(n)} S^{i,(n+1)}_{b\beta}\\ &\quad - 1 + \frac{1}{|V_{\mathcal{F}}|} + \frac{1}{2|V_{\mathcal{F}}|^2}\left\{\sum_{(\alpha,\beta)\in E_{\mathcal{F}}}\frac{d_\alpha^{in}}{d_\beta^{in}d_\alpha^{out^2}} + \sum_{(\alpha,\beta)\in E_{\mathcal{F}1}}\frac{1}{d_\beta^{out}d_\alpha^{out}}\right\}.\end{aligned} \tag{18}$$

3.2.1. Maximization

The maximization step of the EM algorithm can be realized by computing the derivatives of $\hat{\mathcal{A}}$. This step involves a reformulation of both the structure of the supergraph and the assignment variables.

Updating assignment variables. We now commence by computing the partial derivative of the Eq. 18 with respect

to the element of the assignment matrix S^i, which has the form:

$$\frac{\partial\hat{\mathcal{A}}^{(n+1)}}{\partial S^{i,(n+1)}_{b\beta}} = \frac{1}{|\mathcal{G}|}\sum_{a\in V_i}\sum_{\alpha\in V_{\mathcal{F}}} Q^{i,(n)}_{a\alpha} D^i_{ab} M^{(n)}_{\alpha\beta}. \tag{19}$$

As a result, the variables appearing in the assignment matrix S^i can be derived using softmax update rule (Bridle, 1990)

$$S^{i,(n+1)}_{a\alpha} \leftarrow \frac{\exp\left(\frac{\partial\hat{\mathcal{A}}^{(n+1)}}{\partial S^{i,(n+1)}_{a\alpha}}\right)}{\sum_{\alpha'\in V_{\mathcal{F}}}\exp\left(\frac{\partial\hat{\mathcal{A}}^{(n+1)}}{\partial S^{i,(n+1)}_{a\alpha'}}\right)}. \tag{20}$$

Updating supergraph structure. Unlike the case with undirected graphs (Wilson et al., 2015), we consider the complexity of a weighted directed supergraph by using the expression for the von Neumann entropy of a weighted directed graphs. The partial derivative of the Eq. 18 with respect to the entry of adjacency matrix of the weighted directed supergraph $\mathcal{G}$ has the form:

$$\begin{aligned}\frac{\partial\hat{\mathcal{A}}^{(n+1)}}{\partial M^{(n)}_{\alpha\beta}} &= \frac{1}{|\mathcal{G}|}\sum_{G_i\in\mathcal{G}}\sum_{a\in V_i}\sum_{b\in V_i} Q^{i,(n)}_{a\alpha} D^i_{ab} S^{i,(n+1)}_{b\beta} \\ &- \frac{1}{2|V_{\mathcal{F}}|^2}\left\{\sum_{(\alpha,\beta)\in E_{\mathcal{F}}}\left(\frac{d^{in}_\alpha}{(d^{in}_\beta d^{out}_\alpha)^2} + \frac{2d^{in}_\alpha}{T^{in}_\beta d^{out^3}_\alpha}\right)\right. \\ &\left. + \sum_{(\alpha,\beta)\in E_{\mathcal{F}1}}\frac{1}{d^{out}_\beta d^{out^2}_\alpha}\right\}.\end{aligned} \tag{21}$$

Similarly, the softmax update equation takes the form:

$$M^{(n+1)}_{\alpha\beta} \leftarrow \frac{\exp\left(\frac{\partial\hat{\mathcal{A}}^{(n+1)}}{\partial M^{(n)}_{\alpha\beta}}\right)}{\sum_{(\alpha',\beta')\in E_{\mathcal{F}}}\exp\left(\frac{\partial\hat{\mathcal{A}}^{(n+1)}}{\partial M^{(n)}_{\alpha'\beta'}}\right)}. \tag{22}$$

3.2.2. Expectation

We next compute the a posteriori probabilities of the missing correspondence from nodes of an observed sample graph to those of the directed supergraph. This is done by applying the Bayes theorem, and we have

$$\begin{aligned} Q^{i,(n+1)}_{a\alpha} &= \frac{\exp\left(\sum_{b\in V_i}\sum_{\beta\in V_{\mathcal{F}}} D^i_{ab} M^{(n)}_{\alpha\beta} S^{i,(n)}_{b\beta}\right)\pi^{i,(n)}_\alpha}{\sum_{\alpha'\in V_{\mathcal{F}}}\exp\left(\sum_{b\in V_i}\sum_{\beta\in V_{\mathcal{F}}} D^i_{ab} M^{(n)}_{\alpha'\beta} S^{i,(n)}_{b\beta}\right)\pi^{i,(n)}_{\alpha'}}, \\ \pi^{i,(n)}_{\alpha'} &= \langle Q^{i,(n)}_{a\alpha}\rangle_{a'}. \end{aligned} \tag{23}$$

At this point, the updates of both assignment matrices and supergraph structure, and the re-estimation of the a posteriori probabilities can be interleaved and alternately performed until a convergence is reached.

4. The proposed framework for fMRI data

Here the main application of our transfer entropy framework is to analyze fMRI time-series data for various regions of the brain for subjects at different stages in the progression of Alzheimer's disease. The fMRI dataset derives from the publicly available ADNI database. We use data for 114 subjects included in fMRI dataset. These subjects can be divided into four categories in terms of the degree of development of the disease. These are a) a Healthy Control (NC) group of 43 subjects, b) a Healthy Control 2 (NC2) group of 17 subjects, c) an Early Mild Cognitive Impairment (EMCI) group of 17 subjects, and d) a Late Mild Cognitive Impairment (LMCI) group of 38 subjects. The fMRI data for each subject consists of time series of 116 brain regions (aka ROIs, regions of interest). The neural activity of brain regions is measured using time series of the blood oxygenation level-dependent (BOLD) signal, which is characterized by real-valued variables. Considerable effort has been expended aimed at developing effective methods for exploring the functional connectivity between ROIs based on the BOLD signals. These include the use of Pearson's correlation (Chen et al., 2017; Zhang et al., 2016), and partial correlation (Jie et al., 2014). However, these methods are confined to the measurement of undirected causality and result in symmetric relationships.

We, on the other hand, consider a problem of characterizing the functional connectivity of brain regions using transfer entropy. The BOLD signals from each voxel can be divided into multiple overlapping time-series segments using a sliding window approach to capture the non-stationary interactions between ROIs. Specifically, we denote L as the total length of the BOLD signals, W as the length of the sliding window, and t as the sliding step size. The number of segments is $P = \frac{L-W}{t}$. For the p-th segment, we proceed to calculate the transfer entropy between i-th ROI and j-th ROI using the Eq. 3, which denotes as Z^p_{ij}. Then we make use of the root mean square (RMS) to measure the degree of the information transfer between different ROIs, which is given by

$$\tilde{Z}_{ij} = \sqrt{\frac{\sum_p (Z^p_{ij})^2}{P}}. \tag{24}$$

We can, therefore, generate the transfer entropy matrix $\tilde{Z}$ for each subject. The elements of $\tilde{Z}$ imply the degree of the asymmetric connectivity and are real-valued. The transfer entropy matrix can thus be regarded as a representation of a weighted directed graph. The node pairs with weak observed evidence of functional connectivity due to noises of signal detection problems, may though have potential connectivity. Rather than assigning a binary connectivity index (Martin et al., 2016), we do not eliminate weak connections by thresholding. Instead, we iteratively update elements of the matrix via expectation-maximization. In this way we avoid the unnecessary loss of functional connectivity information in the inferred network. With the set of adjacency matrices to hand, we can learn a corresponding supergraph for each class of subjects by the method as demonstrated in Section 3. Such supergraphs enable us to effectively infer the structure of fMRI functional connectivity networks (more details are presented in Section 5).

5. Experiments

In this section we detail both the results and their analysis on a dataset extracted from fMRI scans of human brains for

Table 1: Multi-class classification results of both different methods and TE estimators in fMRI dataset

Method	micro-F1	macro-F1
TECA[Gaussian]	0.6263±0.0133	0.6384±0.016
TECA[Kernel]	0.6789±0.01	0.6058±0.0236
TECA[KSG]	0.6982±0.0295	0.6754±0.0302
Ours[Gaussian]	0.6491±0.0224	0.5933±0.0581
Ours[Kernel]	0.7123±0.03	0.6665±0.0316
Ours[KSG]	**0.7456±0.0196**	**0.7209±0.022**

subjects at various stages in the development of Alzheimer's disease. We commence by studying the convergence properties of the proposed framework, and then report performance on classification tasks compared with our previous work (Wu et al., 2018).

5.1. Convergence

The first aim in this study is to investigate the convergence properties of the proposed framework. We initialize the structural information of the supergraph with the median graph for each class, and the individual correspondence assignment matrices using graduated assignment (Gold and Rangarajan, 1996). Fig. 1 shows the convergence of the weighted directed supergraph for each class with iteration number, when measured in terms of a) the supergraph von Neumann entropy, b) the average data log-likelihood and c) the overall code-length. Fig. 1 indicates that the von Neumann entropy of the weighted directed supergraph increases steadily with iteration number. Moreover, the MCI group (EMCI and LMCI) have a greater von Neumann entropy than the NC group (NC and NC2). This implies that there is a more active functional connectivity between different ROIs in the MCI group. Similarly, the curves of the average of the log-likelihood show a steady increase with iteration number. The overall code-length obtained by Eq. 16 is reduced effectively using developed EM algorithm as illustrated in Fig. 1(c). Together, these preliminary results suggest that the proposed framework is capable of achieving a rapid convergence for weighted directed supergraph learning.

5.2. Classification in fMRI dataset

Our second aim is to evaluate the effectiveness of our weighted directed supergraph model for classifying out-of-sample subjects. The class-label assigned to the out-of-sample subjects is governed by the class supergraph which gives the maximum a posteriori probability computed by Eq. 11. For the fMRI dataset, we aim to 1) classify subjects according to one of the four developmental groups, 2) distinguish between samples belonging to the MCI group from those belonging to the NC group, and 3) distinguish between subjects of different developmental degree of the MCI group. In addition, we aim to determine which of the transfer entropy estimators, i.e., Gaussian, kernel, and KSG estimation, give the best results in the multi-class classification task. To provide some quantitative results for multi-class classification, we measure the fractions of true positive, true negative, false positive, and false negative, i.e. TP, TN, FP, and FN respectively. We have also employed the following two measures of precision and recall (Sokolova and Lapalme, 2009), namely micro-F1 and macro-F1:

$$micro-F1 = \frac{2 * recision_{\mu} * recall_{\mu}}{precision_{\mu} + recall_{\mu}},$$
$$macro-F1 = \frac{2 * recision_{M} * recall_{M}}{precision_{M} + recall_{M}},$$

where

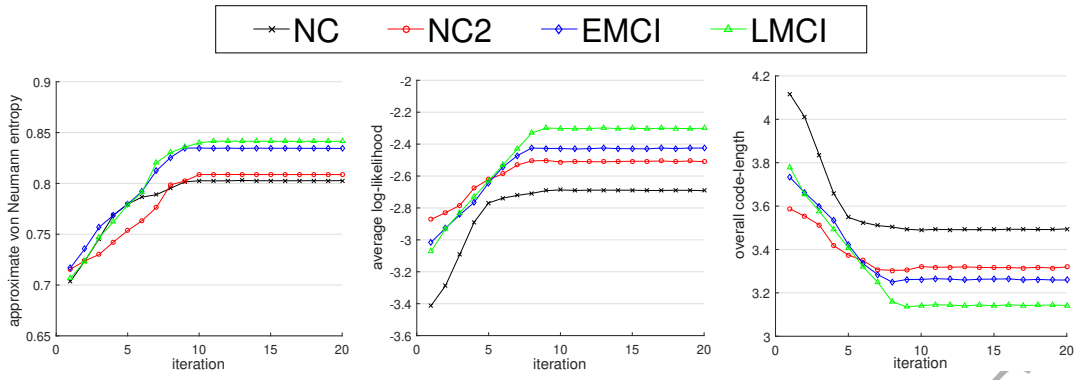
$$precision_{\mu} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l}(TP_i + FP_i)}, \quad recall_{\mu} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l}(TP_i + FN_i)},$$
$$precision_{M} = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i+FP_i}}{l}, \quad recall_{M} = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i+FN_i}}{l}.$$

Here l is the total number of categories or classes and i is the measured index corresponding to the category, e.g., TP_i represents the true positive count of the i-th class. The higher these index value, the better the performance of distinguishing the different degree of disease severity. For the different methods studied, the average micro-F1, macro-F1, and their standard error computed over 5 trials of 5-fold cross validation, resulting from classifying subjects in fMRI dataset, are shown in Tab. 1. The highest metric value is shown in bold. Tab. 1 shows that the newly developed method outperforms the previous approach based on transfer entropy component analysis (TECA) for all estimators of transfer entropy. It should be pointed out that the models based on KSG estimation outperform those based on Gaussian or kernel estimations. There are several possible explanations for this result. Firstly, Gaussian estimation is limited by the assumption of linear interactions between variables. Unfortunately, this assumption fails to capture fMRI time-series data. Secondly, although kernel estimation has can capture nonlinear relationships, it is still sensitive to the parameter choice for r in Eq. 7. KSG estimation, on the other hand, eradicates these limitations and thus achieves significant improvements in classification performance.

For binary classification we employed the following five indices (Sokolova and Lapalme, 2009): accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and F1-score, which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN},$$
$$Sensitivity = \frac{TP}{TP + FN},$$
$$Specificity = \frac{TN}{FP + TN},$$
$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right),$$
$$F1-score = \frac{2 * TP}{2 * TP + FN + FP}.$$

Here accuracy measures the classification rate which gives the fraction of correct samples over all classes and subjects, and the sensitivity and specificity indicate the proportions of positive samples and negative samples correctly classified, respectively. The F1-score denotes the relations between positive labels of



(a) complexity of weighted directed supergraph (b) code-length of observed graphs (c) overall code-length

Fig. 1: Convergence of the proposed framework

Table 2: Performance of binary classification for different methods based on KSG estimator

Method	Accuracy	Sensitivity	Specificity	AUC	F1-score
LMCI vs. EMCI					
TECA[KSG]	0.7333±0.0211	**0.9158±0.0343**	0.3±0.0523	0.6079±0.0229	0.8284±0.0159
Ours[KSG]	**0.7815±0.0606**	0.8±0.0546	**0.7375±0.0815**	**0.7688±0.0659**	**0.837±0.047**
NC vs. MCI					
TECA[KSG]	0.8526±0.019	**0.8433±0.0091**	0.863±0.0361	0.8531±0.0198	0.8578±0.0164
Ours[KSG]	**0.8596±0.0277**	0.763±0.0442	**0.9467±0.0217**	**0.8548±0.0284**	**0.8767±0.0231**
NC vs. EMCI					
TECA[KSG]	0.8265±0.216	0.7913±0.0415	0.8423±0.0325	0.8611±0.0148	0.8238±0.0324
Ours[KSG]	**0.8444±0.0176**	**0.8203±0.0221**	**0.8856±0.0147**	**0.8641±0.0324**	**0.8561±0.0411**

subjects and those given by a classifier. Tab. 2 gives the values of the measures together with their standard error. These were obtained with 5 trials of 5-fold cross validation, in two different binary classification tasks, i.e., LMCI vs. EMCI group, NC vs. MCI group, and NC vs. EMCI group. Results are shown for our novel transfer entropy method and TECA, both based on KSG estimation. For the LMCI vs. EMCI classification task, our proposed model performs better than the alternative method on all the metrics except for the sensitivity. This result may be explained by the fact that the TECA method tends to classify the out-of-sample subjects into the LMCI category. In other words, the TECA method is unable to sidestep the difficulty of the imbalance between samples of the two groups. The NC vs. MCI classification is consistent with the previous work published in (Wu et al., 2018). Regarding the NC vs. EMCI experiment, our proposed method is capable of achieving better performance despite of highly imbalanced samples.

These experimental results confirm the effectiveness of the proposed method and reveal that our method outperforms alternative methods in both multi-class classification and binary classification tasks for fMRI dataset.

6. Conclusion

In this paper, we have developed a weighted directed supergraph based on the von Neumann entropy of a directed graph. We have combined it with transfer entropy to infer a weighted directed network from fMRI time-series data. One of the more significant findings to emerge from this study is that the proposed model can effectively improve the classification accuracy for both multi-class classification and binary classification. The second major finding is that our work offers some important insights into understanding the use of transfer entropy with different estimators to measure asymmetric information flow between time-series variables. Further research should be done to investigate the effect of the proposed framework in larger and more complex datasets.

Acknowledgment

This work is supported by the Research Funds of State Grid Shaanxi Electric Power Company and State Grid Shaanxi Information and Telecommunication Company.

References

Anand, K., Bianconi, G., Severini, S., 2011. Shannon and von neumann entropy of random networks with heterogeneous expected degree. Physical Review E 83, 036109.

Bridle, J.S., 1990. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters, in: Advances in neural information processing systems, pp. 211–217.

Chen, X., Zhang, H., Gao, Y., Wee, C.Y., Li, G., Shen, D., Initiative, A.D.N., 2016. High-order resting-state functional connectivity network for mci classification. Human brain mapping 37, 3282–3296.

Chen, X., Zhang, H., Zhang, L., Shen, C., Lee, S.w., Shen, D., 2017. Extraction of dynamic functional connectivity from brain grey matter and white matter for mci classification. Human brain mapping 38, 5019–5034.

Cover, T.M., Thomas, J.A., 2012. Elements of information theory. John Wiley & Sons.

Figueiredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. IEEE Transactions on pattern analysis and machine intelligence 24, 381–396.

Frenzel, S., Pompe, B., 2007. Partial mutual information for coupling analysis of multivariate time series. Physical review letters 99, 204101.

Gold, S., Rangarajan, A., 1996. A graduated assignment algorithm for graph matching. IEEE Transactions on pattern analysis and machine intelligence 18, 377–388.

Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society , 424–438.

Grünwald, P.D., Myung, I.J., Pitt, M.A., 2005. Advances in minimum description length: Theory and applications. MIT press.

Han, L., Escolano, F., Hancock, E.R., Wilson, R.C., 2012. Graph characterizations from von neumann entropy. Pattern Recognition Letters 33, 1958–1967.

Hinrichs, H., Heinze, H., Schoenfeld, M., 2006. Causal visual interactions as revealed by an information theoretic measure and fmri. NeuroImage 31, 1051–1060.

Jie, B., Zhang, D., Gao, W., Wang, Q., Wee, C.Y., Shen, D., 2014. Integration of network topological and connectivity properties for neuroimaging classification. IEEE transactions on biomedical engineering 61, 576–589.

Jorma, R., 1998. Stochastic complexity in statistical inquiry. volume 15. World scientific.

Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A., Initiative, A.D.N., et al., 2017. Classification of patients with mci and ad from healthy controls using directed graph measures of resting-state fmri. Behavioural brain research 322, 339–350.

Kraskov, A., 2004. Synchronization and Interdependence Maesures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data. Ph.D. thesis. Universität Wuppertal, Fakultät für Mathematik und Naturwissenschaften .

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Physical review E 69, 066138.

Lawrence, I., Lin, K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics , 255–268.

Lizier, J.T., 2014. Measuring the dynamics of information processing on a local scale in time and space, in: Directed information measures in neuroscience. Springer, pp. 161–193.

Luo, B., Hancock, E.R., 2001. Structural graph matching using the em algorithm and singular value decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1120–1136.

Luo, B., Wilson, R.C., Hancock, E.R., 2006. A spectral approach to learning structural variations in graphs. Pattern Recognition 39, 1188–1198.

Martin, T., Ball, B., Newman, M.E., 2016. Structural inference for uncertain networks. Physical Review E 93, 012306.

Schreiber, T., 2000. Measuring information transfer. Physical review letters 85, 461.

Shannon, C.E., 1948. A mathematical theory of communication. Bell system technical journal 27, 379–423.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45, 427–437.

Staniek, M., Lehnertz, K., 2008. Symbolic transfer entropy. Physical Review Letters 100, 158101.

Torsello, A., Hancock, E.R., 2006. Learning shape-classes using a mixture of tree-unions. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 954–967.

Vicente, R., Wibral, M., Lindner, M., Pipa, G., 2011. Transfer entropya model-free measure of effective connectivity for the neurosciences. Journal of computational neuroscience 30, 45–67.

Wibral, M., Rahm, B., Rieder, M., Lindner, M., Vicente, R., Kaiser, J., 2011. Transfer entropy in magnetoencephalographic data: quantifying information flow in cortical and cerebellar networks. Progress in biophysics and molecular biology 105, 80–97.

Wilson, R., Hancock, E., et al., 2015. Generative graph prototypes from information theory. IEEE Transactions on Pattern Analysis & Machine Intelligence , 1–1.

Wu, M., Zeng, Y., Zhang, Z., Hong, H., Xu, Z., Cui, L., Bai, L., Hancock, E.R., 2018. Directed network analysis using transfer entropy component analysis, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer. pp. 491–500.

Ye, C., Wilson, R.C., Comin, C.H., Costa, L.d.F., Hancock, E.R., 2014. Approximate von neumann entropy for directed graphs. Physical Review E 89, 052804.

Zhang, H., Chen, X., Shi, F., Li, G., Kim, M., Giannakopoulos, P., Haller, S., Shen, D., 2016. Topographical information-based high-order functional connectivity and its application in abnormality detection for mild cognitive impairment. Journal of Alzheimer's Disease 54, 1095–1112.