Context Effects in Sentence Reading

Jonathan Mei
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
jmei@cmu.edu

Mariya Toneva
Center for the Neural Basis of Cognition
Carnegie Mellon University
Pittsburgh, PA 15213
mariya@cmu.edu

Abstract

The current work investigates the effect of context on the neural representation of concepts, measured by MEG during reading of sentences in active and passive voice. Two approaches to incorporating context are proposed − one by including estimates of the functional connectivity of the MEG sensors, and the other by creating a joint representation of the sentences using an HMM. The suggested methods are evaluated against non-context models in two tasks − classifying sentences as active or passive and identifying pairs of active and passive sentences that exemplify the same concept. Both supervised and unsupervised learning frameworks are presented, as well as the effect of the number of hidden states in the unsupervised framework on the task accuracies.

1 Introduction

Attaining a comprehensive understanding of how concepts are represented in the brain is crucial for implementing a computer program that understands simple language. Recent work [1-3] lays the foundation for understanding and modeling concept representation by putting forth generative theories of how single-word and simple phrasal concepts are formed. The natural next step is to investigate and model context effects in language that may modulate the neural representation of these concepts. The current work proposes two methods of incorporating context in models of concept representation and evaluates their performance against non-context models.

The models we investigate are based on the neural activity during sentence reading recorded by Magnetoencephalography (MEG). MEG is becoming increasingly popular in studying brain activity in language comprehension because of its high temporal resolution that enables millisecond-by-millisecond tracking of the evolution of neural representations. In line with neuroscience research, we assume that the underlying true representation of a concept gives rise to the MEG activity. In this way, we can think of the MEG activity as a function of the true representation of the concept.

However, the true concept representation is not available, so we must find an appropriate estimate. There are several popular methods for approximating the representation of concepts. The simplest one is to use the raw MEG recordings during the presentation of a concept as the estimate. The drawback of this method is that it is not generative. Another approach is to model the representation of concepts as a combination of the semantic features of the concept. For a single-word noun concept, a semantic feature can be computed as the occurrence of the word within a large text corpus that captures the typical use of this word in English text [1]. These semantic features are motivated by the hypothesis that the neural basis of the semantic representation is related to the distributional properties of nouns in a corpus of the language. While this approach is generative, it uses a set number of semantic features. It is not clear exactly what semantic features and how many of them may be relevant to the brain, so this approach may miss important underlying factors in the representation.

The first method investigated by the current work not only integrates context into the concept representation model, but also avoids the problem of semantic features while maintaining a generative ability. This approach estimates the functional connectivity of the MEG sensors. The functional connectivity encodes information about the state of the brain network and describes its evolution through time. The brain state is modulated by past stimuli, and so the functional connectivity may also be affected by the context created by these past stimuli. This work seeks to explore how usefully functional connectivity can be used to quantify context effects.

Estimating functional connectivity as the concept representation is beneficial not only to neuroscience, but also to machine learning. There exist multiple methods for inferring functional connectivity [4-8] but there is no clear best method. Attempting to address the real-life problem of concept representation will inform the algorithm we select for functional connectivity estimation.

The second proposed method incorporates context through a joint model of the stimulus sentence. While others have considered words in the sentence as largely independent, the current study investigates the benefits of using such a joint model to examine the neural representation of concepts.

2 Methods

2.1 Dataset

The data used in this project was acquired by Tom Mitchell's brain imaging group. This data set includes MEG recordings from 306 sensors for 1 subject who read 480 sentences one word at a time (each word presented for 300ms and followed by 200ms of rest). These 480 sentences include 15 repetitions of 16 pairs of propositions in both active and passive voice (one such pair is: "A dog found the peach" and "The peach was found by a dog"). Thus, the data set is a matrix of dimensions number of sensors $\times$ sample instances for all 480 sentences.

This data was selected because it allows us to investigate the representation of two different sentences (one active and one passive) that seem to illustrate the same concept. Given the nature of this data, the following two tasks are of interest:

1. classify sentences as active or passive
2. identify pairs of active and passive sentences that exemplify the same concepts

While the identification task is more relevant to the goal of investigating concept representations, conducting the classification task allows for a more complete picture of the kind of information that is contained in context.

2.2 Models

To determine the effect of context on modeling concept representation, we need to compare the performance of two models - one with and one without context. First, we describe the general framework of our two models. We then define their supervised and unsupervised instantiations.

Note that because the active sentences have only five words while the passive ones have seven, we consider only the first five words of all sentences in both the context and non-context models. If the full passive sentences are used instead, the models would heavily favor active sentences in the classification task, as five-state sequences are always at least as likely as seven-state sequences.

2.2.1 Hidden Markov Model

For our model that incorporates context, we use HMM to model the entire sentence. The dependencies through time formally capture the information available from context.

We assume that the first word type j is distributed according to the probability mass function (p.m.f.) $p(S_1 = j)$. We assume that word type j follows word k according to a time-invariant p.m.f. $p(S_i = j|S_{i-1} = k)$. We assume that for each word, the MEG data $\mathbf{x}_i$ is generated according to a multivariate Gaussian, $\mathbf{X}_i|S_i \sim \mathcal{N}(\mu_{s_i}, \Sigma_{s_i})$. we let MEG data $\mathbf{x}_i \in \mathbb{R}^N$.

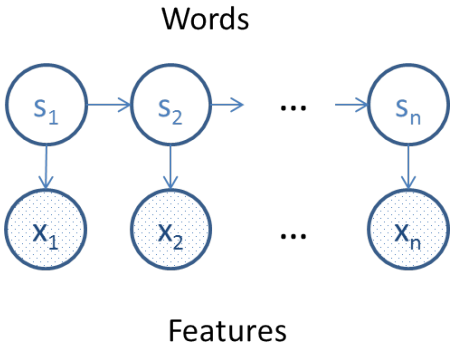

Figure 1: HMM for MEG data based on words in sentence

The graphical model representation of the sentence and the data are shown in figure 1. In our model, the parameters to learn are:

1. Probabilities $\alpha_j = p(S_1 = j)$ for discrete initial p.m.f.
2. Conditional probabilities $\theta_{jk} = p(S_i = j|S_{i-1} = k)$ for discrete transition p.m.f.
3. Means and covariances μ_j and Σ_j for continuous emission p.d.f. $f(\mathbf{X}_i|S_i = j)$

We further assume that Σ_j are diagonal matrices (i.e. corresponding to the assumption that the MEG sensor readings are generated independently). We employ a Maximum Likelihood estimate to compute α_j, θ_{jk}, and use sample means and variances to estimate μ_j and Σ_j. This assumption of diagonal Σ_j is to reduce the computational complexity and variance of the estimation.

2.2.2 Mixture of Gaussians

The model that we choose for comparison is a Mixture of Gaussians (MoG) model. This model choice makes the same assumptions about how data is generated from each state as the HMM but does not account for dependencies across time, which correspond to context. This makes the MoG the most comparable choice to the HMM. Formally, we assume that the word types are distributed as discrete p.m.f. $p(S_i = j)$, and each word is generated from each state according to multivariate Gaussian $\mathbf{X}_i|S_i \sim \mathcal{N}(\mu_{S_i}, \Sigma_{S_i})$.

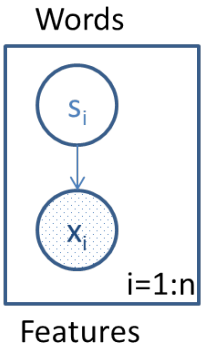

Figure 2: MoG model for MEG data based on words in sentence

The graphical model representation of the sentence and the data are shown in figure 2. The parameters to learn in the MoG model are:

1. Probabilities $\alpha_j = p(S_i = j)$ for discrete p.m.f.
2. Mean and covariances μ_j and Σ_j for continuous p.d.f $f(\mathbf{X}_i|S_i = j)$.

2.3 Learning frameworks

While we do have exact knowledge of true word types for each word, we do not know if there may be some better way to label the words for the two tasks of interest. Thus, we use both supervised and unsupervised learning frameworks to investigate both the performance of the models and their abilities to find the most informative labelings of "word type" for concept representation in sentence reading.

2.3.1 Supervised Framework

In the supervised setting, each sentence is composed of W words w_i of type $s_i \in \mathcal{S} = \{$'det', 'N', 'V', 'was', 'by'$\}$ for $i = 1, \ldots, W$ roughly corresponding to the part of speech of word w_i.

Here, the different word types contain the following words:

1. 'det' $= \{$'a', 'the'$\}$
2. 'N' $= \{$'dog', 'peach', 'student', 'school', 'hammer', 'door', 'doctor', 'monkey'$\}$
3. 'V' $= \{$'found', 'kicked', 'inspected', 'touched'$\}$
4. 'was' $= \{$'was'$\}$
5. 'by' $= \{$'by'$\}$

According to this model, the true form of every active and passive sentence in the data set can be described as follows:

Active sentences: det N V det N

Passive sentences: det N was V by

We use these five states for our supervised HMM model. However, we note that since the third word of the passive sentence is always 'was', while the third word in active sentences is always 'V', we train the MoG to differentiate between 'was' and 'V'. In this case, the MoG model is a simple Gaussian Naive Bayes classifier.

Both the HMM and Naive Bayes are trained on all but one sentence and tested on the held-out sentence, in a leave-one-out fashion. Once the HMM is trained, we use the Viterbi algorithm to estimate the most likely sequence of states for the test sentence. For the Naive Bayes, we use the ratio between the log likelihoods of the features using the estimated parameters for the active and passive sentences.

2.3.2 Unsupervised Framework

To examine the model itself, we turn to an unsupervised setting. Here, we vary the number of word types s_i up to $\mathcal{S} <= 16$ since we only have 16 unique words in our experiment. For both context and non-context models, we randomly initialize the state assignments s_i for all words w_i and can alternate between updating the hidden assignments and learning the model parameters as described in sections 2.2.1 and 2.2.2.

The training for both models is done on all sentences using Expectation-Maximization (EM) for the MoG model and using Gibbs sampling for the HMM. This choice of Gibbs sampling was made in order to allow for easy extension to more sophisticated frameworks we will discuss in section 5. After training, the Viterbi algorithm is again used to estimate the HMM state sequence, and the maximum likelihood estimate of each word is found individually for the MoG model state sequence.

2.4 Functional Connectivity

We determined that we would not use fused lasso [6] since it encourages matrices from neighboring words to be the same. This is not what we want when using the matrices to help distinguish between words, since we know a priori that neighboring words as well as contexts for neighboring time intervals are likely to be different.

Instead, we estimate functional connectivity using cross-correlation [7] and causal Graph Processes (CGP) based on Discrete Signal Processing on Graphs [8] and multivariate autoregressive models. The functional connectivity matrix based on using cross-correlation is computed,

$$[C_f]_{ij} = \frac{1}{T-1}\sum_{t=1}^{T}\frac{(y_i(t)-\bar{y}_i)(y_j(t)-\bar{y}_j)}{\sigma_i\sigma_j}$$

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_i(t)$$

$$\sigma_i = \frac{1}{T-1}\sum_{t=1}^{T}(y_i(t)-\bar{y}_i)^2$$

The connectivity computed using CGP is the result of an optimization problem,

$$\hat{\mathbf{C}}_f = \underset{\mathbf{A}}{\operatorname{argmin}}\frac{1}{T-M}\sum_{t=0}^{T-M}\|\mathbf{y}(t+M)-\bar{\mathbf{y}}(t+M)\|_2^2+\lambda_1\|\mathbf{A}\|_1$$

$$\bar{\mathbf{y}}(t+M) = \sum_{i=1}^{M} h_i(\mathbf{A})\mathbf{y}(t+M-i)$$

$$h_i(\mathbf{A}) = \sum_{j=0}^{i} c_{ij}\mathbf{A}^j$$

This model imparts the functional connectivity matrix $\mathbf{A}$ with a physical meaning as a spatially discretized approximation for a differential operator rather than as a measured statistic. We use an order $M = 2$ model to estimate the functional connectivity matrix using this method.

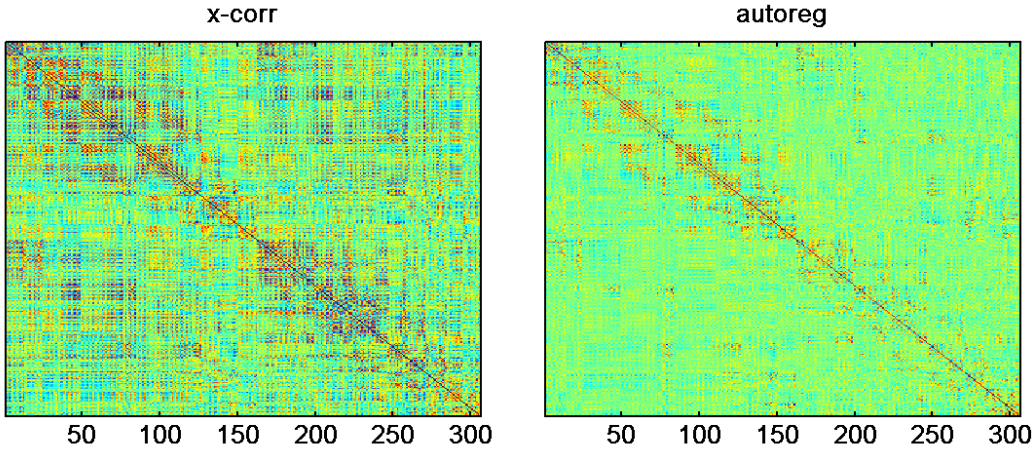

Figure 3: Cross-correlation (x-corr) based and CGP (autoreg) based functional connectivity matrices

Figure 3 shows an example of functional connectivity matrices estimated using the two methods described above. We note that cross-correlation is symmetric and not sparse, and represents indirect interactions rather than direct ones. As expected, the functional connectivity based on CGP provides sparser and more precise information. Note that only the relative values and not the absolute values matter in our modeling.

Using the data $\mathbf{Y} \in \mathbb{R}^{306\times250}$ from each word we estimate the functional connectivity $\mathbf{C}_f \in \mathbb{R}^{306\times306}$. To include the functional connectivity in our model, we then concatenate the vecotrized data and the vectorized functional connectivity matrix to form our feature vector $\mathbf{X}_i \in \mathbb{R}^{306*556\times1}$.

3 Results

3.1 Classifying Sentences as Active or Passive

In the supervised framework using the 5 states that roughly correspond to part-of-speech tags, we compare the Leave-one-out Cross-validation (LOOCV) classification accuracies of the Gaussian Naive Bayes and the HMM. The classification accuracies of the Naive Bayes are obtained by comparing the results of the log likelihood ratio test, as described in section 2.3.1, with the true sentence labels. The classification accuracies for the HMM are computed by calculating the Hamming distances between the predicted sequence for the training example and the true labeled sequence. Here, the Hamming distance between two state sequences equals the number of states that differ between these sequences.

Because there are 15 repetitions of each sentence, it is unclear how many repetitions (if any) should be averaged together into a single test example. The respective accuracies for the number of repetitions averaged together are displayed in Figure 4.

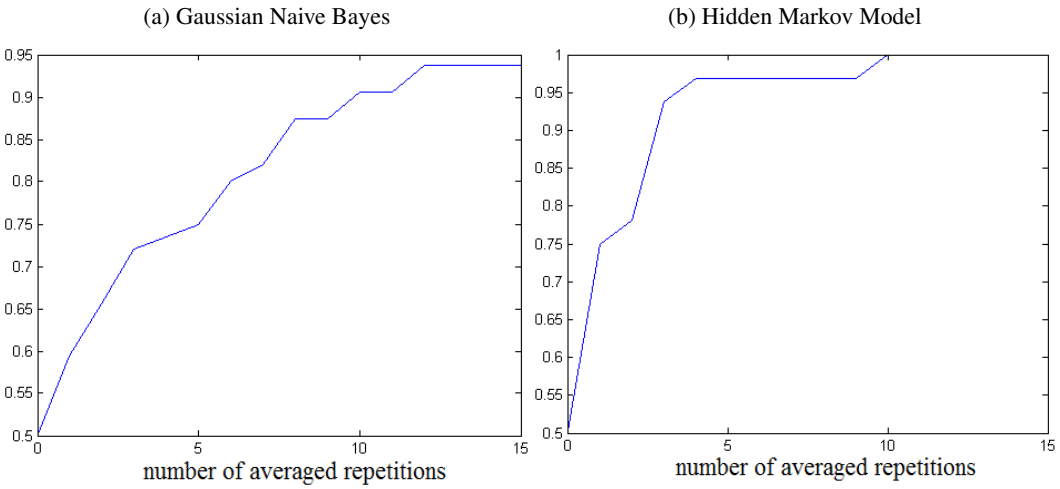

Figure 4: LOOCV classification accuracy verses number of repetitions averaged together into a single test example.

In the unsupervised framework, we first obtain the predicted state sequences for all 32 sentences from both the MoG and the HMM. Then, for each model we compute a distance matrix of the pairwise Hamming distances between the 32 estimated state sequences. Lastly, we perform spectral clustering on these distance matrices to automatically obtain two clusters for each distance matrix. The classification accuracy is computed based on how many passive and active sentences were correctly grouped into different clusters.

To address the question of what number of states is optimal for the classification task, we perform the experiment outlined above for up to 16 states, as described in section 2.3.2.

Evaluating the significance of the classification accuracies is conducted through a permutation test. In this test, we permute the correspondences between sentences and estimated state sequences and perform the outlined experiment. We obtain classification accuracies for 1000 permutation trials. The p-value of the original classification accuracy is computed by dividing the number of permutation trials that produce classification accuracies that are greater than or equal to the original by the number of permutation trials. The results of the permutation tests for those numbers of states which resulted in classification accuracies significantly different from chance are shown in Table 1.

	MoG			HMM		
	accuracy	p-value	states	accuracy	p-value	states
raw MEG only	68.75%	0.056	12	71.88%	0.009	10
raw MEG + x-corr	-	n.s.	-	-	n.s	-
raw MEG + autoreg	-	n.s.	-	-	n.s	-

Table 1: Accuracies for classifying sentences as active or passive using the three types of features, as described in section 2.4. Displayed are only those classification accuracies that are significantly different from chance or approaching significance as determined by a permutation test (n.s. $=$ not significant) and $\alpha = 0.05$.

3.2 Identifying Concept Pairs of Active and Passive Sentences

Because the identification task is likely to rely on finer distinctions between words than parts of speech, we use the unsupervised framework to perform this task. In addition, for this task we measure distance between sentences in a slightly different way because corresponding words in matching active and passive sentence pairs are in different positions in the sentences. We use a bag-of-words based representation, which is a vector of counts of the state sequence of a sentence, to compute pairwise distances between sentences. For a sentence in question, we compute its distance to every other sentence and sort this list of distances. We find the rank of the true match within this sorted list. We repeat this process for each sentence and average the ranks to compute rank accuracy.

Lastly, we perform permutation tests to determine the significance of the rank accuracies. The results of the permutation tests for those numbers of states which resulted in rank accuracies significantly different from chance are shown in Table 2.

	MoG			HMM		
	rank accuracy	p-value	states	rank accuracy	p-value	states
raw MEG only	-	n.s.	-	12.36	0.017	12
raw MEG + x-corr	12.53	0.054	10	12.38	0.048	11
				12.48	0.035	12
raw MEG + autoreg	11.94	0.036	9	-	n.s	-

Table 2: Rank accuracies for identifying a pair of active and passive sentences that exemplify the same concept using the three types of features, as described in section 2.4. Displayed are only those rank accuracies that are significantly different from chance or approaching significance as determined by a permutation test (n.s. $=$ not significant) and $\alpha = 0.05$.

4 Discussion

The LOOCV classification results from the supervised framework show that as the number of repetitions that are averaged into a single test example increases, the classification accuracy also increases. This result is not surprising because the inherent noise in the MEG recordings is increasingly averaged out. However, the interesting outcome is that the Naive Bayes needs 12 averaged repetitions to reach the classification accuracy of the HMM with 4 averaged repetitions per test example. In addition, the HMM reaches 100 accuracy, whereas the Naive Bayes obtains only 93.75%. Considering both of these results, we conclude that by using the joint model for this task, a neuroscientist can collect three times as much data in the same amount of imaging time without compromising accuracy. Because imaging time requires a nontrivial amount of resources while variability in stimuli is still strongly desirable, these results would be of interest to the neuroscience community.

Conducting the same task with the unsupervised framework replicates the supervised results that the joint model can classify sentences more accurately. HMM's classification accuracy is not only higher (71.88%) than the non-context model's (68.75%), but also more significant as shown by the permutation test. However, including the functional connectivity estimates does not result in significant classification accuracies for any number of states. These results suggest that either the

functional connectivity estimates add noise to the raw MEG data or that the way we incorporate them into the features is not sophisticated enough for the classification task.

The concept pair identification task reveals that the joint model continues to outperform the non-context model. Similarly to the classification task, while using the raw MEG features results in significant rank accuracies for the HMM, the same is much less the case for the MoG. Unlike the previous task, the identification task seems to benefit from incorporating the functional connectiity estimates into the features. Even though the cross correlation is a less sophisticated estimate of functional connectivity than the autoregressive method, it seems to perform better on this task. While it is difficult to draw conclusions about the benefit functional connectivity provides, these results suggest that the functional connectivity carries some information that is not readily available in the raw MEG data.

Lastly, another interesting outcome is the number of states that results in significant accuracies. It is notable that they are between 9 and 12, which is neither close to the number of states in the part-of-speech supervised framework (5) nor to the number of different words in the dataset (16). The meaning of these states and their number remain puzzling and warrant further investigation.

5 Further Work

In continuing this work, we have additional data from 3 other subjects available. Performing these analyses will likely show us more about the role of functional connectivity. What could be equally revealing is to study more sophisticated methods of incorporating the functional connectivity in the features for the models. Instead of using the functional connectivity directly as features, it may also make sense to use certain types of estimates of functional connectivity as implicit estimates of a highly structured covariance Σ in the Gaussian emission p.d.f.

It may be worth implementing a hierarchical Dirichlet process (HDP) to automatically estimate the number of states for the HMM and a Dirichlet Process to similarly estimate the number of states for the MoG model. In our current experiment, the number of unique words is limited to 16, but the HDP would allow the model to apply to larger data sets with more and longer sentences with a larger dictionary containing more unique words.

Most importantly, the estimated states in the unsupervised settings should be examined and understood. What these states correspond to may be what opens up additional understanding of how the brain comprehends and represents concepts through reading.

6 Conclusion

While incorporating an estimate of the functional connectivity into the concept representation model is not definitively proven to be beneficial, the joint model is shown to outperform the non-context model on both tasks. The proposed joint model is not only able to classify sentences as active or passive significantly better than chance, but also group corresponding concept pairs of active and passive sentences.

Acknowledgements

The authors would like to thank Tom Mitchell for access to the dataset used in the current work and guidance during the preliminary stages of the project.

References

[1] Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320(5880), 1191-1195.

[2] Fyshe, A., Talukdar, P., Murphy, B., and Mitchell, T. (2013). Documents and Dependencies: an Exploration of Vector Space Models for Semantic Composition. CoNLL-2013, 84.

[3] Chang, K. M. K., Cherkassky, V. L., Mitchell, T. M., and Just, M. A. (2009, August). Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation. In Proceedings of the

Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 638-646). Association for Computational Linguistics.

[4] Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echoplanar mri. Magnetic resonance in medicine, 34(4), 537-541.

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

[6] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1), 91-108.

[7] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. V. Essen, and M. E. Raichle, The human brain is intrinsically organized into dynamic, anticorrelated functional networks, PNAS, vol. 102, no. 27, pp. 96739678, Jul. 2005.

[8] A. Sandryhaila and J. M. F. Moura, Discrete Signal Processing on Graphs, IEEE Transactions on Signal Processing, vol. 61, no. 7, pp. 16441656, Apr. 2013.