Tracking Electroencephalographic Changes Using Distributions of Linear Models: Application to Propofol-Based Depth of Anesthesia Monitoring

Levin Kuhlmann*, Jonathan H. Manton, *Fellow, IEEE*, Bjorn Heyse, Hugo E. M. Vereecke, Tarmo Lipping, *Senior Member, IEEE*, Michel M. R. F. Struys, and David T. J. Liley

***Abstract*— *Objective:* Tracking brain states with electrophysiological measurements often relies on short-term averages of extracted features and this may not adequately capture the variability of brain dynamics. The objective is to assess the hypotheses that this can be overcome by tracking distributions of linear models using anesthesia data, and that anesthetic brain state tracking performance of linear models is comparable to that of a high performing depth of anesthesia monitoring feature. *Methods:* Individuals' brain states are classified by comparing the distribution of linear (auto-regressive moving average—ARMA) model parameters estimated from electroencephalographic (EEG) data obtained with a sliding window to distributions of linear model parameters for each brain state. The method is applied to frontal EEG data from 15 subjects undergoing propofol anesthesia and classified by the observers assessment of alertness/sedation (OAA/S) scale. Classification of the OAA/S score was performed using distributions of either ARMA parameters or the benchmark feature, Higuchi fractal dimension. *Results:* The highest average testing sensitivity of 59% (chance sensitivity: 17%) was found for ARMA $(2, 1)$ models and Higuchi fractal dimension achieved 52%, however, no statistical difference was observed. For the same ARMA case, there was no statistical difference if medians are used instead of distributions (sensitivity: 56%). *Conclusion:* The model-based distribution approach is not necessarily more effective than a median/short-term average approach, however, it performs well compared with a distribution approach based on a high performing anesthesia monitoring measure. *Significance:* These techniques hold potential for anesthesia monitoring and may be generally applicable for tracking brain states.**

***Index Terms*—Anesthesia, Autoregressive Moving Average (Arma) Modeling, Brain Dynamics, Electroencephalography, Model-Based Estimation.**

I. Introduction

MODEL-based estimation methods offer the potential to develop more physiologically motivated approaches to track brain states and infer underlying physiological changes with limited electrophysiological measurements [1]–[3]. Recently, many model-based approaches were developed to predict epileptic seizures in a patient-specific manner [4] and to track anesthetic brain states [5]. The models used are often nonlinear models of complex brain dynamics or linearized approximations of the full nonlinear model. Generally with most estimation frameworks, such as stochastic filtering, the conditional expected values of the model parameters are often generated [6]. Similarly, feature-based approaches to track brain states often take a short-term average feature value over each windowed segment of data, or take the mean of pooled feature values across subjects [7]–[9]. These approaches may not capture the variability of brain dynamics across both time and people, and therefore may limit the ability of a brain-state classifier trained with the model parameter estimates or electrophysiological feature values from one group of subjects to be used to classify brain states in another group of subjects.

For example, this is particularly true in the field of depth of anesthesia monitoring during surgery. The goal of this field is to minimize the possibility of the patient being aware/awake or experiencing pain during surgery [10]–[12], as well as the likelihood of postoperative sequalae that include postoperative nausea and vomiting [13], and postoperative cognitive deficits that are particularly significant in the elderly [14]. Many of the published studies to date track depth of anesthesia using features computed from frontal electroencephalographic (EEG) recordings [7]–[12], [15]–[20]. Analysis of the feature values is usually performed by taking the average of the feature values across all subjects and one looks for monotonic

Manuscript received October 11, 2015; revised April 5, 2016; accepted April 24, 2016. Date of publication June 14, 2016; date of current version March 17, 2017. This work was supported in part by the ARC Linkage under Grant LP120200773 and by the Cortical Dynamics Pvt. Ltd., a depth of anesthesia monitoring device company. *Asterisk indicates corresponding author.*

*L. Kuhlmann is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville 3010, Australia (e-mail: levink@unimelb.edu.au).

J. Manton is with the Department of Electrical and Electronic Engineering, University of Melbourne.

B. Heyse and M. M. R. F. Struys are with the Department of Anesthesia, Ghent University Hospital.

H. E. M. Vereecke are with the Department of Anesthesiology, University of Groningen, University Medical Center Groningen.

T. Lipping is with the Department of Information Technology, Tampere University of Technology.

D. T. J. Liley is with the Brain and Psychological Sciences Research Centre, Swinburne University of Technology.

Digital Object Identifier 10.1109/TBME.2016.2562261

relationships between the mean feature value and anaesthetic concentration or behavioral responsiveness (an indirect measure of awareness, i.e., an awake state). This pooling across all subjects makes it difficult to predict the out-of-sample performance of depth of anesthesia monitoring methods when the methods are to be applied to new patients not included in earlier studies. This is because new patients may have a certain sensitivity to anesthesia that generates slightly different mean feature values for a specific anesthetic brain state when compared to the mean feature values for the group of subjects on which the method was designed. This difference may be enough to lead to inaccurate anesthetic state classification. If instead one considers distributions of feature values estimated from brain dynamics, there may be greater likelihood that there will be overlap between the distributions of feature values of the new patients and the group on which the method was designed, and thus better classification performance.

Here a method is presented which classifies the brain states of an individual using a method that compares the distribution of linear (auto-regressive moving average—ARMA) models [21] estimated from frontal EEG data with distributions of linear models for each brain state. The method is specifically applied for tracking the depth of anesthesia and the method is evaluated using out-of-sample testing of classification performance. A similar depth of anesthesia monitoring method compares the difference in distributions of gamma band (32–64 Hz) discrete wavelet transform coefficients within 1-s periods to a distribution of the same coefficients obtained from participants in the awake state and a distribution of the same coefficients obtained from participants in the fully anaesthetized state to derive a single measure of depth of anesthesia taking values between 0 and 100 [22]. Here, a variation is presented that instead uses distributions of linear model parameters and distributions for six states of awareness/responsiveness as opposed to two. In addition, rather than derive a single measure of depth of anesthesia, classification of a behavioral responsiveness measure (observers assessment of alertness/sedation—OAA/S scale) [23] used by anesthesiologists is performed.

The approach herein is motivated in part by developments in the computational modeling of the EEG using neural field models that have been used to describe data during resting and anaesthetized brain states [5], [24]. Our hypothesis is that the cortical region underlying the frontal EEG recording electrode can be modeled by a nonlinear dynamical system. This nonlinear model can be approximated by the concatenation over time of linearized approximations of the model which are estimated using ARMA modeling and an appropriately defined sliding window [8], [19], [20], [25]. In this paper, the focus is only on the linear ARMA model estimates for tracking depth of anesthesia. This linear modeling approach is supported by the significant experimental evidence that EEG recorded in the presence and absence of anesthesia can be modeled as a random linear process [26], [27]. Here, we seek not only to assess the utility of a distribution-based approach, but also to assess how linear modeling performs compared to a high performing depth of anaesthesia monitoring method evaluated within a distribution-based framework.

The paper is outlined as follows. Section II describes the methods used in the distribution-based tracking approach and how out-of-sample testing performance was evaluated when the method is applied to depth of anesthesia monitoring data. Section III describes the results of the analysis, Section IV discusses the key issues, and Section V concludes the paper.

II. Methods

We have developed a distribution-based approach to classify, or track, the brain states of an individual by comparing the distribution of recently calculated ARMA model parameter estimates, or other EEG features, computed from single-channel frontal EEG data to distributions of the same variables for each brain state that have been computed from data from a set of "training" subjects. A comparison of ARMA model parameter estimation methods using the Broersen technique [21] and Kalman filtering [28] is made to evaluate the effect of different estimation methods on the performance of our distribution-based approach. The Broersen technique is the primary estimation technique underlying the brain anaesthesia response monitor (Cortical Dynamics, Australia) [8], [19], [20], [25] and Kalman filtering represents a time-domain alternative for estimating the parameters. In addition, given that the application domain is depth of anesthesia monitoring, the performance of the distribution-based approach using ARMA model parameters is compared to the benchmark performance of the distribution-based approach using a depth of anesthesia monitoring measure, Higuchi fractal dimension (HFD) [7], that can be regarded as one of the best performing hypnotic measures evaluated to date and has been evaluated over the same dataset as considered here [7].

The methods underpinning our distribution-based classification approach are described in two stages: 1) the methods underlying ARMA modeling, ARMA model estimation, and the HFD; and 2) the depth of anesthesia monitoring EEG data and the distribution-based tracking/classification approach.

A. ARMA Models

ARMA time series models allow for an accurate description of single-channel EEG [21]. The basic form of an ARMA model is as follows:

$$z_t = -\sum_{j=1}^{p} a_t^{(j)} z_{t-j} + \sum_{k=1}^{q} b_t^{(k)} e_{t-k} + e_t \tag{1}$$

where z_t is the observed signal (in this case the frontal EEG), and $a_t^{(j)}$ and $b_t^{(k)}$ are the (time-varying) autoregressive (AR) and moving-average (MA) parameters, respectively, at time t. The constants p and q are the corresponding orders of the AR and MA parts, respectively, and e_t is the observation error or innovation process. The innovation process is assumed to be a Gaussian white noise process with zero mean and variance $\sigma_{e_t}^2$. Based on prior physiological modeling, model orders of $(p = 6, q = 3)$ and $(8, 5)$ were considered. The $(6, 3)$ and $(8, 5)$ model orders correspond to the model orders of transfer functions of linearized versions of the six-dimensional state and eight-dimensional state Liley models of "resting" and

"anaesthetized" EEG, respectively [25], [29]. Note that this is true for all linearizations about any physiologically plausible stable fixed point of the six-dimensional state and eight-dimensional state Liley models. Athough variations of the Liley model have not been directly applied to the modeling of propofol anesthesia, the use of an ARMA $(8, 5)$ model approach to derive a cortical state and cortical input features, and inspired by the Liley model, has been shown to be able to reliably track the propofol-induced anesthetic state [8], [19], [20]. In order to relate our methods to physiology, selection of ARMA model orders that correspond to the model orders of transfer functions of linearized versions of neural models is preferred over the use of model order selection approaches based on information theoretic criteria [30]. These theoretically derived model orders accord with optimal AR ($p = 3 - 14$) and MA ($q = 2 - 5$) orders obtained from resting awake eyes-closed EEG [31]. A $(2, 1)$ model order was also considered, given that lower order models reduce the possibility of overfitting and, therefore, can potentially give better generalization and classification performance.

B. ARMA Estimation With the Broersen Technique

For a given epoch of single-channel data, the method of Broersen was used to estimate an invertible and stationary ARMA model using a variant of Durbin's method with optimal intermediate AR order on zero meaned data [21]. The Broersen method of ARMA model/parameter estimation is well established and implemented directly in the ARMASA MATLAB Toolbox [21]. Subsequent estimates of the innovation variance $\sigma^2_{e_t}$ are calculated as the standard deviation of the zero meaned signal epoch divided by the square root of the power gain of the derived filter/ARMA model. With regard to (1), the AR and MA parameters are considered constant over a finite analysis window (i.e., $a_t^{(j)} \equiv a^{(j)}$ and $b_t^{(k)} \equiv b^{(k)}$).

C. ARMA Estimation With Kalman Filtering

The Kalman filter can be used to estimate ARMA (p, q) models for every sample of the data assuming a drift in the ARMA parameters. The algorithm we employ is based on the suboptimal Kalman filter presented by Tarvainen *et al.* [28], modified to include fading memory [6] and the time-varying estimation of the variance of the innovation process $\sigma^2_{e_t}$. In particular, the measurement noise estimated by the filter is treated as an approximation of the innovation process e_t. A time-varying estimate of the measurement noise covariance, and thus an estimate of the variance of the innovation process $\sigma^2_{e_t}$ is obtained by computing the variance of the estimated measurement noise over the previous two seconds relative to the current data sample. For all the analyses considered here, the state noise/parameter drift covariance matrix as defined by [28] was set to $0.000003I$, where I is the identity matrix. This is comparable to values used in prior studies [28]. Fading memory was used to provide Kalman filter estimates with similar time-varying standard deviations to the Broersen technique and estimation of the innovation standard deviation was employed since this reflects the power of the input to the model. As defined in [6] the fading memory of the filter is controlled by the parameter α. The standard Kalman filter is given by $\alpha = 1$, and $\alpha > 1$ leads to fading memory.

This modified Kalman filter was tested with various simulations of time-varying ARMA models and innovation standard deviations. Tracking performance was comparable to estimation with the Broersen technique (results not shown).

D. Higuchi Fractal Dimension—HFD

HFD is a measure derived from nonlinear dynamics [7]. It can be calculated in the time domain and, therefore, has very low computational complexity. Essentially, HFD estimates the fractal dimension of a time series by measuring the scaling of the length of the time series, when viewed geometrically as a curve, as it is successively subsampled. As described above, HFD performs well as a frontal-EEG-based depth of anesthesia monitoring feature. In particular, best performance has been empirically established for the 6–47-Hz frequency band [7]. HFD was calculated for finite length time series using the method as described by Ferenets *et al.* [7].

E. Depth of Anesthesia Monitoring Data

The data analyzed in this study have been utilized in previous works on depth of anesthesia monitoring [7], [8]. Detailed information regarding patient cohort and anesthetic protocols relevant to this dataset can be found in [7]. The key details are described here. Institutional ethics committee approval was obtained from the Ghent University Hospital (Gent, Belgium) in accord with the Declaration of Helsinki. Informed consent was obtained from 15 patients with American Society of Anesthesiologists physical status I, aged 18 to 60 years old, scheduled to undergo ambulatory surgery. Exclusion criteria included neurological disorder, recent use of psychoactive medication, including alcohol, and weight less than 70% or more than 130% of ideal body weight (as determined using the table of Desirable Weights, Metropolitan Life Insurance, 1983). Frontal EEG data recorded from the 15 subjects undergoing propofol anesthesia from the awake to the anaesthetized state were obtained using the M-Entropy module of the S/5 Anesthesia Monitor (GE Healthcare Finland Oy). The EEG data were passed through a 0.5–118-Hz bandpass filter, sampled at 400 Hz and written to disk. The standard entropy sensor of the S/5 monitor was used with a slightly modified positioning: the two recording electrodes of the sensor were located bilaterally on the forehead approximately 5 cm above the eyebrows and 4 cm from the midline in either direction. The ground electrode was located between the two recording electrodes. This alternative montage was chosen to minimize EMG activity contributions to the calculation of the state entropy and response entropy measures of the S/5 monitor.

Initially, a propofol effect site concentration of 0.75 μg/ml was targeted, increased every 4 min by 0.25–0.30 μg/ml until loss of response to all relevant clinical measures of anesthetic depth was observed. EEG data were only collected during this induction phase. Changes in behavioral responsiveness, including loss of responsiveness, were assessed using the modified OAA/S score, a subjective clinical measure of arousal,

TABLE I
RESPONSIVENESS SCORES OF THE MODIFIED OAA/S SCALE

Score	Responsiveness
5	Responds readily to name spoken in normal tone.
4	Lethargic response to name spoken in normal tone.
3	Responds only after name is called loudly and/or repeatedly.
2	Responds only after mild prodding or shaking.
1	Responds only after painful trapezius squeeze.
0	No response after painful trapezius squeeze.

alertness, and sedation [23]. This scale ranges from 5 for the fully awake state to 0 for complete unresponsiveness to a painful stimulus (see Table I). OAA/S measurements were noted in the 10 s before each propofol concentration step.

The administration of propofol involved using a computer-assisted continuous infusion device to achieve a target effect site concentration (RUGLOOP II; Demed, Temse, Belgium) using a three-compartment model enlarged with an effect site compartment. Infusion of propofol was administered using a Fresenius Modular DPS Infusion Pump connected to a Fresenius Base A (Fresenius Vial Infusion Systems, Bresin, France). RUGLOOP II controlled the pump at infusion rates between 0 and 1200 ml/h. Propofol was infused through a large left forearm vein. No subject received preanesthetic medication and no alternative drugs were given. Approximately, 200-ml crystalloid fluid was given to each patient during the study period and no fluid load was given before induction. Patients maintained spontaneous ventilation through a facemask delivering 6 l/min O_2 during the period of EEG data collection until an OAA/S of 0 was achieved.

F. Data Preprocessing and Artifact Removal

Preprocessing and artifact removal follows in part a previous approach of the authors that involves the estimation of cortical state and cortical input using ARMA models in order to track the depth of anesthesia and analgesia, respectively [8]. In contrast to previous descriptions [8], fixed-order ARMA models were calculated on contiguous 1-s nonoverlapping data epochs, rather than 2 s 50% overlapping segments, because of our use of a Kalman adaptive filter. The 1 s epochs were assessed for artifact as outlined below and artifact-containing epochs were ignored, otherwise epochs were accepted.

Because the Kalman filter is iterative it needs the data to be approximately continuous. The word "approximate" is used here because if a 1-s window is removed or ignored due to artifact then for the next accepted 1-s window the Kalman filter was initialized using values from the end of the last accepted window. Doing so avoids large artifact-induced fluctuations of the parameter estimates that can affect the numerical stablity of the algorithm and classification sensitivity. High correlation between ARMA coefficients and the OAA/S score is expected to give better separability of the ARMA coeffcient probability distributions corresponding to the different OAA/S classes and high classification sensitivity [32]. Therefore, to assess the benefits of applying artifact removal during Kalman-based estimation the correlation between the estimated ARMA coefficient time series and the nearest neighbor interpolated OAA/S score time series was analyzed using the Pearson correlation coeffcient ρ. This was done for the cases with and without artifact removal of the data when applying Kalman estimation.

Fixed-order ARMA model parameters were also calculated using the Broersen method for purposes of direct comparison. Broersen method estimates obtained from 1-s windows and no overlap were found to be statistically indistinguishable from estimates obtained from 2-s windows with 50% overlap (results not shown). The Broersen method operates on a window-by-window basis and, therefore, estimation for the current window is not affected by estimates in the previous window. Therefore, the correlation analysis with and without artifact removal was not applied to the Broersen method.

It is worth noting that here the Broersen method provides one estimate of the parameter values for each 1-s window, whereas the Kalman method provides a number of estimates that is equivalent to the sampling rate. Thus, the Kalman method provides more samples per 1-s window to contribute to the construction of distributions of parameter estimates than the Broersen method.

For ARMA modeling, prior to windowing, the data were resampled from 400 to 80 Hz using a finite impulse response antialiasing filter with a sharp cutoff at 40 Hz and the transition band made sufficiently sharp to minimize aliasing. This was performed to avoid spurious fitting to 50-Hz spectral peaks or any low-pass filter band edges. For the original EEG time series, the electromyogram (EMG—detected as the total power between 70 and 110 Hz excluding a notch at 98–102 Hz due to 50 Hz electric power harmonic at 100 Hz) was calculated [8], [19], [20]. The root mean square (rms) amplitude was also calculated from the resampled electroencephalogram time series. Subsequently, an automated artifact rejection method was used to classify all windows based on the original and resampled electroencephalogram time series. Windows were excluded from further analysis if any of the following occurred: total EMG power greater than approximately 400 μV^2 or less than approximately 0.004 μV^2, RMS amplitude less than 5 μV or greater than 150 μV, or amplitude distributions were not normal (based on Lilliefors test at $P < 0.01$).

The HFD analysis was calculated on similarly segmented and artifact rejected EEG data. However, instead of downsampling the raw data as was done for the ARMA analysis, the raw data were filtered using a 2024-order 6–47 Hz pass-band linear phase equiripple filter with sharp stop band cutoffs of 5.5 Hz (60-dB attentuation) and 47.5 Hz (80-dB attenuation) to minimize gamma band EMG content as previously described [7]. The HFD was then calculated on 1-s windows of this filtered data. In prior depth of anesthesia monitoring studies, the HFD was calculated on a 15-s window [7] whereas here we chose to calculate it on contiguous 1-s windows for direct comparison with our ARMA approach. Moreover, short windows are needed to sufficiently characterize the distribution of HFD values. In addition, it has been demonstrated that for the EEG the HFD produces stable values over window lengths from 1 to 15 s [33]. This was also verified for the anesthesia data considered here by looking at Pearsons correlation coefficient values between the HFD time series calculated on 1-s windows with zero overlap and 15-s windows with 14-s overlap for individual subjects full

TABLE II
TRAINING AND TESTING SETS OF CROSS-VALIDATION ANALYSIS

Set	Subject indices	# OAA/S epochs						# patients per OAA/S					
		0	1	2	3	4	5	0	1	2	3	4	5
Train 1	2, 3, 5, 6, 8, 10, 12, 13, 14, 15	9	2	8	10	31	49	9	2	7	4	10	10
Test 1	1, 4, 7, 9, 11	4	2	4	6	15	22	4	2	3	3	5	5
Train 2	1, 2, 4, 7, 8, 9, 10, 11, 13, 15	8	2	8	11	30	48	8	2	7	5	10	10
Test 2	3, 5, 6, 12, 14	5	2	4	5	16	23	5	2	3	2	5	5
Train 3	1, 3, 4, 5, 6, 7, 9, 11, 12, 14	9	3	8	11	31	45	9	3	6	5	10	10
Test 3	2, 8, 10, 13, 15	4	1	4	5	15	26	4	1	4	2	5	5

OAA/S epochs: number of OAA/S epochs used for each OAA/S score for each set; # patients per OAA/S: number of patients contributing data for each OAA/S score for each set.

recordings. Correlation values in the order of 0.9 were obtained (results not considered further).

G. Tracking/Classifying OAA/S State Using Distributions

The tracking/classification of the OAA/S score was performed using discrete distributions of either the estimated ARMA parameters (see Sections II-B and II-C) or HFD (see Section II-D). Our overall goal is to track the OAA/S score for an arbitrary individual undergoing general anesthesia during surgery. One way to know if a chosen approach is suitable for an arbitrary patient population is to apply an out-of-sample cross validation. This involves breaking up the data into training and testing sets that provide a balanced number of patients and number of OAA/S epochs for each OAA/S score consistent with the number of training and testing subjects. Moreover, the training sets were chosen such that not more than half of the patients are the same in each training set. This provides a balance between training and testing sets, while preserving some distance between the training sets. Three-fold cross validation was applied, meaning three train and test set pairs were created for validation. The data contained in each of these sets are summarized in Table II. OAA/S epochs were defined to be 30-s segments prior to the OAA/S measurement times. This duration corresponds approximately to the time it takes to obtain an OAA/S measurement. Only these OAA/S epochs were used in the cross-validation classification performance analysis, however, in practice the same training distributions can be used for tracking of brain states with continuous data. We also consider an example of tracking where new parameter estimates are obtained every second and the current distribution of parameter estimates is computed over the last 30 s of data and is updated every second. Here, "current distribution" essentially refers to the distribution of the most recent parameter estimates obtained from scrolling EEG data.

For the case of ARMA models, training distributions of ARMA parameters for each OAA/S score $P_{\mathrm{ijk}}^{(o)}$ were obtained from the OAA/S epochs for all of the subjects in a training set pooled together for each OAA/S score, where o indexes the OAA/S score and ijk are the dimensions of the discrete probability distribution, which correspond to the ARMA parameters. Thus, training does not involve any form of learning; it is simply the creation of distributions for each OAA/S score using the training set data. Then, the current OAA/S score is estimated by comparing the distributions of the ARMA parameters of a current 30-s (OAA/S) epoch $P_{\mathrm{ijk}}^{(t)}$ to six training distributions of the ARMA parameters corresponding to the 6 OAA/S scores. Specifically, the estimated OAA/S score for a current individual OAA/S epoch is taken to be the minimum of the total variation (TV) between the current distribution and each of the 6 OAA/S score training distributions:

$$\mathrm{TV}_t^{(o)} = \sum_{i,j,..,k} \frac{1}{2} \left| P_{ij..k}^{(o)} - P_{ij..k}^{(t)} \right| \tag{2}$$

$$\hat{O}_t = \mathrm{argmin}_o \mathrm{TV}_t^{(o)}. \tag{3}$$

TV gives a value between 0 for exact distribution match and 1 for zero overlap between distributions. The OAA/S index o with the smallest TV is chosen to be the estimate $\hat{O}_t$. This estimate is compared with the true OAA/S score for each epoch in order to assess the classification sensitivity for the training and testing sets.

For the sake of comparison, Jensen–Shannon divergence (JSD) as defined by [34] was also used to compute the distance between probability distributions instead of using TV.

Definition of discrete probability distributions: Discrete probability distributions were defined as multivariate functions of estimated parameters. To create the discrete probability distributions, each parameter dimension was divided into N bins where the bin centers were chosen to uniformly span the range of values defined by the training set. In addition, all bin widths were equal except for bins on the edges of the distribution which were allowed to cover the remaining possible range of values that could be encountered in an arbitrary test set (i.e., $-\infty$ to ∞). Generally, $N = 5$ bins were considered for each parameter dimension, however, values of $N = 3$ and $N = 7$ were also considered to analyze the effect of bin number on classification sensitivity. Memory issues are encountered if one tries to create a discrete probability distribution for an ARMA (8, 5) model since it has 14 dimensions (13 coefficients + 1 innovation standard deviation) and this involves 5^{14} bins for $N = 5$. Therefore, the analysis focuses on distributions with at most six dimensions. This still allows us for the estimation of an ARMA (8, 5) model, however, only the lower order coefficients are used when creating the probability distributions. To check that this was not a problem, an analysis was performed to see if the lower order coefficients vary the most with changes in depth of anesthesia and OAA/S score. This was tested by looking at the correlation between the estimated coefficient time series and the nearest neighbor interpolated OAA/S score time series using the Pearson correlation coeffcient ρ.

For the case of the analysis with HFD alone, since there is only one feature dimension, this HFD dimension was divided into 20 bins to compute the discrete probability distribution.

H. Tracking/Classifying OAA/S State Using Medians

For comparison, instead of computing the effective distance between the current distribution and each of the six OAA/S score training distributions, the normalized Euclidean distance between the median of the parameter estimates of the current OAA/S epoch and the medians of the parameter estimates defining each of the six OAA/S score training distributions was

employed separately to also classify OAA/S state. Here, normalized refers to each parameter value being divided by the parameter's maximal value across the training data before calculating the Euclidian distance. This normalization ensures no parameter can bias the distance estimate more than any other parameter. In this case, the OAA/S index o with the smallest normalized Euclidean distance between medians was chosen to be the OAA/S estimate $\hat{O}_t$. This analysis acts as a control to test if the distribution approach gives higher performance than an approach based on the average or the median. In this case, the median was considered instead of the mean as it is less sensitive to noisy outliers.

I. Evaluation of Classification Performance and Statistical Analysis

Performance evaluation focuses on total sensitivity, as well as sensitivity for each OAA/S class. Sensitivities are given as the proportion of correctly classified OAA/S epochs and take values between 0 and 1. Given that there are six OAA/S classes, the chance level performance for each OAA/S score sensitivity is $1/6 = 0.17$. The "nearest neighbor" sensitivity is also shown because of strong overlap in the distributions of ARMA parameters for adjacent OAA/S scores and nearest neighbor sensitivity accounts for this. Nearest neighbor sensitivity is defined by checking if the current OAA/S score estimate is within the true OAA/S score ± 1. In the following, training sensitivity and testing sensitivity refer to the sensitivities obtained for the training and testing sets, respectively.

Statistical comparison of the total testing sensitivity for the different methods was performed using SPSS Statistics 23 (IBM, Armonk, New York) and repeated measures general linear modeling (GLM) [35] with the different methods set as within-subject factors. A given sample corresponded to an individual's testing senstivity, where the sensitivities are derived from three classification "models," i.e., from the three training sets, and five testing individuals for each "model." To gain statistical power, the data were pooled over the "models" to give 15 independent individual testing sensitivities as the population sample for each method. A repeated measures model was used because the different methods were applied to the same testing data. Pairwise comparisons of methods involved a simple contrast where the method with the highest average sensitivity or the the method of interest was set as the reference depending on the set of methods being compared as described in the results. Statistical significance was assessed at the 0.05 level. The Benjamini–Hochberg procedure for correction of multiple comparisons based on controlling the false discovery rate at a level of 0.05 and taking into account the number of contrasts was applied [36]. The resultant p-values p, the F-statistic F, partial effect size η^2, and, in specific cases, power $P_\beta = 1 - \beta$ are reported.

To complement statistical tests, an *a priori* power analysis using G* Power 3.1.9.2 [37] was performed to find the number of subjects needed to gain standard statistical power of $P_\beta = 0.8$ for a given a significance level of 0.05 and the partial effect size computed by SPSS for specific underpowered repeated measures GLM tests indicated in the results.

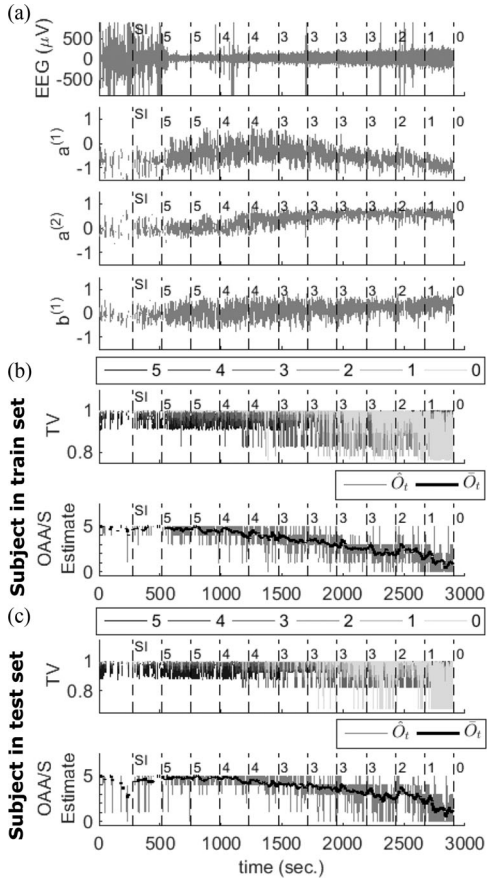

Fig. 1. Example of tracking depth of anesthesia using the ARMA (2, 1) parameter distribution approach with the Broersen technique for subject 9. (a) Raw EEG signal and ARMA (2, 1) parameter estimates vary with decreasing OAA/S scores. Tracking-based estimation of OAA/S score when the subject's data (b) contributes to the training set and (c) when it does not. Vertical-dashed lines and numbers indicate the time of OAA/S measurment and OAA/S score, respectively. SI: start of anaesthetic induction. White gaps in all graphs indicate periods where artifact has been removed. TV: total variation.

III. Results

A. Example of Tracking Depth of Anesthesia

An example of tracking depth of anesthesia using the ARMA (2, 1) parameter distribution approach with the Broersen technique is shown in Fig. 1 for subject 9. In Fig. 1(a), it can be seen that the subject's EEG (including strong artifacts at the start of the recording) varies with depth of anesthesia and associated reductions in responsiveness as quantified by the OAA/S score. The ARMA (2, 1) model parameters $a^{(1)}, a^{(2)}, b^{(1)}$ also vary with the level of responsiveness. Fig. 1(b) and (c) shows tracking-based estimation of the OAA/S score when the subject's data contributes to the training set (set 3 in Table II) and when it does not (set 1 in Table II, i.e., the subject is in the testing set), respectively. Here the current distribution of parameter estimates spans the most recent 30 s and is updated every second. In Fig. 1(b) and (c), the six different TV time series corresponding to each OAA/S score $\mathrm{TV}_t^{(o)}$ also vary with level of responsiveness, with the TV time series corresponding to the current true OAA/S score typically taking on the lowest value. Following (3), the time series of the OAA/S estimate $\hat{O}_t$ is

determined based on the minimum TV value for each 1-s window. In Fig. 1(b) and (c), the MA of the OAA/S estimate determined over the last 30 s $\bar{O}_t$ as well as the OAA/S estimate $\hat{O}_t$ decrease with the level of responsiveness.

B. Correlations and Low Versus High-Order ARMA Coefficients

As outlined in the methods, for the case of higher order ARMA models discrete probability distributions were calculated using the lower order coefficients. To show that this is sufficient, here it is shown by an example that for ARMA $(8, 5)$ models the lower order ARMA coefficients vary more than the higher order coefficients with changes in depth of anesthesia as reflected by the OAA/S score. The results given are for Kalman-based estimation ($\alpha = 1$) but also hold for Broersen-based estimation (results not shown). For the same participant in Fig. 1 (but with a higher model order), the Pearson correlation coeffcient ρ values between the 13 estimated coeffcients of the ARMA $(8, 5)$ model and the OAA/S score were 0.81, −0.82, 0.45, 0.15, −0.07, 0.27, −0.25, and −0.07 for the AR parameters $a^{(1)}$ to $a^{(8)}$, respectively, and 0.15, 0.68, 0.30, 0.53, 0.03 for the MA parameters $b^{(1)}$ to $b^{(5)}$, respectively. This example highlights that, generally, the lower order model coeffcients are best correlated with the OAA/S score. Similar results were obtained for other participants (results not shown). Therefore given the tradeoff with the computation time required to use probability distributions with several dimensions, it is sufficient to use only lower order coefficients in the probability distributions to track brain states using higher order ARMA models.

C. Correlations and Artifact Removal

The results of the Pearson correlation coeffcient analysis in the above example correspond to the case with artifact segments removed such that for the current accepted 1-s window the Kalman filter is initialized using values from the end of the last accepted window. If artifact segment removal was not applied and the Kalman filter, therefore, processed all data windows in a continous manner, for the same case in the previous paragraph, then the Pearson correlation coeffcient ρ values between the 13 estimated coeffcients of the ARMA $(8, 5)$ model and the OAA/S score were 0.50, −0.69, 0.37, −0.04, 0.06, 0.17, −0.00, and 0.15 for the AR parameters $a^{(1)}$ to $a^{(8)}$, respectively, and 0.04, 0.27, 0.34, 0.41, 0.19 for the MA parameters $b^{(1)}$ to $b^{(5)}$, respectively. Through the examples in this, and the previous paragraph, it can be observed that Pearson correlation coeffcient values between the ARMA coeffcients and the OAA/S score are higher for Kalman-based estimation when artifact segment removal is applied. Similar results were obtained for other participants (results not shown).

D. Distribution-Based Approach Cross-Validation Analysis

The results for OAA/S score classification sensitivity of the distribution-based approach for different ARMA model orders and a comparison between Kalman and Broersen estimation methods and the HFD are shown in Tables III and IV for training and testing data, respectively. These results correspond to the cross validation described in Section II-G. Moreover, the ARMA analyses correspond to the case when there are $N = 5$ bins per dimension. In Tables III and IV, the rows with bold sensitivities correspond to the best performing method. Moreover, "distribution parameters" refers to the variables defining the dimensions of the parameter/feature distribution.

TABLE III
AVERAGE TRAINING SENSITIVITY; $N = 5$ FOR ARMA, $N = 20$ FOR HFD

Method	Model order p, q	Parameters	Sensitivity Tot.	OAA/S class 0	1	2	3	4	5
B	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.71	0.76	1	0.75	0.84	0.4	0.88
B	2, 1	a^1, a^2, b^1	0.67	0.92	0.83	0.54	0.68	0.39	0.86
K, $\alpha = 1$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.68	0.84	0.92	0.54	0.63	0.50	0.81
K, $\alpha = 1.01$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.61	0.44	0.75	0.38	0.5	0.49	0.83
K, $\alpha = 1.02$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.61	0.54	0.75	0.33	0.53	0.48	0.79
B	6, 3	a^1, a^2, b^1	0.62	0.84	1	0.38	0.53	0.36	0.82
B	8, 5	$a^{1-3}, b^{1-2}, \sigma_{e_t}$	**0.73**	**0.93**	**1**	**0.88**	**0.85**	**0.41**	**0.88**
B	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.66	0.81	1	0.46	0.59	0.41	0.86
B	8, 5	a^1, a^2, b^1	0.63	0.88	1	0.5	0.5	0.37	0.82
K, $\alpha = 1.02$	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.58	0.62	0.83	0.33	0.49	0.41	0.77
HFD	NA	HFD	0.57	0.5	0.33	0.25	0.46	0.40	0.85
			Nearest Neighbor Sensitivity						
Method	Model order p, q	Parameters	Tot.	OAA/S class 0	1	2	3	4	5
B	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	**0.92**	**0.92**	**1**	**0.88**	**0.94**	**0.88**	**0.95**
B	2, 1	a^1, a^2, b^1	0.91	0.96	1	0.79	0.97	0.85	0.95
K, $\alpha = 1$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.87	0.96	1	0.79	0.81	0.83	0.92
K, $\alpha = 1.01$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.87	0.59	1	0.88	0.88	0.88	0.93
K, $\alpha = 1.02$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.89	0.93	1	0.79	0.81	0.88	0.94
B	6, 3	a^1, a^2, b^1	0.89	0.92	1	0.75	0.94	0.85	0.93
B	8, 5	$a^{1-3}, b^{1-2}, \sigma_{e_t}$	**0.92**	**1**	**1**	**1**	**0.91**	**0.83**	**0.96**
B	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.89	0.96	1	0.75	0.84	0.86	0.95
B	8, 5	a^1, a^2, b^1	0.88	0.96	1	0.71	0.85	0.83	0.94
K, $\alpha = 1.02$	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.89	0.92	1	0.79	0.8	0.88	0.92
HFD	NA	HFD	0.9	0.96	1	0.67	0.81	0.88	0.96

Distribution parameters: indicates parameters used to define distributions;
B: Broersen echnique; K: Kalman filtering with forgeting factor α;
HFD: Higuchi fractal dimension; NA: not applicable; Bold font indicates best performing case on average.

Generally from Tables III and IV it can be observed that the use of lower order ARMA models gives better testing sensitivity, while higher order models give better training sensitivity. In particular, in Table IV it can be seen that the distribution-based approach with the Broersen technique for the ARMA (2, 1) model and distribution parameters $a^{(1)}, a^{(2)}, b^{(1)}$, produced the highest average total testing sensitivity of 0.59 $[0.49, 0.68]$ (numbers in square brackets indicate the 95% confidence interval; "average" refers to average over cross-validation sets; "total" refers to total over all OAA/S classes). On the other hand, in Table III it can be seen that the Broersen technique applied to the higher order ARMA (8, 5) model and using parameters $a^{(1)}, a^{(2)}, a^{(3)}, b^{(1)}, b^{(2)}, \sigma_{e_t}$ produced the highest average total training sensitivity of 0.73 $[0.62, 0.84]$.

Statistical comparison of the testing sensitivity of the different methods in Table IV using repeated measures GLM revealed that when the ARMA (2, 1) model method that gave the highest average total testing sensitivity was set as the simple contrast reference and compared against the other ten methods, p-values below 0.05 were found for comparisons against

TABLE IV
AVERAGE TESTING SENSITIVITY; $N = 5$ FOR ARMA, $N = 20$ FOR HFD

Method	Model order p, q	Parameters	Sensitivity Tot.	OAA/S class 0	1	2	3	4	5
B	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.55	0.4	0.75	0.25	0.42	0.26	0.89
B	2, 1	a^1, a^2, b^1	**0.59**	**0.75**	**0.5**	**0.42**	**0.31**	**0.39**	**0.85**
K, $\alpha = 1$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.52	0.75	0.5	0.42	0.36	0.19	0.80
K, $\alpha = 1.01$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.52	0.3	0.0	0.17	0.23	0.43	0.84
K, $\alpha = 1.02$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.53	0.38	0.75	0.17	0.36	0.33	0.84
B	6,3	a^1, a^2, b^1	0.53	0.83	0.0	0.17	0.3	0.34	0.80
B	8, 5	$a^{1-3}, b^{1-2}, \sigma_{e_t}$	0.49	0.7	0.75	0.08	0.5	0.22	0.79
B	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.46	0.55	0.25	0.17	0.24	0.26	0.77
B	8, 5	a^1, a^2, b^1	0.49	0.92	0.25	0.25	0.24	0.28	0.72
K, $\alpha = 1.02$	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.54	0.52	0.5	0.25	0.37	0.45	0.73
HFD	NA	HFD	0.51	0.62	0.75	0.17	0.23	0.3	0.82
Nearest Neighbor Sensitivity									
Method	Model order p, q	Distribution parameters	Tot.	OAA/S class 0	1	2	3	4	5
B	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	**0.9**	**1**	**1**	**0.67**	**0.87**	**0.87**	**0.95**
B	2, 1	a^1, a^2, b^1	0.87	0.83	1	0.58	0.87	0.85	0.97
K, $\alpha = 1$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.86	1	0.75	0.83	0.67	0.87	0.88
K, $\alpha = 1.01$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	0.83	0.38	0.5	0.83	0.73	0.89	0.93
K, $\alpha = 1.02$	2, 1	$a^1, a^2, b^1, \sigma_{e_t}$	**0.9**	**0.83**	**1**	**0.75**	**0.8**	**0.93**	**0.95**
B	6, 3	a^1, a^2, b^1	0.86	0.92	1	0.67	0.93	0.8	0.92
B	8, 5	$a^{1-3}, b^{1-2}, \sigma_{e_t}$	0.83	1	0.75	0.67	0.62	0.83	0.9
B	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.82	0.92	1	0.75	0.62	0.83	0.86
B	8, 5	a^1, a^2, b^1	0.88	0.92	0.75	0.75	0.87	0.83	0.95
K, $\alpha = 1.02$	8, 5	$a^1, a^2, b^1, \sigma_{e_t}$	0.81	0.75	1	0.67	0.82	0.83	0.87
HFD	NA	HFD	0.88	1	1	0.58	0.87	0.83	0.97

Distribution Parameters: indicates parameters used to define distributions;
B: Broersen Technique; K: Kalman filtering with forgeting factor α;
HFD: Higuchi fractal dimension; NA: not applicable; Bold font indicates best performing case on average.

the ARMA (6, 3) model (average total testing sensitivity of 0.53 $[0.42, 0.63]$, $p = .005$, $F = 10.8$, $\eta^2 = .436$), the ARMA (8, 5) model using the Broersen technique and distribution parameters $a^1, a^2, b^1, \sigma_{e_t}$ (average total testing sensitivity of 0.46 $[0.37, 0.55]$, $p = .012$, $F = 8.3$, $\eta^2 = .371$) or a^1, a^2, b^1 (average total testing sensitivity of 0.49 $[0.38, 0.60]$, $p = .01$, $F = 8.8$, $\eta^2 = .386$), and the HFD (average total testing sensitivity of 0.51 $[0.41, 0.61]$, $p = .027$, $F = 6.1$, $\eta^2 = .304$). The above comparisons were significant after Benjamini–Hochberg correction for multiple comparisons, except for the comparison between ARMA (2, 1) and HFD which only trended significance with an adjusted significance level of 0.02 for the corresponding p-value of 0.027. It is also worth noting there was no statistical difference between the ARMA (2, 1) model performance when the Broersen or Kalman techniques were applied. For the repeated measures GLM test of the 11 methods, the obtained overall effect size was $\eta^2 = .064$ and observed power was $P_\beta = 0.49$. *A priori* power analysis for this effect size suggests that power of $P_\beta = 0.8$ can be obtained with 26 subjects.

When considering nearest neighbor sensitivity, statistical comparison of the methods in Table IV using repeated measures GLM with Broersen estimation applied to the ARMA (2, 1) model with distribution parameters a^1, a^2, b^1 (average total nearest neighbor testing sensitivity of 0.87 $[0.75, 0.98]$) set as the contrast reference, a p-value below 0.05 was only found for the comparison with the ARMA (8, 5) model using the Kalman filter with forgetting ($\alpha = 1.02$) and distribution parameters $a^1, a^2, b^1, \sigma_{e_t}$ (average total nearest neighbor testing sensitivity of 0.81 $[0.68, 0.94]$, $p = .018$, $F = 7.2$, $\eta^2 = .338$). This comparison was not significant after Benjamini–Hochberg correction.

Statistical comparison of the different Kalman methods applied to the ARMA (2, 1) model in Table IV using repeated measures GLM with standard Kalman filtering ($\alpha = 1$) set as the contrast reference, revealed no statstically significant differences between testing sensitivities of the Kalman methods.

1) Effect of Bin Number on Performance: The effect of bin number per parameter dimension of the probability distributions N on classification sensitivity was analyzed for two ARMA model cases: 1) the case giving the best test sensitivity for $N = 5$—the Broersen technique for model order (2, 1) with the distributions defined by the three parameters $a^{(1)}, a^{(2)}, b^{(1)}$; and 2) the highest considered model order case—the Broersen technique for model order (8, 5) with the distributions defined by the six parameters $a^{(1)}, a^{(2)}, a^{(3)}, b^{(1)}, b^{(2)}, \sigma_{e_t}$.

For the statistical comparison with the ARMA (2, 1) model for the three different bin numbers using repeated measures GLM with the $N = 5$ method (average total testing sensitivity of 0.59 $[0.49, 0.68]$) set as the contrast reference, a p-value below 0.05 was obtained for the comparison against $N = 3$ (average total testing sensitivity of 0.51 $[0.41, 0.61]$, $p = .042$, $F = 5.0$, $\eta^2 = .264$) but not for the comparison against $N = 7$ (average total testing sensitivity of 0.58 $[0.48, 0.68]$). Moreover, the comparison for $N = 3$ was not significant after Benjamini–Hochberg correction for the number of contrasts considered. Thus, more bins (i.e., $N = 7$) appear to give comparable performance, while less bins (i.e., $N = 3$), although not statistically significant, trend toward giving weaker performance.

For the second case, the ARMA (8, 5) model for the three different bin numbers using repeated measures GLM with the $N = 5$ method set as the contrast reference revealed no statistically significant differences between the performance for $N = 5$ (average total testing sensitivity of 0.49 $[0.39, 0.58]$) and $N = 3$ (average total testing sensitivity of 0.49 $[0.39, 0.58]$) or $N = 7$ (average total testing sensitivity of 0.49 $[0.42, 0.57]$). This suggests for higher order ARMA models with large numbers of parameters defining the distributions (6 in this case) that $N = 5$ bins per dimension adequately characterizes the required distributions to obtain reasonable testing performance despite the fact that the distributions contain a large number of bins (5^6 in this case).

2) Effect of Distribution Distance Measure on Performance: For sake of comparison, an alternative measure to compute the distances between probability distributions, JSD was used in place of TV to perform the cross-validation analysis. As such, repeated measures GLM applied for testing senstivities for the ARMA (2, 1) model with the Broersen technique focused on parameters $a^{(1)}, a^{(2)}$, and $b^{(1)}$ revealed no statistically significant differences between when TV (average total testing sensitivity of 0.59 $[0.49, 0.68]$) and JSD (average total testing sensitivity of 0.58 $[0.48, 0.68]$) were used.

3) Comparing Classification Using Distributions or Medians: To check whether using distributions to track brain state is more effective than using average or median values, the normalized Euclidean distance (see methods) between the median of the parameter estimates of the current OAA/S epoch and the medians of the parameter estimates defining each of the six OAA/S score training distributions was employed separately to also classify OAA/S state. For the ARMA (2, 1) model with the Broersen technique focused on parameters $a^{(1)}, a^{(2)}$, and $b^{(1)}$, repeated measures GLM applied for testing senstivities revealed no statistically significant differences in performance between the distribution approach (average total testing sensitivity of 0.59 $[0.49, 0.68]$) and the median approach (average total testing sensitivity of 0.56 $[0.44, 0.69]$). For this repeated measures GLM test, the obtained effect size was $\eta^2 = .021$ and observed power was $P_\beta = 0.08$. *A priori* power analysis for this effect size suggests that power of $P_\beta = 0.8$ can be obtained with 369 subjects.

IV. Discussion

A. Model-Based Tracking of Brain States

A method to track brain states from the EEG based on characterizing the distribution of estimated ARMA model parameters has been demonstrated here. This approach shows promise when compared with a high performing depth of anesthesia monitoring feature, HFD, which has been shown to be as good as, or better than, three entropy-based features, two commercial entropy-based features, and three features used in the most common commercial depth of anesthesia monitoring device, the BIS monitor (Covidian, Ireland) [7]. Studies with larger numbers of subjects and a variety of anesthetics with different molecular modes of action [38] will be required to properly evaluate the performance of this approach when compared to a high performing depth of anesthesia monitoring features. Generally, it was found that a lower order ARMA (2, 1) model estimated with the Broersen windowed ARMA method gave higher average testing sensitivity when the three model parameters defined the distributions. On the other hand, a higher order ARMA (8, 5) model gave better training sensitivity when six model parameters were used to construct the discrete probability distributions. This might reasonably be expected as it is well known in the pattern classification literature [32] that too many model parameters or features can lead to overfitting during training, and thus better training performance but poorer generalization ability and testing performance. Although after correction for multiple comparisons there was no statistically significant difference between the performance of the Broersen technique using the distribution approach and the Broersen technique using the median approach, or between the performance of the best-on-average ARMA method and HFD (the difference only trended significance), it can still be noted that the average total testing sensitivity for the ARMA methods and HFD at least appears to be comparable. Moreover, the *a priori* power analysis based on effect sizes of the statistical tests suggests that greater statistical power and reduction of the likelihood of concluding there is no effect when there is an effect can be gained from having more subjects. It is worth noting that the statistical analysis results presented here are likely affected by the variability of testing sensitivities across subjects as indicated by the 95% confidence intervals noted in the results. This variability may reflect the differing susceptibility of different subjects to anesthesia and suggests that a more patient-specific approach may be warranted. This notion of individual specific analysis is supported by studies of slow-wave activity and propofol [39]. However, it is difficult to perform out-of-sample testing with a patient-specific approach as it is potentially unhealthy for individuals to be anesthetized too often. Therefore, a hybrid approach may be desirable where a monitor trained on several subjects is calibrated using a given individual's resting EEG in order to then track that individual's anesthetic brain state.

Although low-order ARMA (2, 1) models gave better testing performance than the model orders (6, 3) and (8, 5), which correspond to the model orders of physiologically based models of EEG [25], [29], these higher orders still gave reasonable testing sensitivity. This suggests that approaches based on physiologically based models may also work, however, one needs to remain aware of the effects of parameter overfitting on out-of-sample testing performance since these models can have large numbers of parameters [1], [2].

Model-based approaches for tracking brain states and brain connectivity in real-time are increasingly being seen as necessary and, thus, a number of methods are beginning to emerge [1], [2], [5]. At present, these methods rely on a range of stochastic approaches, such as unscented Kalman filtering [1], [5] and Bayesian estimation [2], or deterministic observers [3]. Recently, a model-based approach was developed to predict epileptic seizures in a patient-specific manner [4]. In this case, it was based on a modification of the Jansen–Rit model and involved a window-based Bayesian inversion method to estimate the model parameters. Although high sensitivity and low specificity was obtained, seizure prediction is a difficult problem with limitations regarding adequate numbers of seizures and interseizure data to enable proper method evaluation [4]. For the case of anesthesia investigated here, there are different difficulties. In particular, developing a method that is not necessarily tailored to a specific individual and can also classify six OAA/S states of responsiveness (as opposed to just two: preseizure and nonpreseizure). The linear modeling methods considered here will provide a useful reference point for studies using nonlinear neural models to track anesthetic brain states [5].

Given that the OAA/S score is only an indirect, and arguably subjective, measure of awareness, which is somewhat tenuously conceived as continuous [40], it is expected that there will be a significant overlap in the brain states and associated model parameter values corresponding to neighboring OAA/S scores. We provided tentative support for such a notion by showing that the nearest neighbor sensitivity is much higher than the standard sensitivity score, while both are much higher than chance sensitivity. Anesthesia EEG data also comes with its limitations. In particular, for this dataset there are many more epochs corresponding to OAA/S scores of 4 or 5 as compared to lower values except for OAA/S 0. This is because the anesthetic propofol is titrated in fixed steps that may skip certain OAA/S scores for different individuals based on their sensitivity to the drug.

Despite this bias, for the best case considered nearest neighbor testing sensitivity was found to be high for all OAA/S scores, and standard sensitivity was generally above chance for all scores.

Given that there is no gold standard for the assessment of anesthetic brain state or hypnotic level [11], the OAA/S score has been used as a surrogate marker or correlate of the anesthetic state. Whether linear or nonlinear brain models can describe specific levels of anesthesia remains an open problem, although significant progress has been made in the modeling of EEG and anesthesia [41]–[43]. Given that there is no clear way to define anesthetic brain states, apart from operational measures such as the OAA/S score, the true yard stick in this field is more likely to be whether or not certain approaches lead to positive outcomes for patients [10], [12], [17].

B. Relationship to Other Distribution-Based Methods

Our approach for estimating the distributions of model parameters is similar to the wavelet-coefficient distribution approach used by Zikov *et al.* [22] and is somewhat different to the more commonly instantiated Bayesian or particle filters [6]. The main reason for taking our approach is to extend our existing methods for depth of anesthesia monitoring using ARMA modeling and the Broersen technique of windowed ARMA estimation [8]. Further our approach is expected to be significantly more computationally efficient than particle filtering as it can be run in real-time without need for more complex and optimized parallel programing techniques. Moreover, with regards to efficiency, while the two considered measures of the distance between two probability distributions, TV and JSD, gave similar testing sensitivity, TV is more computationally efficient and therefore preferred in practice.

C. Depth of Anesthesia Monitoring

Generally, previous depth of anesthesia monitoring methods have been evaluated by pooling data across all subjects and correlating depth of anesthesia features, such as HFD, with the OAA/S state, other behavioral measures, or drug concentration [7]–[9]. However, this somewhat simplistic correlational analysis will not reveal how well a given processed EEG feature will classify a new individual patient's brain state. The three-fold out-of-sample cross-validation analysis of epochs preceding OAA/S measurements presented here overcomes this problem by evaluating out-of-sample performance.

In terms of feature-based classification, only the HFD was considered. It is possible to combine HFD with other features in order to boost classification performance [16], [18], however, the main goal here is to demonstrate that a model-based distribution approach is comparable to an approach based on a high performing depth of anesthesia monitoring feature.

Although the subjects here were brought to loss of consciousness, it would be interesting to consider the effects of deeper propofol concentrations and also emergence from anesthesia. For the subjects considered here, the deepest predicted effect-site propofol concentration [7] experienced was 4.25 μg/ml. In other recent studies looking at multichannel EEG signatures of propofol anesthesia and focusing on induction and emergence [44], subjects experienced propofol levels as deep as 5 μg/ml. These studies have demonstrated EEG signatures based on spectral, coherence, and phase-amplitude information [44] that delinate the boundaries of propofol-induced loss of consciousness, recovery of consciousness, and deep unconsciousness. However, a method for tracking anesthetic brain state and evaluating its tracking performance based on this approach still needs to be performed. This approach, if consistent across large numbers of subjects, offers the potential for more timely detection or prediction of anesthetic state changes, in particular in the transitions from consciousness to unconsciousness and back again. The distribution-based approach presented here evaluates the current distribution of parameter estimates or feature values over the previous 30 s of data in order to create adequately sampled distributions, and this means the approach is less responsive to immediate transient changes in anesthetic state. However, the use of a distribution-based approach should be robust in the clinical environment where transient artifacts are common.

Regarding deeper levels of propofol anesthesia, that include EEG features such as burst suppression [43], linear modeling approaches like ARMA may not suffice and it may be more appropriate to consider anesthetic state tracking using more realistic (nonlinear) neural models of EEG and anesthesia [5], [41]–[43]. Such neural models may yield improvements over ARMA-based approaches as they take into account known physiology [5]. It should be stated, however, that computational models of neural activity will always be an approximation of the real system. Thus, the main point will be whether a model can be simple enough to allow us for computationally efficient real-time tracking of brain states, but also complex enough to provide accurate tracking of brain states, while at the same time potentially providing the estimates of underlying anesthetic-induced physiological changes, such as population average membrane potential or postsynaptic potential changes, based on model state and parameter estimates. Although such estimates may be the approximations of the true underlying physiological variables, they may still provide information that has a useful physiological interpretation in clinical anesthesia.

Given that the EEG is not very sensitive to deep brain sources [45], the application of model-based approaches to the EEG signal may not be sensitive to the effects of deeper regions which could be responsible for observed cortical effects. This in part depends on the model, as one can always consider modeling deep brain structures in more detail, such as the thalamus [41]. At the same time, however, models of cerebral cortex alone have been able to demonstrate many of the key EEG features of anesthesia [42], [43]. Anesthesia studies involving magnetoencephalography or functional magnetic resonance imaging [46] may assist in overcoming the limitations of EEG with respect to deep brain areas.

Regarding other depth of anesthesia monitoring approaches [7], [8], [11], [12], [15]–[20], although commercial monitors can improve anesthetic delivery and postoperative recovery [17] more work is required to reliably reduce intraopertive awareness [12] and to ensure the methods can track the effects of anesthetics with different molecular modes of action [47]. Recent approaches reflecting reconfiguration of brain activity during anesthesia using single-channel EEG [18], [48], [49] or

multichannel EEG, and in some cases functional magnetic resonance imaging, [44], [46], [50], [51] hold significant promise for advancing depth of anaesthesia monitoring.

In single-channel-EEG studies involving nonlinear analysis measures like HFD, it has been shown that measures of order pattern analysis such as permutation entropy outperform other nonlinear measures like approximate entropy [48], [49]. Likewise HFD has been shown to outperform entropy measures such as approximate entropy [7]. It would be possible to compare order of pattern analysis measures to HFD here, however, as mentioned above the primary goal of this paper is to explore different ARMA model-based approaches and demonstrate that they can compete with a high performing depth of anesthesia monitoring measure like HFD.

Another important aspect of depth of anesthesia monitoring not considered in detail here is the recent demonstration that the most common commercial depth of anesthesia monitoring device, the BIS monitor (Covidian, Ireland), is sensitive to increasing levels of neuromuscular blockade and associated EMG changes [52], suggesting that its ability to track anesthetic depth is confounded when neuromuscular blockade is combined with general anesthesia. Here, we detected EMG in the 70–110 Hz band as a surrogate for broadband EMG that can be present above 30 Hz [53]. We found that power detected in the 70–110 Hz band was highly correlated to power in the 30–110 Hz band across the anesthetic levels and that the EMG power in the 30–110 Hz and 70–110 Hz bands only correlated with the estimated OAA/S state of the ARMA (2, 1) model method at OAA/s levels of 5, before the anesthetic started to have effects and when both EMG power and OAA/S estimate were relatively flat (methods/results not shown here). This supports the use of power in the 70–110 Hz band as a surrogate for detecting broadband EMG and suggests that the ARMA modeling approach presented here is not heavily influenced by the EMG content. However, clear conclusions regarding EMG effects can only be made by applying the methods here to EEG recordings obtained during neuromuscular blockade [52].

V. Conclusion

This paper suggests (for propofol-based anesthesia) that linear model-based distribution approaches can at least achieve similar performance when compared to a distribution approach based on an existing high performing depth of anesthesia monitoring measure. Moreover, the linear model-based distribution approach is not necessarily more effective at classifying depth of anesthesia than a median approach. Future studies should consider more subjects in order to guarantee adequate power for statistical comparisons, deeper levels of anesthesia in order to better characterize the utility of these methods at all clinically important anesthetic concentrations, and the effects of different anesthetics. The same techniques presented here may be generally applicable to the tracking of other brain states using electrophysiological measurements.

Acknowledgment

D. T. J. Liley holds an unvalued equity stake in Cortical Dynamics Pvt. Ltd.

References

[1] D. Freestone *et al.*, "Patient-specific neural mass modelling: Stochastic and deterministic methods, recent advances in predicting and preventing epileptic seizures," in *Recent Advances Predicting Preventing Epileptic Seizures*. Boca Raton, FL, USA: CRC Press, 2013, pp. 63–82.

[2] R. Moran *et al.*, "Neural masses and fields in dynamic causal modeling," *Frontiers Comput. Neurosci.*, vol. 7, 2013, Art. no. 57.

[3] M. S. Chong *et al.*, "Parameter and state estimation of nonlinear systems using a multi-observer under the supervisory framework," *IEEE Trans. Automat. Control*, vol. 60, no. 9, pp. 2336–2349, Sep. 2015.

[4] A. Aarabi and B. He, "Seizure prediction in hippocampal and neocortical epilepsy using a model-based approach," *Clin. Neurophysiol.*, vol. 125, pp. 930–940, 2013.

[5] L. Kuhlmann *et al.*, "Neural mass model-based tracking of anesthetic brain states," *NeuroImage*, vol. 133, pp. 438–456, 2016.

[6] D. Simon, *Optimal State Estimation: Kalman, H-infinity and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley, 2006.

[7] R. Ferenets *et al.*, "Behavior of entropy/complexity measures of the electroencephalogram during propofol-induced sedation: Dose-dependent effects of remifentanil," *Anesthesiology*, vol. 106, no. 4, pp. 696–706, 2007.

[8] D. Liley *et al.*, "Propofol and remifentanil differentially modulate frontal electroencephalographic activity," *Anesthesiology*, vol. 113, pp. 1–13, 2010.

[9] J. C. Sigl and N. G. Chamoun, "An introduction to bispectral analysis for the electroencephalogram," *J. Clin. Monit.*, vol. 10, no. 6, pp. 392–404, 1994.

[10] P. Myles *et al.*, "Bispectral index monitoring to prevent awareness during anaesthesia: The b-aware randomised controlled trial," *Lancet*, vol. 363, no. 9423, pp. 1757–1763, 2004.

[11] J. Bruhn *et al.*, "Depth of anaesthesia monitoring: What's available, what's validated and what's next?" *Brit. J. Anaesthesia*, vol. 97, no. 1, pp. 85–94, 2006.

[12] G. A. Mashour *et al.*, "Prevention of intraoperative awareness with explicit recall in an unselected surgical population: A randomized comparative effectiveness trial," *Anesthesiology*, vol. 117, no. 4, pp. 717–725, 2012.

[13] S. S. Liu, "Effects of bispectral index monitoring on ambulatory anesthesiaa meta-analysis of randomized controlled trials and a cost analysis," *J. Amer. Soc. Anesthesiologists*, vol. 101, no. 2, pp. 311–315, 2004.

[14] M. T. Chan *et al.*, "Bis-guided anesthesia decreases postoperative delirium and cognitive decline," *J. Neurosurgical Anesthesiol.*, vol. 25, no. 1, pp. 33–42, 2013.

[15] M. Struys *et al.*, "Performance of the ARX-derived auditory evoked potential index as an indicator of anesthetic depth: A comparison with bispectral index and hemodynamic measures during propofol administration," *Anesthesiology*, vol. 96, no. 4, pp. 803–816, 2002.

[16] D. Jordan *et al.*, "EEG parameters and their combination as indicators of depth of anaesthesia/EEG-parameter und deren kombination für das narkosemonitoring," *Biomedizinische Technik*, vol. 51, no. 2, pp. 89–94, 2006.

[17] Y. Punjasawadwong *et al.*, "Bispectral index for improving anaesthetic delivery and postoperative recovery," *Cochrane Library*, no. 6, 2014.

[18] G. Schneider *et al.*, "Monitoring depth of anesthesia utilizing a combination of electroencephalographic and standard measures," *Anesthesiology*, vol. 120, no. 4, pp. 819–828, 2014.

[19] M. Shoushtarian *et al.*, "Comparisons of electroencephalographically derived measures of hypnosis and antinociception in response to standardized stimuli during target-controlled propofol-remifentanil anesthesia," *Anesthesia Analgesia*, vol. 122, no. 2, pp. 382–392, 2016.

[20] M. Shoushtarian *et al.*, "Evaluation of the brain anaesthesia response monitor during anaesthesia for cardiac surgery: A double-blind, randomised controlled trial using two doses of fentanyl," *J. Clin. Monit. Comput.*, 2015, to be published.

[21] P. M. Broersen, "Automatic spectral analysis with time series models," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 2, pp. 211–216, May 2002.

[22] T. Zikov *et al.*, "Quantifying cortical activity during general anesthesia using wavelet analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 4, pp. 617–632, Apr. 2006.

[23] D. A. Chernik *et al.*, "Validity and reliability of the observer's assessment of alertness/sedation scale: Study with intravenous midazolam," *J. Clin. Psychopharmacol.*, vol. 10, no. 4, pp. 244–251, 1990.

[24] I. Bojak and D. T. Liley, "Modeling the effects of anesthesia on the electroencephalogram," *Phys. Rev. E*, vol. 71, 2005, Art. no. 041902.

[25] D. T. Liley *et al.*, "Dissociating the effects of nitrous oxide on brain electrical activity using fixed order time series modeling," *Comput. Biol. Med.*, vol. 38, pp. 1121–1130, Oct. 2008.

[26] D. T. Liley *et al.*, "Drug-induced modification of the system properties associated with spontaneous human electroencephalographic activity," *Phys. Rev. E*, vol. 68, no. 5, 2003, Art. no. 051906.
[27] C. Jeleazcov *et al.*, "Electroencephalogram monitoring during anesthesia with propofol and alfentanil: The impact of second order spectral analysis," *Anesthesia Analgesia*, vol. 100, no. 5, pp. 1365–1369, 2005.
[28] M. Tarvainen *et al.*, "Estimation of nonstationary EEG with Kalman smoother approach: An application to event-related synchronization (ERS)," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 3, pp. 516–524, Mar. 2004.
[29] D. Buente *et al.*, "Complex dynamics for a reduced model of human EEG: Implications for the physiological basis of brain activity," *BMC Neurosci.*, vol. 12, 2011, Art no. 198.
[30] P. Stoica and Y. Selen, "Model order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.
[31] S. Tseng *et al.*, "Evaluation of parametric methods in EEG signal analysis," *Med. Eng. Phys.*, vol. 17, pp. 71–78, 1995.
[32] R. O. Duda *et al.*, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.
[33] A. Accardo *et al.*, "Use of the fractal dimension for the analysis of electroencephalographic time series," *Biol. Cybern.*, vol. 77, no. 5, pp. 339–350, 1997.
[34] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[35] K. Kim and N. Timm, *Univariate and Multivariate General Linear Models: Theory and Applications With SAS*. Boca Raton, FL, USA: CRC Press, 2006.
[36] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Series B (Methodological)*, vol. 57, pp. 289–300, 1995.
[37] F. Faul *et al.*, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior Res. Methods*, vol. 41, no. 4, pp. 1149–1160, 2009.
[38] U. Rudolph and B. Antkowiak, "Molecular and neuronal substrates for general anaesthetics," *Nature Rev. Neurosci.*, vol. 5, pp. 709–720, 2004.
[39] R. N. Mhuircheartaigh *et al.*, "Slow-wave activity saturation and thalamocortical isolation during propofol anesthesia in humans," *Sci. Translational Med.*, vol. 5, no. 208, pp. 208ra148–208ra148, 2013.
[40] G. Tononi, "Consciousness as integrated information: A provisional manifesto," *Biol. Bull.*, vol. 215, no. 3, pp. 216–242, 2008.
[41] R. Hindriks and M. J. van Putten, "Meanfield modeling of propofol-induced changes in spontaneous EEG rhythms," *Neuroimage*, vol. 60, no. 4, pp. 2323–2334, 2012.
[42] A. Hutt and L. Buhry, "Study of GABAergic extra-synaptic tonic inhibition in single neurons and neural populations by traversing neural scales: Application to propofol-induced anaesthesia," *J. Comput. Neurosci.*, vol. 37, no. 3, pp. 417–437, 2014.
[43] I. Bojak *et al.*, "Emergence of spatially heterogeneous burst suppression in a neural field model of electrocortical activity," *Frontiers Syst. Neurosci.*, vol. 9, 2015.
[44] P. L. Purdon *et al.*, "Electroencephalogram signatures of loss and recovery of consciousness from propofol," *Proc. Nat. Acad. Sci.*, vol. 110, no. 12, pp. E1142–E1151, 2013.
[45] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. New York, NY, USA: Oxford, 2006.
[46] D. Jordan *et al.*, "Simultaneous electroencephalographic and functional magnetic resonance imaging indicate impaired cortical top-down processing in association with anesthetic-induced unconsciousness," *Anesthesiology*, vol. 119, no. 5, pp. 1031–1042, 2013.
[47] K. Hirota, "Special cases: Ketamine, nitrous oxide and xenon," *Best Pract. Res. Clin. Anaesthesiol.*, vol. 20, no. 1, pp. 69–79, 2006.
[48] D. Jordan *et al.*, "Electroencephalographic order pattern analysis for the separation of consciousness and unconsciousness," *Anesthesiology*, vol. 109, no. 6, pp. 1014–22, 2008.
[49] E. Olofsen *et al.*, "Permutation entropy of the electroencephalogram: A measure of anaesthetic drug effect," *Brit. J. Anaesthesia*, vol. 101, no. 6, pp. 810–821, 2008.
[50] L. Kuhlmann *et al.*, "Modulation of functional EEG networks by the NMDA antagonist nitrous oxide," *PloS one*, vol. 8, no. 2, 2013, Art. no. e56434.
[51] U. Lee *et al.*, "Disruption of frontal-parietal communication by ketamine, propofol, and sevoflurane," *Anesthesiology*, vol. 118, no. 6, pp. 1264–1275, 2013.
[52] P. Schuller *et al.*, "Response of bispectral index to neuromuscular block in awake volunteers," *Brit. J. Anaesthesia*, vol. 115, no. suppl 1, pp. i95–i103, 2015.
[53] V. Bonhomme and P. Hans, "Muscle relaxation and depth of anaesthesia: Where is the missing link?" *Brit. J. Anaesthesia*, vol. 99, no. 4, pp. 456–460, 2007.

Authors photographs and biographies not available at the time of publication.