33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

An Unsupervised Method for Identifying Regions that Initiate Seizures on Intracranial EEG

Drausin Wulsin*, *Student Member IEEE*, and Brian Litt, *Senior Member IEEE*

Abstract— **Epilepsy patients who do not respond to pharmacological treatments currently have only brain surgery as a major alternative therapy. Identifying which brain areas to remove is thus of critical importance for physicians and the patient. Currently, this process is almost entirely manual, can vary greatly between clinical experts and centers, and depends only on qualitative EEG features, all of which may help explain the only modest success of extratemperal lobe epilepsy surgery. In this study, we explore an unsupervised, quantitative method for identifying seizure onset regions. A Gaussian mixture model (GMM) was used to cluster 500 ms epochs of intracranial electroencephalogram (EEG) prior to (preictal) and during (ictal) seizures in week-long continuous recordings from three patients during evalulation for epilepsy surgery. The GMM learning paradigm determines the optimal number of clusters for each patient. For the two patients whose epochs sorted into two clusters, we found that one cluster was predominantly composed of seizure epochs, and a subset of the channels made brief "forays" into that cluser in the time leading up to seizure onset. This observation is in keeping with the clinical hypothesis that certain brain areas may be the initiators of seizure activity, and we find that the channels independently labeled by physicians as seizure onset zones (SOZs) are statistically overrepesented in the seizure-defined cluster. Nevertheless, we also find that a subset of channels not labeled as SOZs has similar properties as those labeled SOZs. In this study we have tried to avoid many of the assumptions commonly made about what features and events are indicative of epileptogenic activity and believe that such analysis can help avoid many of the pitfalls of manual, non-objective human SOZ marking.**

I. INTRODUCTION

Those who suffer from frequent epileptic seizures have two main treatment options: anti-epileptic drugs, and if those are ineffective, surgery, where part of the brain thought to be starting the seizures is removed. For those patients for whom surgery is the only remaining treatment option, the primary question for doctors is what areas of the brain to remove. Thus, determining the *seizure onset zone* (SOZ) is of critical importance. In current clinical practice, board-certified epileptologists look at the intracranial electroencephalogram (EEG) from implanted electrodes on the surface of the cortex or penetrating into deeper brain structures like the hippocampus or amygdala. The physicians manually examine the EEG around the patient's seizures and try to identify specific electrodes (also known as channels) where the seizures or seizure-like activity seems to start under the working hypothesis that removing them may reduce or eliminate a patient's seizures.

In current clinical practice, this process is almost entirely manual, with physicians paging through raw EEG traces and noting distinguishing and unusual characteristics like spikes and higher-frequency activity in a subset of the channels. In surgery, the physicians try to remove these channels in hopes that it will reduce the number and severity of seizures in the patients, and for some patients (approximately 35% in non-lesional extratemporal epilepsy [1]) this sugery leaves them seizure-free.

Why do the other half of patients still have seizures after surgery? In addition to the diversity and complexity of the disease and the crudeness of the surgical procedure, the sheer mass of EEG data that the epileptologists must wade through to make their seizure onset zone predictions is a potential source of error. Furthermore, the lexicon of qualitative EEG features used by epileptologists may be insufficient to fully identify SOZs. While humans tend to be very good at noticing patterns in data, they are usually not as good at meticulously combing through large amounts of it. Thankfully, automated algorithms tend to do well on such tasks, and we believe they deserve a prominant role in understanding how different channels (and thus areas of brain) are involved in the dynamics leading up to a seizure.

In this paper, we present an unsupervised clustering algorithm that models short (500 ms) epochs across channels leading up to and during patients's seizures. We use a measure called the gap statistic [2] to assess how many clusters are appropriate, including only one cluster. In two of the three patients studied, we found that pre-seizure and during-seizure epochs in individual channels separated into two clusters. In these patients, we found that the cluster prodiminantly defined by seizure epochs also contains a larger-than-expected fraction of epochs in channels independently labeled by epileptologists as within the seizure onset zone. This work is an attempt to characterize the behavior of channels involved in pre-seizure activity without relying as much on potenially fallible human judgements or assumptions.

II. METHODS

We used continuous intracranial EEG from three neocortical epilepsy patients undergoing pre-surgical evaluation in the epilepsy monitoring unit at the Mayo Clinic, Rochester, MN. Data was sampled at 500 Hz and EEG clips from all channels were extracted from five minutes before to one

This work was supported in part by the National Institutes of Health Grants (NINDS RO1-NS041811, NINDS R01 NS 48598) and the Dr. Michel and Mrs. Anna Mirowski Discovery Fund for Epilepsy Research.

D. Wulsin (corresponding author): Department of Bioengineering University of Pennsylvania, Philadelphia, PA wulsin@seas.upenn.edu

B. Litt: Departments of Neurology and Bioengineering University of Pennsylvania, Philadelphia, PA littb@mail.med.upenn.edu

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

TABLE I

NUMBER OF EPOCHS ANALYZED FOR EACH PATIENT.

	No. epochs
patient 1	966,336
patient 2	253,088
patient 3	431,400

minute after the start of each seizure for each patient, who had seven, five, and four seizures, respectively. We used a 500 ms sliding window (with 250 ms overlap) in each channel of each seizure of each patient to extract short-term "epochs." Table 1 shows the number of epochs extracted and analyzed for each patient across all channels. Under the hypothesis that seizure onset channels can sometimes demonstrate seizure-like activity preictally, we extracted five features meant to separate seizure from non-seizure activity: line length [3], and log-power in the delta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-100 Hz) EEG frequency bands, with the assumption that seizures and seizure-like activity tends to have higher frequency content (in the beta and gamma bands) than "normal" activity. For each patient individually, we normalized the epochs to have zero mean and unit variance before applying Principal Components Analysis (PCA), reducing the five features down to two dimentions (referred to as "pc1" and "pc2") while retaining 89.8%, 90.5%, and 96.4% of the variance for the three patients, respectively. Fig. 1 shows scatter plots of a random sample of epochs in the two-dimensional PCA space for each patient. Each epoch was given a corresponding label of whether or not it came from a channel previously marked by a human as a seizure onset zone (SOZ) channel.

We fit a Gaussian mixture model using the expectation maximization (EM) algorithm [4]. In traditional mixture-modeling, one must predefine the number of models in the mixture, which can be a somewhat ad hoc procedure. We used a measure call the gap statistic described by Tibshirani *et al.* [2] to determine the optimal number of clusters, including just a single cluster. The data was clustered with the gap statistic in twenty independent trials for each patient to confirm that the number of clusters found by the gap statistic remained constant across the trials. Note, the seizure-onset channel labels were not used in any way during the unsupervised clustering, which only used the two PCA dimensions in representing each event.

For those patients whose epochs sorted into more than one cluster, we examined the population of epochs found in the cluster (which we call the "second" cluster) dominated by epochs during the seizure. Specifically, we compared the observed fraction of epochs that came from SOZ-labeled channels to their expected fraction using a 1000-trial permutation test over the human SOZ labels. Finally, to examine how similar each channel's pre-seizure epochs were to the aggregate in-seizure activity, we tabulated how many epochs from each channel occurred in cluster two in the minutes leading up to seizures.

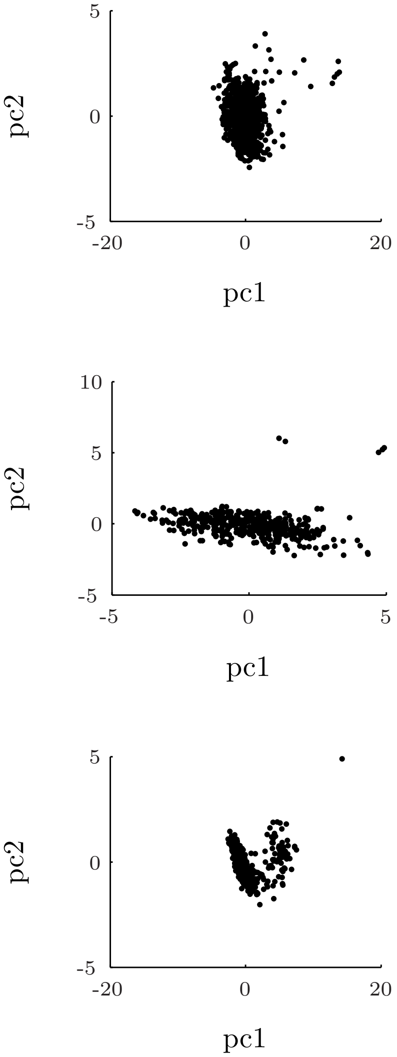

Fig. 1. Scatter plots of epochs drawn at random from patients 1 (top), 2 (middle), and 3 (bottom) shown in two-dimensional PCA-space.

III. RESULTS AND DISCUSSION

Interestingly, we found that the epochs leading up to and during seizure did not separate into distinct clusters for all of the patients. Two event clusters were found for two of the three patients (patients 1 and 3), and one cluster for patient 2. We see patient 2's single cluster as evidence that our model is not making undo assumptions about the data, however reasonable they may seem. Furthermore, it is evidence that the data can sometimes tell a different story than we humans expect, which here is that epochs in seizure are always statistically different from those just before a seizure. The single cluster does not mean that the epochs in seizures are indistinguishable from those before seizures, only that they do not comprise their own statistically significant sub-population.

In the two patients where two distinct sub-populations of

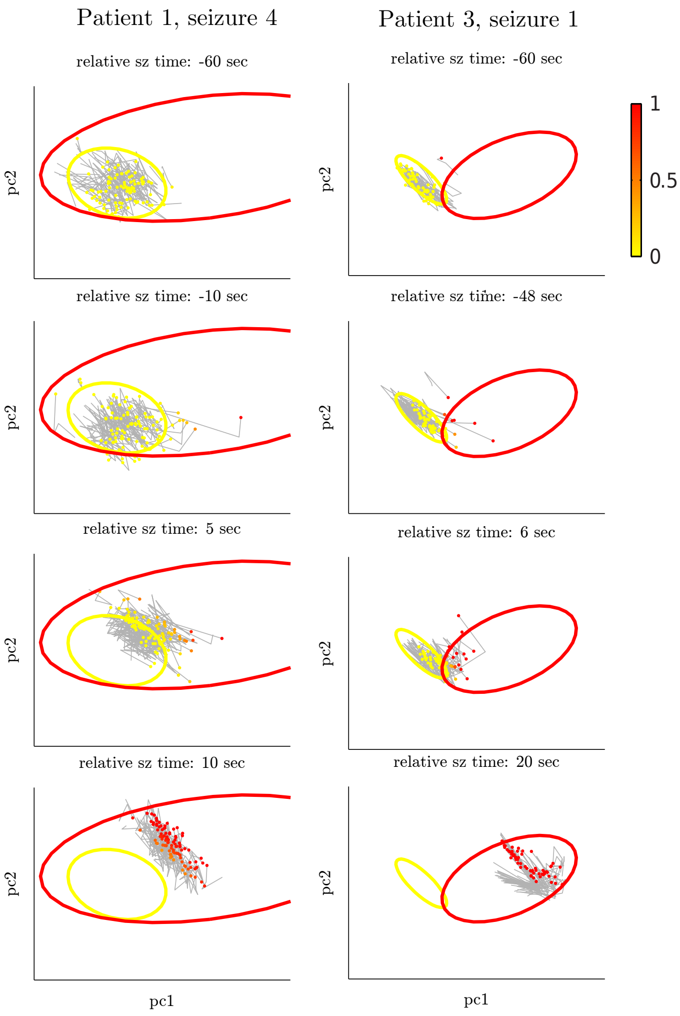

Fig. 2. Event dynamics leading up to and during a seizure for patient 1 (left) and patient 3 (right). Outlines of the majority cluster (yellow) and seizure-dominated (red) cluster delineating 90% of the probability mass of each Gauassian. Part of the red, seizure-dominated Gaussian in patient 1 is cut off by the plot on the right. Each dot represents a channel at a given point in time, and its gray tail shows its path from the previous three time points. The dots are colored by their posterior probability (0 to 1) of being in the seizure-dominated cluster (red). The rows show different time points leading up to and during the start of a seizure.

epochs were found, one of the clusters contained the large majority (95% to 97%) of the probability mass of the mixture model. The other cluster was defined predominantly by epochs during the seizure, as one might expect. We noticed, however, that leading up to the seizure, a handleful of channels would occasionally make "forays" into the second cluster. Fig. 2 shows this behavior for both patients. As the seizure begins, a subset of channels "lead the way" into the seizure-dominated cluster, where all of the channels end up by the middle of the seizure.

These observations about the dynamics of pre-seizure channel epochs are consistent with the clinical hypotheses that certain channels can be the initiators of epileptic activity that ultimately expands to include most or all of the other channels. Our hypothesis that some of these leader channels (heretofore with blinded labels) would be the same channels labeled by a human as part of the SOZ was confirmed by the permutation tests, shown in Fig 3. The observed fraction of SOZ-labeled channels was significantly higher than chance ($p < 0.001$), indicating than the SOZ-channels are overrepresented in cluster 2 for both patients. Given that the majority of epochs in cluster 2 are still from channels not labeled SOZs (85-87%) we do not claim that cluster 2 is the "SOZ-cluster," merely that it is predominantly defined by in-seizure epochs (as seen in the bottom row of Fig. 2) and that epochs from SOZ-labeled channels are overrepresented in this cluster.

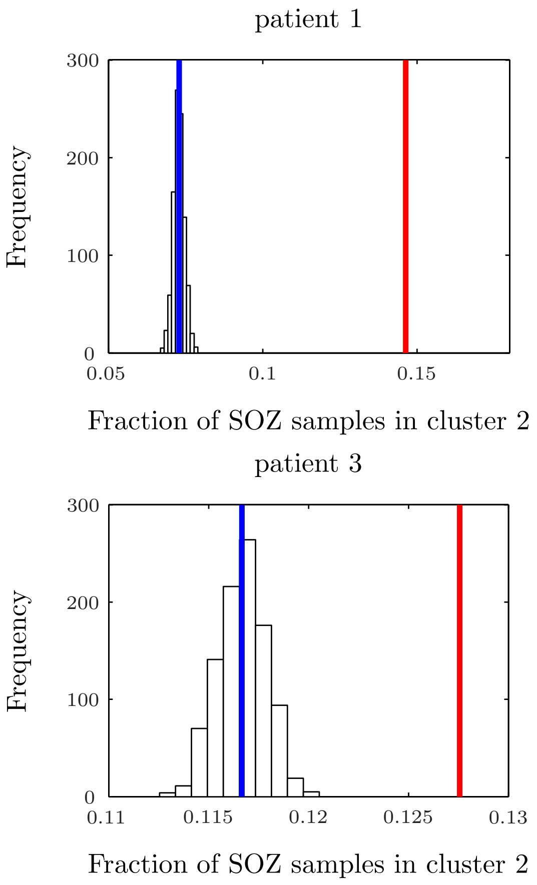

Fig. 3. The observed (red line) fraction of channels labels as SOZ in the seizure-dominated cluster 2 compared to the chance expectation (blue line) and its distribution (histogram) as determined by 1000 trials over the labels from a permutation test.

Our final analysis, seen in Fig. 4, shows the log-frequency that each channel spent in the seizure-dominated cluster 2 in the time before (but not during) the seizures. We notice that in both patients the human-labeled SOZ channels are not unique in how much time they spent in cluster 2. In the case of patient 3, the SOZ channels are not even the ones that spent the most time in cluster 2. While these non-SOZ-labeled channels in the aggregate are proportionally underrepresented in cluster 2 (as seen in Fig. 3 by the fact that the SOZ-labeled channels are overrepresented), some individuals spend an amount of time on par with those channels labeled as SOZs.

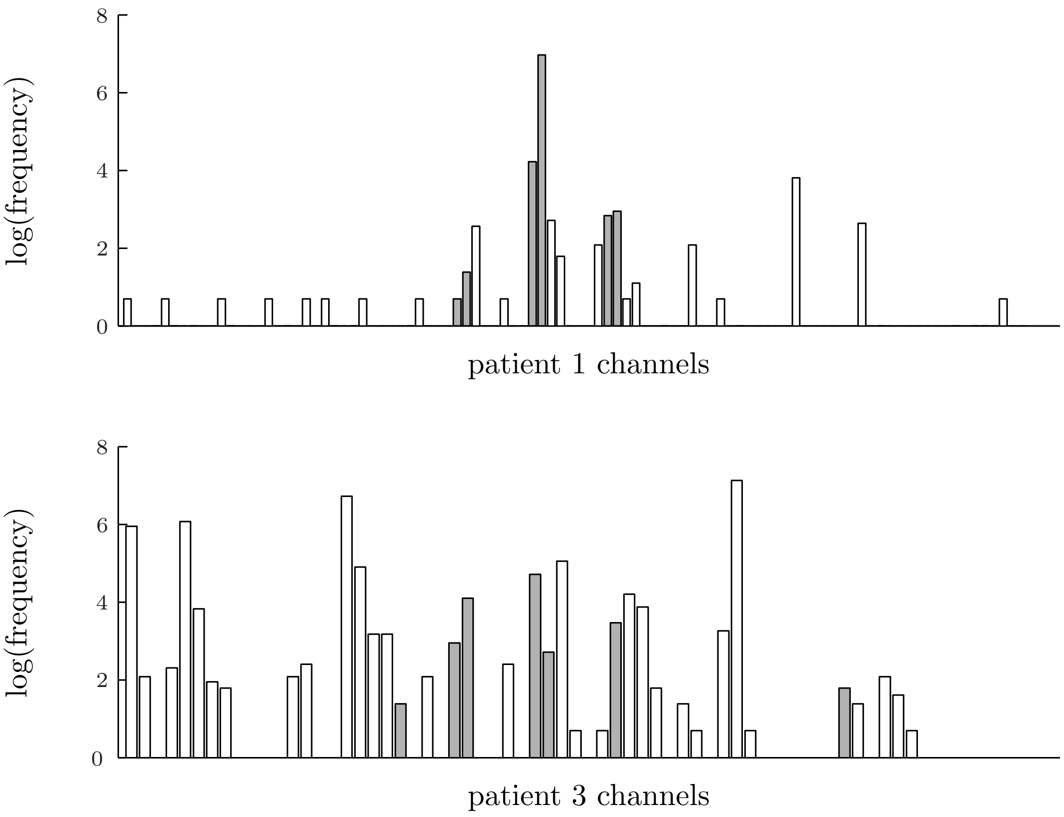

Fig. 4. The log frequency of epochs occuring in the seizure-dominated cluster 2 for each channel of patients 1 and 3. Grey bars denote the channels labeled as SOZs by a human marker, and white bars denote those not labeled as SOZs.

These analyses seem to indicate that a subset of channels that physicians do *not* label as SOZs can behave very similarly to those channels labeled as SOZs.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we have modeled the epochs leading up to and during seizures using unsupervised methods (i.e., no *a priori* human SOZ labelings were used in the model). We have tried to involve as few assumptions as possible in this model, for example assumptions about what types of epochs are correlated with seizure onset zones, assumptions that epochs in seizures are always statistically discinct from those before seizures, and assumptions that the channels labeled SOZs are the "gold standard" for what channels truly are the precursors and instigators of epileptic activity.

We found the following results: 1) not all patients have epochs in seizures that statistically distinct from pre-seizure epochs, 2) in those patients with distinct clusters, a few channels seem to make pre-seizure "forays" into the state-space defined by seizure-activity, in keeping with clinical hypotheses about epileptogenic brain areas, 3) these leading channels are statistically overrepesented by those labeled by humans as SOZs but 4) are not the only channels that spend a large amount of time in the seizure-dominated cluster.

These analyses and observations are still very exploratory and certainly merit future study. We plan to increase the number of patients as well as used channels subsequently resected in surgery combined with patient outcomes to assess whether the human-labeled SOZ channels and/or the channels that spend more time in cluster 2 have any predictive power for estimating patient outcome after resection. We believe that only a quantitative, objective analysis of channel features and epochs will have the discriminative power in predicting patient outcome after surgery.

REFERENCES

[1] N. Wetjen, R. Marsh, F. B. Meyer, G. D. Cascino, E. So, J. Britton, S. M. Stead, and G. A. Worrell, "Intracranial electroencephalography seizure onset patterns and surgical outcomes in nonlesional extratemporal epilepsy," *Journal of Neurosurgery*, vol. 110, no. June, pp. 1147–1152, 2009.

[2] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, May 2001.

[3] R. Esteller, J. Echauz, T. Tcheng, B. Litt, and B. Pless, "Line length: an efficient feature for seizure onset detection," *Proceedings of the 23rd EMBS Conference*, no. 3, pp. 1707–1710, 2001.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics, M. I. Jordan, J. Kleinberg, and B. Scholkopf, Eds. Springer, 2006, vol. 16, no. 4.