Decoding Word Semantics from Magnetoencephalography Time Series Transformations

Alona Fyshe[1,2,3], Gustavo Sudre[2], Leila Wehbe[1,2], Brian Murphy[1], and Tom Mitchell[1,2]

[1] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Center for Neural Basis of Cognition, Carnegie Mellon University,Pittsburgh, PA 15213
[3] afyshe@cs.cmu.edu

Abstract. Neuroimaging techniques such as Magnetoencephalography have facilitated the careful study of perceptual and motor systems. These processes are largely feed-forward and bottom up, and the evoked responses are very consistent. Gaining a similarly strong understanding of higher-level cognitive thought has proven more difficult. The processes involved in higher-order thought appear to be spatially distributed, involve top-down cognitive influence and are not as tightly coupled to the stimulus. To deal with these complications and inconsistencies, we need a robust method for processing the MEG signal. In this study we explore several methods of processing the MEG signal and evaluate their utility for decoding the higher order cognitive process of noun comprehension.

1 Introduction

The processes involved in perceptual and motor neural systems are well studied, thanks to neuroimaging techniques such as Magnetoencephalography (MEG) and the use of animal models. The neural systems of perception and motor response are largely feed-forward and bottom up, and thus are very consistent in nature. This stability in response has allowed scientists to pinpoint the location and timing of neural responses, which remain fairly stable across subjects [3, 10].

Attempts to gain a similar understanding of higher-level cognitive thought have been less successful. Depending on the task, there may be no reasonable animal model. To further complicate matters, the processes involved with higher-order thought (e.g. attention) appear to involve a distributed network of brain areas, involve top-down cognitive influence, and incorporate neural activity not coupled to a stimulus [20, 4]. To deal with the inconsistent nature of the neural response we need robust processing methods. In this study we explore several processing techniques (transformations) of the MEG signal. We evaluate the utility of each transformed MEG signal for decoding the stimulus noun a subject is reading. The transformations we consider provide different representations of the underlying neural activity (e.g. phase of neural oscillations, average firing in a time window). As we uncover the relation between the stimulus and the recorded brain signals we begin to understand how the brain encodes information, and thus we elucidate the *neural code*. This paper's contributions are: 1) the exposition of a model that can, with high probability, correctly identify the word a person is reading based on the MEG signal 2) A statistical analysis of the effect of several MEG signal processing techniques on the accuracy of the word-prediction model.

We collected MEG data while 9 subjects viewed 60 concrete noun word/pictures, with 20 interleaved repetitions (single trials). During each of the 20 repetitions the subjects were asked to consider a different semantic question selected from the 218 semantic features described below. A full description of the data, including preprocessing steps, can be found at http://www.ml.cmu.edu/research/dap-papers/dap_fyshe.pdf. We also collected 218 semantic features for each of the 60 words used in this study. The semantic features are ratings $[1 \dots 5]$ in response to questions like “Do you hold it to use it?” and “Is it alive?”. This projection of the 60 words into a semantic space allows us to decompose the neural representation of a word into components related to each of the semantic features. Past work has shown that word properties (e.g. size, animacy) have differing effects on the MEG signal, across brain areas and points in time [18, 21]. We will explore when and where these effects happen, and also whether transformations of the MEG signal have an effect on the accuracy with which different semantic features can be decoded.

Zero Shot Learning While we have MEG data for only 60 words, there are tens of thousands of words in the English language, and new words are added continually. We would like to develop a system that can predict the word a person is reading even when we have not previously analyzed a MEG recording of that word. We accomplish this by decomposing our words into 218 semantic features and utilizing *Zero Shot Learning* [13].

Typically for this sort of word prediction problem, one might learn a n-way predictor $f(X) \rightarrow w$ so that we predict the word w based on the MEG data X. In our case, where $n = 60$, training such a predictor can be very difficult. Instead, we utilize a known mapping that transforms word w to its $m = 218$ dimensional semantic representation $\{s_1 \dots s_m\}$. In our case, s_i can take on 5 values that represent ratings. Then, we train m independent functions $f_1(X) \rightarrow s'_1, \dots, f_m(X) \rightarrow s'_m$ where s' represents the *predicted* value of a semantic feature. The predicted ratings are concatenated to produce a predicted semantic vector: $\vec{s'} = \{s'_1, \dots, s'_m\}$. We then define a function $d(\{s'_1 \dots s'_m\}, \{s_1, \dots, s_m\})$ that quantifies the dissimilarity between two semantic vectors. Any distance metric could be used here; we will use cosine distance. The final prediction is the word w with the semantic vector $\vec{s}_w$that minimizes the cosine distance: $w = \underset{w}{\operatorname{argmin}}\{d(\vec{s}_w, \vec{s'})\}$.

We train Zero Shot classifiers using each of the MEG transformations (described in the following sections) where regularized regression is used to learn the f_i. We examine the decodability of semantic features at different time intervals relative to stimulus onset, and at different frequencies, by training distinct classifiers for distinct time-frequency windows in the MEG data. Finally, we can combine all of the MEG transformations together to train one large regressor. Cross-validation and L_2 regularized regression are used to minimize overfitting problems.

Related Work Past work has explored a variety of MEG signal transformations for various tasks. For example, the power in gamma frequency bands (>30 Hz) has been shown to be an indicator of attention and memory, as well as the coordination of brain areas [8]. A common strategy when dealing with low SNR (signal to noise ratio) is to take a time windowed average of the signal. In [2], the average amplitude in several time windows were used as input to an SVM classifier which could distinguish living vs. non-living things and individual words.

The continuous wavelet transform has been used to decode the movements made by a subject [14] or to detect networks of neuronal activity related to movement [1]. Though these studies did not involve language, it has been proposed that language processing may be a distributed task that, for example, involves the motor cortex when the language being perceived is movement related [9]. The idea of a distributed system of semantic representation has stirred up controversy (see [15]). Still, we would like to leave open the possibility of involvement of motor cortex and visual cortex (and their associated rhythmic activity) when learning our semantic prediction functions f. A detailed examination of the usefulness of wavelet transforms for EEG data is given in [19], where wavelets were used to classify between three different cognitive tasks (multiplication, mental rotation of 3D object, silent letter composition) and 6 motor tasks (imagined movements). This study found that the information in several different wavelet types could be used to successfully differentiate between cognitive and imagined motor tasks.

2 MEG Transformations

Zero Shot Learning has provided us with a mechanism to map from a multivariate MEG time series X, to a vector of semantic features $\vec{s'}$, to arrive at a predicted word w. Now we define transformations g on the MEG time series X that may provide additional information to the learned functions f so that the mapping $f_i(g(X)) \dots f_m(g(X)) \rightarrow \{s'_1 \dots s'_m\}$ increases the chance that the word w that minimizes $d(\vec{s}_w, \vec{s'})$ is the correct word label for the MEG recording X. The space of possible functions g is infinite. In this section we select and define 5 functions from the infinite space of functions for further exploration.[4] We call the output of these functions *MEG transformation types.*

Magnitude of the Signal: We create the Magnitude MEG transformation by calculating a single time series which is the mean of all 20 MEG time series obtained from the 20 repeated presentations of each individual stimulus. If the magnitude of the signal is the best MEG transformation, then it is the magnitude of the magnetic field (or gradient of the field, in the case of gradiometers) that best encodes the semantic features of the words.

Average in a Window (Windowed Mean): The Windowed Mean MEG transformation is the average of all trials for a given word, further averaged within a 50 ms windows (with 25 ms overlap). If the Windowed Average outperforms the Magnitude of the Signal, it indicates that averaging reduces the noise more than it reduces the signal.

Frequency Power and Phase: To create Phase and Power transformations, we apply a short time Fourier transform (STFT) to single trials, and use the Fourier coefficients to calculate the power and phase in each frequency band (0.5 to 50 Hz). We average the phase and frequency across all single trials. The STFT windows have width 100 ms, with 50 ms overlap.

The rhythmic coordination of many neurons to form oscillations is a topic of great interest amongst MEG researchers (e.g. [16], [11], [6]). If the STFT Power transformation performs best, then the rhythmic activity of neurons firing at a particular rate is what encodes information. If the STFT Phase transformation performs best then it is not the strength of the oscillations in a frequency band, but their synchronization to the stimulus that is important.

Continuous Wavelet Transformation: The Continuous Wavelet Trasformation (CWT) is an adaptation of the Discrete Wavelet Transformation (DWT) that alters the size and position of the wavelet in small increments, which creates coefficients that are smoother in time and frequency spaces. This smoothness creates coefficients that, unlike the DWT, are more robust to noise and small shifts in time. Smooth coefficients are useful for inconsistent signals, like MEG recordings of higher-order cognition. For an in depth description of the DWT and CWT see [23]. Previous exploration by our group has found little difference between different mother wavelets (Morlet, Haar, Daubechies). Because of its simplicity, we chose to use the Haar wavelet, which is a simple step function.

The wavelet transformation has windows at multiple time scales, and so, unlike the STFT, it can handle signals with time-varying phase and amplitude frequency components (non-stationary signals). This makes it a particularly attractive candidate for analyzing MEG data. For this study, we use wavelet scales $[4 \dots 64]$, which corresponds to pseudo-frequencies $[3.1 \dots 49.8]$. We average the wavelet coefficients returned by the CWT across all of the single trials for a given word.

The wavelet transformation represents many different types of information. The CWT is an over-complete representation of the signal, so like the discrete case, one can fully recreate the

[4] For additional transformations, see `http://www.ml.cmu.edu/research/dap-papers/dap_fyshe.pdf`

original signal with a linear combination of the wavelet coefficients. The wavelet coefficients can also extract features related to the frequency and phase of the signal. If the Continuous Wavelet Transformation performs the best, then it may be a combination of phase, frequency power and the magnitude of the signal that contributes to the decoding of semantics. In addition, the robustness of the Continuous Wavelet Transform to noise may contribute to its performance.

All transformations: Finally, we can use all of the transformations described in the paragraphs above, append them into one large training set, and create one large regressor. This results in a training data vector with length on the order of 3.8 million. If the individual MEG transformation types contain complementary information about each of the semantic features, using all of the transformations together may produce better predictions. Regularized regression and cross-validation reduce the chance that we will overfit the data.

3 Prediction Framework

We turn now to the methods we use to evaluate the utility of each MEG transformation type. To learn the 218 independent functions f_i, we employ L_2 regularized regression, or Ridge Regression, which has several nice properties. Firstly, regularization automatically down-weights less useful features and avoids over-fitting. Secondly, L_2 regression has a closed form and can be solved without gradient descent methods. Thirdly, because L_2 regularization produces a linear predictor, we can employ Generalized Cross Validation (GCV) to choose our Lambda parameter [7] separately for each of the cross validation folds.

Our long term goal is to predict the word a person is reading from a large set of candidate words. To simulate this task, we use a larger list of 940 words for which we have collected the same 218 semantic features, but no MEG data. Given a predicted semantic feature vector, we can rank the 941 semantic feature vectors (940 new + 1 true) by their distance to the predicted semantic feature vector. We sort the distances and find the position of the true semantic feature vector. Rank accuracy is the percentage of the list of 941 words ranked lower than the correct word:

$$\text{rank accuracy} = (1 - \frac{r_i}{W}) \times 100$$

where r_i is the position of the true semantic feature vector in the list of sorted distances and $W = 941$ is the total length of the sorted list. Under this schema, higher scores are better. We use 30-fold cross validation, and test the rank accuracy for each held out word.

To train the regressors we use 750 ms of MEG signal beginning immediately after the onset of the stimulus. The generally agreed upon time at which semantic processing has finished is 750 ms [17]. For transformations created from MEG signals using a window, we used those windows with midpoints between 0 and 750 ms after stimulus onset. We standardize the semantic features so that each has mean 0 and standard deviation 1.

Testing for Significance Since we are measuring the performance of 5 MEG transformations across many subjects, we must correct our results for multiple comparisons. For this we use a combination of Fisher's Method and Bonferroni correction. Fisher's Method combines several p-values into one variable with a χ^2 distribution that we can convert into a p-value.

To test the statistical significance of a result, we must calculate the probability that a result was obtained when there is no connection between the MEG data and the semantic features (the null hypothesis). To estimate a distribution drawn from data satisfying the null hypothesis, we perform more than 100 permutation tests in which we shuffle the labels of each of the 1200 MEG single trials, and use that relabeled data as input to the MEG transformations. We then learn regressors with the same procedure outlined above, and use the rank accuracies obtained by training on this permuted data to empirically estimate the distribution of rank accuracies obtained under the null hypothesis.

Using the rank accuracy distribution under the null hypothesis, we can estimate the probability of a rank accuracy value obtained with the correct data (create a p-value). We do this for all rank accuracy values calculated from regressors trained on the non-permuted data. We choose the standard cutoff of $\alpha = 0.05$, which we Bonferroni correct to $\alpha = 0.05/5 = 0.01$ to correct for the 5 MEG signal transformations tested.

4 Results

Table 1. Median rank accuracy over 941 words for all 9 subjects and 5 transformation types, as well as all transformations appended. Higher scores are better. All 5 transformation types performed better than chance. A Wilcoxon test shows that the Wavelet transformation produces more accurate predictions than Windowed Mean, and that there is no significant difference in the performance of the Wavelet transformation and All Transformations.

MEG Transformation	S1	S2	S3	S4	S5	S6	S7	S8	S9	Mean
Magnitude	90.1	93.8	88.2	95.4	90.9	87.7	87.1	88.7	90.1	**90.2**
Windowed Mean	88.2	92.2	88.8	95.5	91.4	88.0	88.2	89.2	92.0	**90.4**
Phase	79.6	80.3	53.6	82.9	77.3	56.3	74.4	63.6	61.2	**69.9**
Power	58.0	82.6	59.5	87.7	84.7	61.7	80.8	68.5	66.5	**72.2**
Wavelets	95.3	96.0	90.1	96.5	91.6	90.7	88.3	87.8	95.8	**92.5**
All Transformations	95.0	95.9	90.4	96.7	91.7	89.8	88.8	87.6	95.6	92.4

Table 1 shows the median rank accuracy for each of the 9 subjects and 9 MEG transformation types. Fisher's Method shows that all 5 MEG transformation types produce rank accuracies that are statistically significantly better than chance across subjects ($p < 10^{-13}$), which easily passes our Bonferroni correction. We can test wether two transformation types have statistically different performance by recording the distance of each predicted vector to the true semantic feature vector. If there is no consistent advantage in using one feature type over the other, we would expect the distributions of the distances to be very similar. One can test wether the two distances could have been drawn from distributions with the same median using the Wilcoxon rank sum test. Using the Wilcoxon test we deduced that there was no significant difference between the Windowed mean and Magnitude transformations ($p = 0.87$). There is also no significant difference between Phase and Power transformations ($p = 0.58$). However, there is a significant difference between the Windowed mean and the Wavelet transformations ($p = 2.2 * 10^{-4}$).

Interestingly, a regressor trained on all appended transformations has mean performance slightly worse than the regressor trained on just the wavelet transformation. For all 9 subjects, the rank accuracy of the "All Transformations" predictions are within a percent of the "Wavelets" predictions, so it is not surprising that the Wilcoxon test fails to reject the null hypothesis that either transformation type is significantly better ($p = 0.6966$). Under some circumstances, a regressor trained on many features may not outperform regressors trained on feature subsets. If the information available in the additional features is not useful or complimentary, the addition of such features will not improve performance. We have confirmed with simulated data that the performance of an L_2 regularized regressor degrades quickly in the face of many irrelevant features or many noisy replicates of the same feature (corroborated also in [12]). In this case, using the Wavelet transformation alone is as good as using the Wavelet transformations in conjunction with many other sub-optimal transformations. Using the wavelet features alone reduces the length of the feature vectors by about 1 million elements.

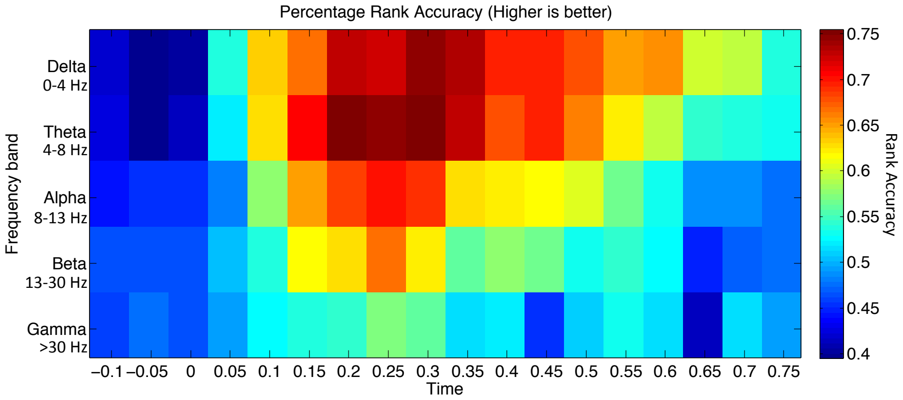

Fig. 1. The rank accuracy for different scales and time segments of the coefficients produced by the Continuous Wavelet Transform (CWT) for all sensors. Delta corresponds to frequencies 0-4Hz, theta 4-8Hz, alpha 8-13Hz, beta 13-30Hz, gamma >30 Hz (truncated at 50 Hz due to CWT parameters). The plot shows that the most useful coefficients are focused between 200 and 300 ms after stimulus onset, and between the delta to alpha frequency bands.

Analysis of Wavelet Features The Continuous Haar Wavelet MEG transformation is the top performing amongst the transformations we have evaluated here. But there are two dimensions to this transformation: frequency and time. Which dimension carries the most decoding power? Figure 1 shows the rank accuracy as a function of time and frequency for the Wavelet transformations using the signal from all sensors. From this plot we can see that the majority of the decoding power is focused between 200 and 300 ms after stimulus onset, and in the delta, theta and alpha frequency bands. This is in line with the timing of semantic decodability seen in previous studies [21]. The high performance in the low frequency domain is also consistent with previous results that used only the average magnitude of the MEG signal for classification, in which the dominant features would be the low frequency components.

5 Conclusion

We have explored several MEG signal transformations for handling the sometimes inconsistent MEG signals that result from higher-order cognitive processing. We built a model that predicts the concrete noun a person is reading based on MEG recordings, explored several MEG signal transformations, and detailed the effect of transformations on the model's predictive performance. We have shown that the Continuous Haar Wavelet Transform is the best MEG transformation of those considered here. Previous work has shown the usefulness of the Wavelet transform for MEG and EEG data with a focus on visual and motor tasks [22, 1, 14, 19]. While the task explored here focused on language and semantics, the conclusions are consistent: the Continuous Wavelet Transform provides a markedly better representation of the underlying signal.

The Continuous Wavelet Transform has several advantages over the other transformations explored here: it represents the frequency and phase (with respect to the stimulus onset) of the MEG signal, it can handle signals with time-varying frequency components, and it is robust to noise and to shifts in time. The STFT transformations also capture information related to the frequency and phase, but performed poorly in this study. Perhaps the frequency components of the MEG signal are not stable in the window size selected for the STFT. Shifts in frequency and phase that occur at too fine a scale will be lost in the STFT, which reduces comprehensiveness when compared to the Continuous Wavelet Transform. One could tune the window size of the STFT, but the same information is available without tuning when one chooses to use the Wavelet Transform.

In the future we would like to extend this work to incorporate multivariate transformation types (those that combine sensors together). In particular, measures of functional connectivity may hold great promise, as indicated in [5]. In addition, we would like to use the wavelet transform to perform single trial analysis, where only one trial per word is available. We hope that the robustness of the wavelet transform to noise will prove advantageous in this challenging low SNR scenario. We also plan to extend this work to phrases and full sentences. We expect that the increases in accuracy found here will extend to multi-word paradigms as well.

References

1. Danielle S Bassett, Andreas Meyer-Lindenberg, Sophie Achard, Thomas Duke, and Edward Bullmore. Adaptive reconfiguration of fractal small-world human brain functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51):19518–23, December 2006.
2. Alexander M Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S Cash. Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage*, 54(4):3028–39, February 2011.
3. D Cheyne and H Weinberg. Neuromagnetic fields accompanying unilateral finger movements: pre-movement and movement-evoked fields. *Experimental brain research*, 78(3):604–12, January 1989.
4. A K Engel, P Fries, and W Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature reviews. Neuroscience*, 2(10):704–16, October 2001.
5. Alona Fyshe, Emily Fox, David Dunson, and Tom Mitchell. Hierarchical Latent Dictionaries for Models of Brain Activation. In *The fifteenth international conference on Artificial Intelligence and Statistics*, volume XX, pages 409–421, 2012.
6. Avniel Singh Ghuman, Jonathan R McDaniel, and Alex Martin. A wavelet-based method for measuring the oscillatory dynamics of resting-state functional connectivity in MEG. *NeuroImage*, 56(1):69–77, January 2011.
7. Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.
8. Ole Jensen, Jochen Kaiser, and Jean-Philippe Lachaux. Human gamma-frequency oscillations associated with attention and memory. *Trends in neurosciences*, 30(7):317–24, July 2007.
9. Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, 5(1):e8622, January 2010.
10. Ryusuke Kakigi, Minoru Hoshiyama, Motoko Shimojo, Daisuke Naka, Hiroshi Yamasaki, Shoko Watanabe, Jing Xiang, Kazuaki Maeda, Khanh Lam, Kazuya Itomi, and Akinori Nakamura. The somatosensory evoked magnetic fields. *Progress in Neurobiology*, 61(5):495–523, August 2000.
11. Huan Luo and David Poeppel. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–10, June 2007.

12. Andrew Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine Learning (ICML)*, 2004.
13. Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M Mitchell. Zero-Shot Learning with Semantic Output Codes. *Advances in Neural Information Processing Systems*, 22:1410–1418, 2009.
14. G Pfurtscheller and F H Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 110(11):1842–57, November 1999.
15. David C Plaut and James L McClelland. Locating object knowledge in the brain: comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological review*, 117(1):284–8, January 2010.
16. F Pulvermüller, N Birbaumer, W Lutzenberger, and B Mohr. High-frequency brain activity: its possible role in attention, perception and language processing. *Progress in neurobiology*, 52(5):427–45, August 1997.
17. Riitta Salmelin. Clinical neurophysiology of language: the MEG approach. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 118(2):237–54, February 2007.
18. Riitta Salmelin and Jan Kujala. Neural representation of language: activation versus long-range connectivity. *Trends in cognitive sciences*, 10(11):519–25, November 2006.
19. Jesse Sherwood and Reza Derakhshani. On Classifiability of Wavelet Features for EEG-Based Brain-computer Interfaces. In *International Joint conference on Neural Networks*, pages 1–8, 2009.
20. PN Steinmetz, A Roy, PJ Fitzgerald, S. S. Hsiao, K. O. Johnson, and E. Niebur. Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404:187–190, 2000.
21. Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking Neural Coding of Perceptual and Semantic Features of Concrete Nouns. *NeuroImage*, 62(1):463–451, May 2012.
22. C Tallon-Baudry, O Bertrand, C Delpuech, and J Pernier. Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 16(13):4240–9, July 1996.
23. Brani Vidakovic. *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics. Wiley, 1999.