A LOCATION-MIXTURE AUTOREGRESSIVE MODEL FOR ONLINE FORECASTING OF LUNG TUMOR MOTION

By Daniel Cervone[*], Natesh S. Pillai[*], Debdeep Pati[†], Ross Berbeco[‡] and John Henry Lewis[‡]

Harvard University[*], *Florida State University*[†], *Brigham and Women's Hospital, Dana-Farber Cancer Institute and Harvard Medical School*[‡]

Lung tumor tracking for radiotherapy requires real-time, multiple-step ahead forecasting of a quasi-periodic time series recording instantaneous tumor locations. We introduce a location-mixture autoregressive (LMAR) process that admits multimodal conditional distributions, fast approximate inference using the EM algorithm, and accurate multiple-step ahead predictive distributions. LMAR outperforms several commonly used methods in terms of out-of-sample prediction accuracy using clinical data from lung tumor patients. With its superior predictive performance and real-time computation, the LMAR model could be effectively implemented for use in current tumor tracking systems.

1. Introduction. Real-time tumor tracking is a promising recent development in External Beam Radiotherapy (XRT) for the treatment of lung tumors. In XRT, a compact linear accelerator is used to deliver photon radiation to the tumor locations in a narrow beam, minimizing exposure to nearby healthy tissue. As the location of the lung tumor is in constant motion due to respiration, some patients who undergo this treatment are implanted with a small metal marker (known as a fiducial) at the location of a tumor. During XRT, X-ray imaging reveals the location of the fiducial, thus providing the desired target of the radiation beam. Tumor tracking is an advanced technology that minimizes normal tissue exposure by moving the radiation beam to follow the tumor position [Rottmann, Keall and Berbeco (2013), D'Souza, Naqvi and Cedric (2005), Schweikard et al. (2000)]. However, there is a system latency of 0.1–1.0 seconds (depending on the equipment used) that causes the aperture of the radiation beam to lag behind the real-time location of the tumor. This latency is estimated empirically by comparing the motion history of the fiducial and radiation beam aperture. For tumor

Keywords and phrases: lung tumor tracking, external beam radiotherapy, nonlinear time series, mixture autoregressive process, time series motifs, likelihood approximation, multiple-step prediction

tracking XRT to be successful, hardware and software system latencies must be overcome by the introduction of a predictive algorithm.

As accurate radiotherapy is essential for both minimizing radiation exposure to healthy tissue and ensuring the tumor itself is sufficiently irradiated, the subject of predicting tumor motion to overcome the system latency has received a good deal of attention in the medical community. Any possible forecasting approach must provide k-step ahead predictive distributions in real-time, where k is approximately equal to the system latency multiplied by the sampling frequency of the tumor tracking imagery. Real-time forecasting requires that a (k-step ahead) prediction be made before any further data on the tumor's motion has been recorded.

Statistical methods for tumor prediction in the literature include penalized linear models (e.g., Sharp et al. (2004) and many others), the Kalman filter [Murphy, Isaakson and Jalden (2002)], state-space models [Kalet et al. (2010)], and wavelets [Ernst, Schlaefer and Schweikard (2007)]; machine learning methods include kernel density estimation [Ruan and Keall (2010)], support vector regression [Riaz et al. (2009), Ernst and Schweikard (2009)], and neural networks [Murphy, Isaakson and Jalden (2002), Murphy and Dieterich (2006)]. All of these examples include simulations of out-of-sample prediction using real patient data in order to assess forecasting accuracy. Because predictive performance varies considerably from patient to patient and across different equipment configurations, of particular importance to the literature are comparisons of different prediction methods for the same set of patients with the same conditions for data preprocessing [Sharp et al. (2004), Krauss, Nill and Oelfke (2011), Ernst et al. (2013)]. While standard, "off-the-shelf" time series forecasting models can be applied to lung tumor tracking, better predictive performance can be achieved with a model that explicitly incorporates the dynamics of respiratory motion.

We propose a novel time series model which we call a location-mixture autoregressive process (LMAR). A future observation (Y_n) given the observed history of the time series is assumed to follow a Gaussian mixture,

$$Y_n | Y_{n-1}, Y_{n-2}, \ldots \sim \sum_{j=1}^{d_n} \alpha_{n,j} \mathrm{N}(\mu_{n,j}, \sigma^2), \tag{1.1}$$

where $\sum_{j=1}^{d_n} \alpha_{n,j} = 1$, and $\mu_{n,j}$ is of the form

$$\mu_{n,j} = \tilde{\mu}_{n,j} + \sum_{l=1}^{p} \gamma_l Y_{n-l}. \tag{1.2}$$

We refer to this as a location-mixture autoregressive model because the autoregressive part of the component means, $\sum_{l=1}^{p} \gamma_l Y_{n-l}$, is the same for all j, and only the location parameter, $\tilde{\mu}_{n,j}$, changes across the components in (1.1). Our model differs from other time series models that yield mixture-normal conditional distributions (e.g., the class of threshold autoregressive models [Tong and Lim (1980)], including Markov-switching autoregressive models [Hamilton (1989)] and the mixture autoregressive models of Wong and Li (2000)) in that $\tilde{\mu}_{n,j}$ in (1.2) depends on an unknown subseries of the time series, at least p observations in the past. The mixture weights, $\{\alpha_{n,j}\}$, also depend on the entire history of the observed time series, and the number of mixture components in our model, d_n, increases with n.

Another noteworthy characteristic of our model is that all parameters in (1.1) are obtained from a single, unknown, $(p+1) \times (p+1)$ positive definite matrix. This parsimonious parameterization is motivated in part by the need for real-time parameter estimation and forecasting. Compared with other mixture autoregressive models, LMAR is simpler to fit and admits accurate closed-form expressions for k-step ahead predictive distributions. While the data application we consider shows the promise and appeal of the LMAR model, we believe a thorough treatment of its theoretical properties (a future endeavor) is necessary before the LMAR model is a viable "off-the-shelf" method for diverse data sets.

We motivate our model in the context of time series motifs, which offers a geometric interpretation of the components in our model. In general terms, motifs catalog recurring pattens in time series and are commonly used in data mining tasks for which a symbolic representation of a time series is useful, such as event detection and time series clustering or classification [Lin et al. (2002), Ye and Keogh (2009), Tanaka, Iwamoto and Uehara (2005), Fu (2011)]. For the purposes of forecasting, predictive state representations [Littman, Sutton and Singh (2002), Shalizi (2003), Boots and Gordon (2011)] categorize time series motifs not as subseries of the observed data, but as equivalence classes of conditional predictive distributions.

Section 2 of this paper discusses the important features of the data we use and graphically motivates our model. Section 3 formally introduces the LMAR model and describes parameter estimation and forecasting using principled methods that are feasible in real-time. Section 4 describes the procedure for comparing out-of-sample prediction error under our model with competing forecasting methods for tumor tracking, including the selection of tuning parameters. The results of this comparison are discussed in Section 5, and Section 6 summarizes and points out future directions.

2. Tumor tracking data. We have data on 11 patients treated at the Radiation Oncology Clinic at the Nippon Telegraph and Telephone Corporation Hospital in Sapporo, Japan. A detailed discussion of the conditions and instruments involved in the data aquisition is available in Berbeco et al. (2005). The data is derived from observations of the position of gold fiducial markers implanted into the tumors of lung cancer patients. The marker position is determined via stereoscopic x-ray imaging conducted at 30 Hz. In each of the two stereoscopic images, the marker position is automatically detected using thresholding and edge detection. The position of the marker in these two images is used to triangulate its position in 3D space relative to the radiation beam. Data consists of tumor positions measured over one or multiple days of radiotherapy treatment delivery (range 1-12), and for multiple sequences on each day, denoted *beams*. In our data set, there are a total of 171 such distinct sequences, with lengths varying from 637 observations (about 21 seconds at 30 observations per second) to 8935 observations (about 5 minutes).

Note that this paper focuses on within-beam forecasting—that is, each beam is treated independently and there is no information sharing between patients or within different beams from the same patient. Developing methodology for combining prediction models from distinct time series (both within and across patients) is an important area for further research.

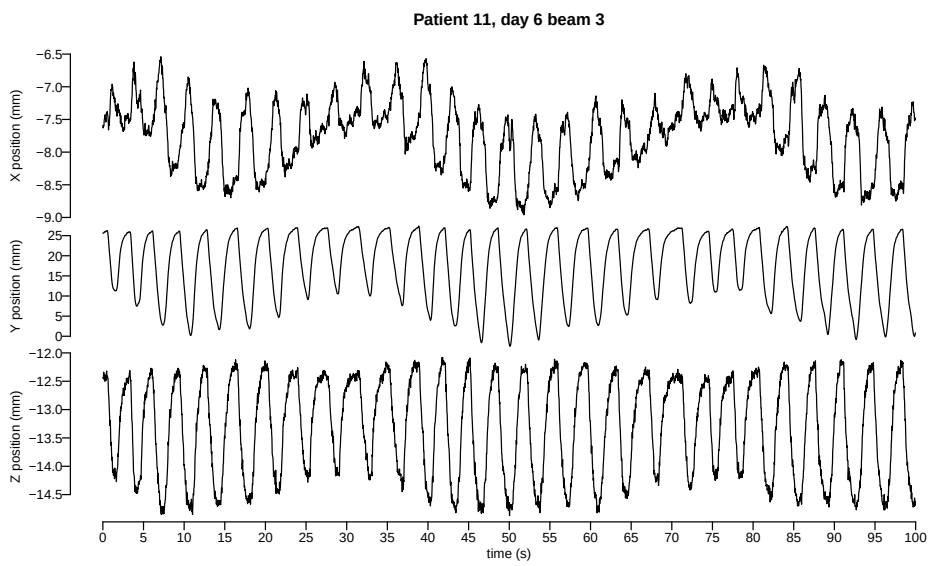


FIG 1. *Sample time series of 3D locations of lung tumor. The X axis is the lateral-medial (left-right) direction, Y axis superior-inferior, and Z axis anterior-posterior*

2.1. *Features of the data.* Each observation in each sequence is a point in $\mathbb{R}^3$, representing the real-time 3D location of the lung tumor. The X axis is the lateral-medial (left-right) direction, the Y axis is superior-inferior, and the Z axis is anterior-posterior, with all measurements in millimeters[1]. Figure 1 shows the motion in each dimension during the first 100 seconds of a particular observation sequence. As expected with respiratory motion, the pattern is approximately periodic, with inhalation closely corresponding to decreasing values in the Y direction. However, the amplitude of each breath varies considerably (in Figure 1 the variation seems periodic, though this is not a typical feature of the data). The curves undergo gradual baseline location shifts, and, while it may not be visually discerned from Figure 1, it is common for respiratory cycles to change periodicity, either sporadically or gradually over time. Table 1 shows the variability in period and amplitude of the respiratory traces, both within and between patients.

Patient	Total beams	Total time (s)	Amplitude (mm)		Period(s)	
			mean	SD	mean	SD
1	4	212.27	14.57	6.98	3.66	1.16
2	2	136.87	13.74	1.84	3.89	1.06
3	2	80.93	9.84	3.16	3.97	0.56
4	38	2502.67	8.86	1.35	2.88	0.31
5	26	2769.33	7.90	1.66	3.61	0.68
6	28	2471.93	10.07	2.51	2.58	0.55
7	11	1661.37	9.66	2.41	5.05	1.09
8	8	832.80	14.38	4.02	3.15	1.18
9	15	2599.90	11.45	1.61	3.09	0.41
10	15	3497.67	14.88	3.65	3.77	0.64
11	22	3674.77	21.81	5.05	3.38	0.52

TABLE 1
Summary statistics for the first principal component of respiratory trace data, at the patient level

Due to the extremely high correlations between series of observations from different dimensions, it is useful to consider a lower-dimensional representation of the 3D process. Transforming each 3D sequence into orthogonal components using principal component analysis (PCA) loads the periodic respiratory dynamics onto the first component, representing about 99% of the total variance in the 3D data. The last two principal components still

[1]The origin is set to the isocenter, which is the center of rotation for the linear accelerator axis motions. During treatment, the patient is positioned so that this coincides with the centroid of the region being treated. However, there is uncertainty in determining this point, so the data is best thought of as relative tumor motion on each day.

exhibit some periodic behavior (see Figure 2), but the signal is weak relative to the noise[2]. In addition to dimension reduction and useful interpretability, the PCA transformation prevents any loss of statistical efficiency if models are fit independently for each component. Ruan and Keall (2010) compared independent-component prediction before and after PCA using kernel density estimation, finding smaller 3D root mean squared prediction error when using the PCA-transformed data for prediction. When comparing several algorithms for predicting lung tumor motion, both Ernst et al. (2013) and Krauss, Nill and Oelfke (2011) used the principal components, then transformed their predictions to the original linear basis of the data.

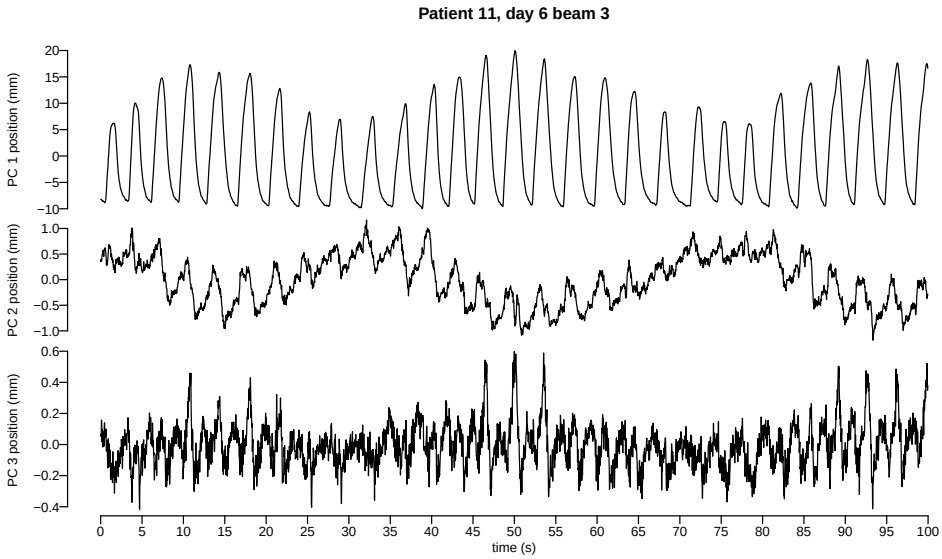


FIG 2. *Time series of principal components. Components 2 and 3 exhibit periodic behavior, but with much smaller magnitude.*

For the remainder of this study, we focus on modeling the first principal component only, as it encodes such a large portion of the system dynamics. In

[2]A referee pointed out that while the first principal component gives the linear combination of the 3D data with maximum variance, it is not necessarily the most *forecastable* linear combination. Alternative linear transformations (e.g., forecastable components [Goerg (2013a)]) may load additional periodic features to the first component than we observe with PCA. In choosing an appropriate transformation, the goal is to find an orthogonal basis in which componentwise predictions have the smallest error when transformed back to the original basis. We do not explore this issue here; however, one advantage in using the first principal component is that the signal-to-noise ratio will be high, allowing for forecast procedures that aren't well suited for measurement error in the observed data.

clinical implementation, we would forecast independently on each orthogonal component and transform back to the original linear basis in order to inform the location of the radiation treatment beam.

2.2. *Time series motifs for forecasting: a graphical example.* Because the data are quasi-periodic, it is useful to look at short patterns that recur at possibly irregular intervals, which we call motifs (we provide a more rigorous definition of time series motifs in Section 3.2). Figure 3 highlights different motifs in the first principal component at the end of the exhale (start of the inhale) for a particular observation sequence. The highlighted areas appear to be heartbeats, which affect the location of the tumor differently depending on the real-time location of the tumor relative to the heart.

Observing repeated patterns within each time series in the data suggests a modeling/prediction framework that leverages this structure. In general, if the recent past of the time series resembles a motif we have observed previously in the data, then the shape of this motif should inform our predictions of future observations; this idea is formalized through predictive state representations [Littman, Sutton and Singh (2002), Shalizi (2003)]. For a graphical illustration, consider predicting 0.4s (12 steps) ahead for the first principal component of the curve displayed in Figure 2. We have observed 100 seconds of the process, and it appears as though we have just observed the start of the exhale; the current observation at time $t = 100$ seconds, as well as the previous 12 observations, are colored orange in Figure 4. Colored

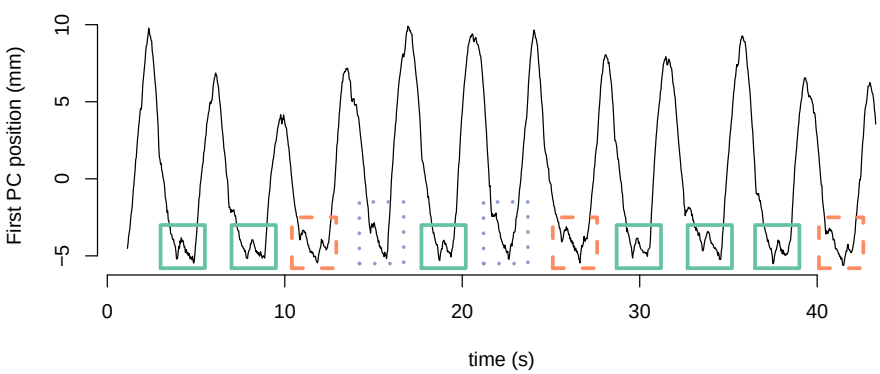


FIG 3. *Recurring patterns (coded by color and line type) in the first principal component of patient 10, day 1, beam 3. Areas boxed by lines of the same color/line type resemble one another. The behavior highlighted in these motifs is most likely caused by the patient's heartbeat.*

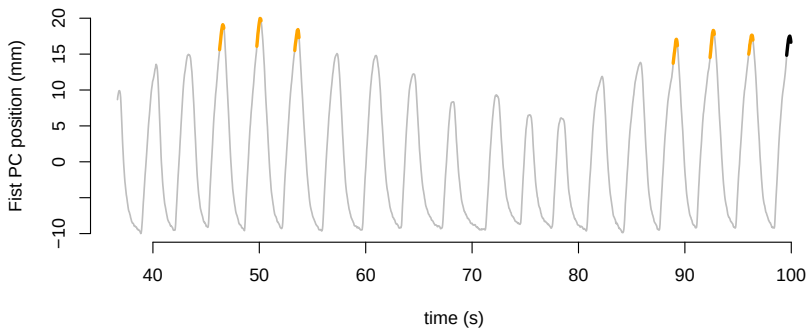


FIG 4. *The most recent 0.43s (13 observations) are in black. The thicker, orange segments share similar local history.*

in black are segments earlier in the time series that resemble the current motif (specifically, we highlighted subseries of length 13 where the tenth point has the largest magnitude, and the 11th–13th points are decreasing).

To predict future observations, we can incorporate the points immediately succeeding the endpoints of black motifs. Figure 5 shows these trajectories (in gray), and the actual current trajectory of the process is shown in orange,

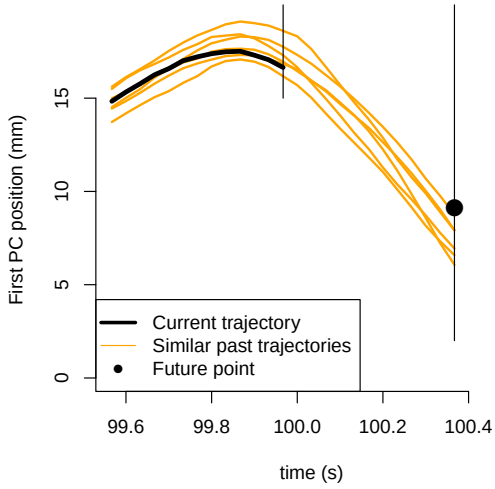


FIG 5. *The recent history of the process (thick black line) instantiates a motif. Previous instances of this motif, and their subsequent evolutions, are in orange and provide reasonable predictions for future points (black dot).*

with a point giving the value 0.4s in the future. The gray curves provide reasonable forecasts for the future evolution of the time series, and indeed the actual future value is close to where these trajectories predict.

Our model, formally introduced in Section 3, implements the forecasting approach sketched in this subsection using an autoregressive model for the data-generating process.

3. Location-mixture autoregressive processes. Here, we define the LMAR process and provide computationally efficient algorithms for parameter estimation and k-step ahead forecasting. To establish terminology, we denote a *time series* as an ordered sequence of real numbers $\{Y_i \in \mathbb{R}, i = 0, \pm 1, \pm 2, \ldots\}$ measured at regular, equally spaced intervals. Also, a *subseries* of length $p+1$ is a subset of a time series $\{Y_i, i = 0, \pm 1, \ldots\}$ comprised of consecutive observations, $Y_i, Y_{i+1}, \ldots, Y_{i+p}$. For notational ease, we will denote subseries as $Y_{i:(i+p)}$, or equivalently $Y_{i+0:p}$.

3.1. *A model for the data-generating process.* Let $\{Y_i, i = -m, \ldots, n\}$ be a time series. Also, assume Σ is a $(p+1) \times (p+1)$ symmetric, nonnegative definite matrix, where Σ_{11} is the upper-left $p \times p$ submatrix, Σ_{22} is the single bottom-right element, and Σ_{21} and Σ_{12} are the respective off-diagonal row and column vectors. p is assumed to be fixed and known. For notational ease, let $\gamma = \Sigma_{11}^{-1}\Sigma_{12}$, $\sigma^2 = \Sigma_{22} - \gamma'\Sigma_{12}$, and $\mathcal{J}_i = \{p+1, \ldots, i+m-p\}$. Lastly, let

$$V_{ij} = \begin{pmatrix} Y_{i-p} - Y_{i-j-p} \\ \vdots \\ Y_{i-2} - Y_{i-j-2} \\ Y_{i-1} - Y_{i-j-1} \end{pmatrix}.$$

As in (1.1), we assume that the distribution of Y_i given $Y_{-m}, \ldots, Y_{i-1}$ is a normal mixture:

$$Y_i | Y_{(-m):(i-1)} \sim \sum_{j \in \mathcal{J}_i} \alpha_{i,j} \, \mathrm{N}(\mu_{i,j}, \sigma^2), \tag{3.1}$$

$$\text{where} \qquad \alpha_{i,j} = \frac{\exp\left(-\frac{1}{2} V_{ij}' \Sigma_{11}^{-1} V_{ij}\right)}{\sum_{l \in \mathcal{J}_i} \exp\left(-\frac{1}{2} V_{il}' \Sigma_{11}^{-1} V_{il}\right)}$$

$$\text{and} \qquad \mu_{i,j} = Y_{i-j} + \gamma' V_{ij}.$$

The model in (3.1) defines the location-mixture autogressive process with parameter Σ (abbreviated LMAR(Σ)). We can recognize the location-mixture

form originally given in (1.1) by writing $\mu_{i,j} = \tilde{\mu}_{i,j} + \sum_{l=1}^{p} \gamma_l Y_{i-l}$ where

$$\tilde{\mu}_{i,j} = Y_{i-j} - \sum_{l=1}^{p} \gamma_l Y_{j-l} \tag{3.2}$$

and $(\gamma_p \ \gamma_{p-1} \ \ldots \ \gamma_1)' = \gamma$. Thus, the distribution for $Y_i | Y_{(-m):(i-1)}$ is a normal mixture with $|\mathcal{J}_i|$ different mean components—each sharing a common autoregressive component but different location parameter—equal variance across components (σ^2), and data-driven mixture weights ($\alpha_{i,j}$). We assume (3.1) for all $i \geq 0$, but we do not make any distributional assumptions about $Y_{(-m):(-1)}$.

As Σ parameterizes the entire mixture distribution, the component means and mixture weights are linked through a common parameter which encourages self-similarity in the data-generating process. If two subseries, $Y_{(i-p):(i-1)}$ and $Y_{(i-p-j):(i-1-j)}$ resemble one another in that $V_{ij}' \Sigma_{11}^{-1} V_{ij}$ is small, then we have a large weight on the mixture component with mean $Y_{i-j} + \gamma' V_{ij}$. This means that the next observation of the process, Y_i, is centered near a previous value of the series Y_{i-j} inasmuch as the subseries of observations preceding Y_i and Y_{i-j} have a similar shape. Simply put, if Y_i and Y_{i-j} are preceded by similar values, then the components of V_{ij} will be close to 0. This drives up the mixture weight $\alpha_{i,j}$, implying the mean of Y_i will be close to $\mu_{i,j}$ (which itself is close to Y_{i-j}).

The dimension of Σ, $p+1$, can in principle be chosen using standard model selection methods (e.g., Bayes factors), though if the goal of fitting a LMAR model is prediction, we recommend cross-validation or hold-out testing for choosing p. For quasi-periodic time series, a reasonable choice for p might be anywhere between one tenth and one third of the average number of observations per period. Larger values of p increase the computational load in estimating Σ while favoring sparser component weights.

The model (3.1) specifies the role of time series motifs in the data-generating process, which was informally discussed in Section 2.2. To illustrate this, we introduce a latent variable M_i that takes values in $\mathcal{J}_i$, such that for all $j \in \mathcal{J}_i$,

$$\mathbb{P}(M_i = j | Y_{(-m):(i-1)}) \propto \exp\left(-\frac{1}{2} V_{ij}' \Sigma_{11}^{-1} V_{ij}\right). \tag{3.3}$$

Then, given $M_i = j$, we induce the same distribution for Y_i as in (3.1) by assuming

$$Y_i \mid [M_i = j, Y_{(-m):(i-1)}] \sim \mathrm{N}(Y_{i-j} + \gamma' V_{ij}, \sigma^2). \tag{3.4}$$

Expression (3.4) can be used to define a motif relation: each subseries of length $(p+1)$ is a *motif*, and $Y_{(i-p):i}$ is an *instance* of motif $Y_{(i-p-j):(i-j)}$ if $M_i = j$ (thus yielding (3.4)). We denote this by writing

$$\text{(motif) } Y_{(i-p-j):(i-j)} \to Y_{(i-p):i} \text{ (instance).}$$

Note that our indexing set $\mathcal{J}_i$ is defined in such a way that instances of a particular motif cannot overlap (share a common component Y_j) with the motif itself.

Our definition of motifs is atypical of the literature for data mining tasks [Lin et al. (2002)] and predictive state representations of time series [Littman, Sutton and Singh (2002)]. For instance, the relationship that instantiates motifs (notated $\to$) is not symmetric, and is not an equivalence relation; for this reason we have defined a motif instance distinctly from a motif. Also, we define motifs as observed subseries of the data and motif instances as latent states (we do not observe M_i). For most data mining tasks, time series motifs represent an equivalence class of observed subseries of the data (possibly transformed) [Fu (2011)] whereas predictive state representations of time series treat motifs as latent equivalence classes of predictive distributions [Shalizi (2003)].

However, our definition of motifs preserves the interpretation of geometric similarity we sketched in Section (2.2). From (3.3), we have $M_i = j$ (meaning $Y_{(i-p-j):(i-j)} \to Y_{(i-p):i}$) with high probability if V_{ij} is small with respect to the Σ_{11}^{-1} inner product norm. Our model thus expects a subseries that is an instance of a particular motif to be close to the motif, and Σ parameterizes this distance metric.

3.2. *Comparison with other mixture autoregressive processes.* We may compare the LMAR(Σ) to a general form of regime-switching autoregressive models, for which we can write the distribution function of Y_i conditional on all available history of the process $Y_{(-m):(i-1)}$ as

$$F(y|Y_{(-m):(i-1)}) = \sum_{j=1}^{d} \alpha_{i,j} \Phi\left(\frac{y - (\beta_{0,j} + \sum_{l=1}^{p} \beta_{l,j} Y_{i-l})}{\sigma_j}\right), \tag{3.5}$$

where $\sum_{j=1}^{d} \alpha_{i,j} = 1$ for all i and Φ denotes the standard normal CDF. Models satisfying (3.5) can be represented in the framework of threshold autoregressive models [Tong (1978), Tong and Lim (1980); see Tong (1990) for a book-length treatment], which represent (3.5) using an indicator series

$\{M_i\}$ taking values on $\{1, \ldots, d\}$, such that

$$Y_i = \beta_{0,M_i} + \sum_{l=1}^{p} \beta_{l,M_i} Y_{i-l} + \sigma_{M_i} \epsilon_i, \tag{3.6}$$

where $\{\epsilon_i\}$ are i.i.d. standard normals. Generally, M is not observed, although there are notable exceptions such as the self-exciting threshold AR model of Tong and Lim (1980).

A canonical model of this form is the mixture autoregressive model of Le, Martin and Raftery (1996) and Wong and Li (2000), which assumes $\{M_i\}$ are i.i.d. and independent of Y. Another special case of (3.6) is when M is a Markov chain, such as in the Markov-switching autoregressive models of Hamilton (1989) and McCulloch and Tsay (1994). More general stochastic structure for M is considered by Lau and So (2008), as well as in mixture-of-experts models in the machine learning literature [Carvalho and Tanner (2005)]. These models seem favorable over the mixture autoregressive models of Wong and Li (2000) when the data is seasonal or quasi-periodic, as is the case with the time series we consider.

The LMAR(Σ) process differs from (3.5) in that the mixture means, following (3.1)–(3.2), are given by

$$\begin{aligned} \mu_{i,j} &= \tilde{\mu}_{i,j} + \sum_{l=1}^{p} \gamma_l Y_{i-l} \\ &= Y_{i-j} + \sum_{l=1}^{p} \gamma_l Y_{i-l} - \sum_{l=1}^{p} \gamma_l Y_{j-l}, \end{aligned}$$

instead of $\mu_{i,j} = \beta_{0,j} + \sum_{l=1}^{p} \beta_{l,j} Y_{i-l}$ as in (3.5). Thus, for LMAR(Σ), the autoregressive coefficients (γ) are fixed, and the normal-mixture form of the conditional distribution is induced by a location shift that is a function of a random subseries of past observations, $\tilde{\mu}_{i,j}$. The normal-mixture form of (3.5), however, is induced by a mixture distribution for autoregressive coefficients of the same lagged values of the time series. The mixture weights of the LMAR(Σ) process are also strongly data-driven, depending on the entire history of the process. Unlike many forms of mixture autoregressive models, there is no prior distribution or conditional dependence structure assumed for M; the distribution of M is supplied entirely by the data.

Another key difference is that LMAR(Σ) does not assume a fixed number of mixture components, as is clear from (3.1). But because the same autoregressive coefficient vector (γ) parameterizes all mean components $\mu_{i,j}$, we actually have a much smaller parameter space than all the instances of (3.5)

cited above, which include the parameters for the mixture components (d vectors of length $p+1$ for the means) as well as for the distribution of M. A small parameter space is advantageous in the context of our data application as it facilitates rapid updating. Also, time constraints will not allow for any goodness-of-fit or model selection procedures for choosing structural parameters such as d or p in (3.5), or structural parameters for M. The only structural parameter in the LMAR(Σ) model is p, and in our analysis of this data set we found that predictive distributions were quite stable for different choices of p.

The most important distinction of the LMAR(Σ) model is the existence of good approximations for k-step ahead predictive distributions, for $k \leq p$, which are given in Section 3.4. Closed-form predictive distributions for $k > 1$ are not available for many models of the form (3.5) (the exception is the Markov-switching autoregressive models of Hamilton (1989); for a discussion see Krolzig (2000)). Wong and Li (2000) recommended Monte Carlo estimates of k-step ahead predictive distributions, although Boshnakov (2009) found for them a closed-form representation as a normal mixture. Calculating the mixture component parameters for moderate k, however, is quite laborious. For the general model (3.5), De Gooijer and Kumar (1992) discussed the difficulty in k-step ahead forecasting and question whether predictive performance is improved over classes of linear time series models (also see Tong and Moeanaddin (1988) for a discussion of the robustness of medium-to-long range forecasts using threshold autoregressive models).

3.3. *Parameter estimation.* In order to be able to adjust radiotherapy treatments in real-time to the patient's breathing pattern, we seek estimation procedures that are fast enough to run online (in less than a few seconds). As a general rule, this favors approximate closed-form solutions to estimating equations over exact numerical or Monte Carlo methods. To estimate Σ, which is the only unknown parameter of this model, we take a conditional likelihood approach based on the conditional distribution $Y_{0:n}|Y_{(-m):(-1)}$. We assume the full-data likelihood can be written as

$$L(\psi, \Sigma) = L_1(\psi, \Sigma) L_2(\Sigma),$$

where $L_1(\psi, \Sigma) \propto \mathbb{P}(Y_{(-m):(-1)}; \psi, \Sigma)$ and $L_2(\Sigma) \propto \mathbb{P}(Y_{0:n}|Y_{(-m):(-1)}; \Sigma)$. The distribution of the first m observations, and thus L_1, is left unspecified, and all information for Σ comes from L_2. If L_1 depends on Σ, there will be some loss of efficiency when using only L_2 for inference versus the complete-data likelihood, though under mild conditions the maximum conditional likelihood estimate is consistent and asymptotically efficient [Kalbfleisch and Sprott (1970)].

The conditional likelihood, $L_2(\Sigma)$, can be written as

$$L_2(\Sigma) = \prod_{i=0}^{n} \frac{1}{\sigma} \left[\sum_{j \in \mathcal{J}_i} \exp\left(-\frac{1}{2\sigma^2}(Y_i - Y_{i-j} - \gamma' V_{ij})^2 \right) \times \left(\frac{\exp(-V_{ij}' \Sigma_{11}^{-1} V_{ij}/2)}{\sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2)} \right) \right]. \tag{3.7}$$

To maximize (3.7), we augment the data to $\{Y_{0:n}, M_{0:n}\}$, with M_i as in (3.3). This invites the use of the Expectation-Maximization (EM) algorithm [Dempster, Laird and Rubin (1977)] to estimate Σ. The augmented-data (complete-data) conditional likelihood is

$$L_{2,\text{com}}(\Sigma) = \prod_{i=0}^{n} \frac{1}{\sigma} \prod_{j \in \mathcal{J}_i} \left[\exp\left(-\frac{1}{2\sigma^2}(Y_i - Y_{i-j} - \gamma' V_{ij})^2 \right) \times \left(\frac{\exp(-V_{ij}' \Sigma_{11}^{-1} V_{ij}/2)}{\sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2)} \right) \right]^{\mathbf{1}[M_i = j]}.$$

This can be simplified further. Let $W_{ij}' = (V_{ij}' \quad Y_i - Y_{i-j})$, and recalling the notation for σ and γ, we have

$$L_{2,\text{com}}(\Sigma) = \prod_{i=0}^{n} \frac{\exp\left(-\frac{1}{2} \sum_{j \in \mathcal{J}_i} \mathbf{1}[M_i = j] W_{ij}' \Sigma^{-1} W_{ij} \right)}{\sigma \sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2)}. \tag{3.8}$$

The term $\sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2)$ can be viewed as an approximation of a Gaussian integral; if we assume that, for all i, $\{V_{il}, l \in \mathcal{J}_i\}$ resemble $|\mathcal{J}_i|$ i.i.d. draws from some distribution $V \sim \mathrm{N}(0, \Omega)$, then we have

$$\begin{aligned} \sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2) &\approx |\mathcal{J}_i| \int \exp(-V' \Sigma_{11}^{-1} V/2) \frac{\exp(-V' \Omega^{-1} V/2)}{(2\pi)^{p/2} |\Omega|^{1/2}} dV \\ &= |\mathcal{J}_i| \left(\frac{|(\Sigma_{11}^{-1} + \Omega^{-1})^{-1}|}{|\Omega|} \right)^{1/2} \\ &= |\mathcal{J}_i| \left(\frac{|\Sigma_{11}|}{|\Sigma_{11} + \Omega|} \right)^{1/2}. \end{aligned} \tag{3.9}$$

Noting that $\sigma |\Sigma_{11}|^{1/2} = |\Sigma|^{1/2}$, and ignoring multiplicative constants, we arrive at an approximate augmented-data conditional likelihood:

$$L_{2,\text{com}}(\Sigma) \approx \left(\frac{|\Sigma_{11} + \Omega|}{|\Sigma|} \right)^{(n+1)/2} \exp\left(-\frac{1}{2} \sum_{i=0}^{n} \sum_{j \in \mathcal{J}_i} \mathbf{1}[M_i = j] W_{ij}' \Sigma^{-1} W_{ij} \right).$$

Typically $\Sigma_{11} \ll \Omega$, meaning

$$\begin{aligned}\partial\left(\log(|\Sigma_{11}+\Omega|)-\log(|\Sigma|)\right) &= \mathrm{Tr}((\Sigma_{11}+\Omega)^{-1}\partial\Sigma_{11})-\mathrm{Tr}(\Sigma^{-1}\partial\Sigma)\\ &\approx -\mathrm{Tr}(\Sigma^{-1}\partial\Sigma)\end{aligned}$$

as $\partial\log(|\Sigma|)$ dominates $\partial\log(|\Sigma_{11}+\Omega|)$. This justifies the approximation $\log(|\Sigma_{11}+\Omega|)-\log(|\Sigma|)\approx-\log(|\Sigma|)$ in the augmented-data conditional log-likelihood, as it will admit nearly the same maximizer. Thus, we have

$$\log(L_{2,\mathrm{com}}(\Sigma))\approx-\frac{n+1}{2}\log(|\Sigma|)-\frac{1}{2}\sum_{i=0}^{n}\sum_{j\in\mathcal{J}_i}\mathbf{1}[M_i=j]W_{ij}'\Sigma^{-1}W_{ij}. \tag{3.10}$$

While (3.10) is much easier to work with than the logarithm of the exact conditional likelihood (3.8), the assumptions of this approximation are somewhat tenuous. Under this model (3.1), both conditional and marginal distributions of observations at each time point follow a normal mixture, meaning for l randomly chosen from $\mathcal{J}_i$, we have a difference of normal mixtures (itself a normal mixture) for V_{il}, instead of i.i.d. normals as (3.9) suggests. We nevertheless proceed with approximation (3.10) in place of (3.8), noting that convergence of the EM algorithm needs to be more carefully monitored in this instance.

At each iteration of the EM algorithm, we maximize the so-called Q function:

$$\begin{aligned}Q^{(t)}(\Sigma) &= \mathbb{E}_{\Sigma^{(t)}}[\log(L_{2,\mathrm{com}}(\Sigma))|Y]\\ &\approx -\frac{n+1}{2}\log(|\Sigma|)-\frac{1}{2}\sum_{i=0}^{n}\sum_{j\in\mathcal{J}_i}\omega_{ij}W_{ij}'\Sigma^{-1}W_{ij},\end{aligned} \tag{3.11}$$

with $\Sigma^{(t)}=\mathrm{argmax}(Q^{(t-1)}(\Sigma))$ and $\omega_{ij}=\mathbb{E}_{\Sigma^{(t)}}[\mathbf{1}[M_i=j]|Y]$. Clearly,

$$\omega_{ij}=\frac{\exp(-W_{ij}'[\Sigma^{(t)}]^{-1}W_{ij}/2)}{\sum_{l\in\mathcal{J}_i}\exp(-W_{lj}'[\Sigma^{(t)}]^{-1}W_{lj}/2)}.$$

The maximizer of (3.11) can be found in closed form as a weighted sample covariance matrix,

$$\Sigma^{(t+1)}=\frac{1}{n+1}\sum_{i=0}^{n}\sum_{j\in\mathcal{J}_i}\omega_{ij}W_{ij}W_{ij}'. \tag{3.12}$$

Again, due to several different approximations used in maximizing the original conditional likelihood (3.7), it is necessary to monitor the convergence to a suitable (if slightly sub-optimal) solution, as the log-likelihood is not guaranteed to increase at each iteration.

3.4. *A prediction model for fast implementation.* Exact closed-form expressions for k-step ahead predictive distributions are not available for the model (3.1). Because of the need for real-time forecasting of many steps ahead, we explore approximations to k-step ahead predictive distributions that are available in closed form. An immediate approach to doing so is to explore whether the approximate complete-data conditional log-likelihood used for inference (3.10) corresponds to a probabilistic model (perhaps misspecified) that admits closed-form predictive distributions. In other words, if the previous section derives an approximate log-likelihood (3.10) from an exact model (3.1), here we treat (3.10) as exact and explore corresponding approximate models.

Let $Z_i = (Y_{i-p} \ \dots \ Y_{i-1} \ \ Y_i)'$ for $0 \le i \le n$. Since $W_{ij} = Z_i - Z_j$, we may arrive at the likelihood expression (3.10) by assuming $Z_i \sim \mathrm{N}(Z_{i-M_i}, \Sigma)$ independently. This is obviously a misspecification, since for any $k \le p$, Z_i and Z_{i+k} contain duplicate entries and thus cannot be independent. But assuming the $\{Z_i\}$ independent, and further assuming $\mathbb{P}(M_i = j) = 1/|\mathcal{J}_i|$ independently for all i, we can write the (conditional) likelihood for a independent multivariate normal mixture model, denoted L_a to distinguish from $L_{2,\mathrm{com}}$:

$$L_a(\Sigma) = \prod_{i=0}^{n} \prod_{j \in \mathcal{J}_i} \left[|\Sigma|^{-1/2} \exp\left(-\frac{1}{2} W_{ij}' \Sigma^{-1} W_{ij} \right) \right]^{\mathbf{1}[M_i = j]}. \tag{3.13}$$

Indeed, we see that $L_a(\Sigma)$ is equal to the approximation of $L_{2,\mathrm{com}}(\Sigma)$ given in (3.10). Thus, the misspecified independent mixture model for Z_i yields the same likelihood (L_a) as the approximation to L_2, the exact (conditional) likelihood corresponding to the data-generating process. Also, recall that $M_i = j$ denotes Z_i as an instance of motif Z_j. The implied relation in (3.13) is that

$$Z_j \to Z_i \text{ if } Z_i | Z_j \sim \mathrm{N}(Z_j, \Sigma), \tag{3.14}$$

and indeed this relation is closely connected to the one defined in (3.4). They appear equivalent, as (3.4) is recovered by assuming $Z_i | Z_j \sim \mathrm{N}(Z_j, \Sigma)$, and then considering the conditional distribution $Y_i | Y_{(-m):(i-1)}$. However, for (3.14) to hold for all i requires the impossible assumption of Z_i being independent of Z_{i-1}, while the relation in (3.4) does not.

The corresponding Q function for this complete-data conditional likelihood (3.13) is

$$Q_a^{(t)}(\Sigma) = \sum_{i=0}^{n} -\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{j \in \mathcal{J}_i} \mathbb{E}_{\Sigma^{(t)}}[\mathbf{1}[M_i = j] | Z] W_{ij}' \Sigma^{-1} W_{ij}.$$

Working $\mathbb{E}_{\Sigma^{(t)}}[\mathbf{1}[M_i = j]|Z] = \omega_{ij}$, we see that $Q_a^{(t)}$ is identical to $Q^{(t)}$ given in (3.11), confirming that the "same" Σ parametrizes both the original data-generating process assumed in (3.1), and its degenerate approximation that we will use to make predictions (3.13). We may also think of maximizing Q as inferring motif instances given by the relation (3.14), i.e., minimizing a distance metric.

The independent multivariate mixture distribution of $\{Z_i\}$ considered here very easily provides k-step predictive distributions for $k \leq p$. If we have observed the process up to Y_n and wish to predict Y_{n+k}, this is equivalent to having observed Z up to Z_n and wishing to predict the last component of Z_{n+k}. Having observed Z_n completely, we have observed the first $p-k+1$ components of Z_{n+k}, and thus by the (misspecfied) independence assumed for $\{Z_i\}$, the predictive distribution for Y_{n+k} depends only on these $p-k+1$ values. To write this, we denote $\tilde{Z}_n^k$ as the first $p-k+1$ components of Z_{n+k} (or the last $p-k+1$ components of Z_n); also let $\tilde{W}_{nj}^k = \tilde{Z}_n^k - \tilde{Z}_j^k$ and partition Σ into Σ_{11}^k as the upper-left $(p-k+1) \times (p-k+1)$ submatrix, Σ_{22}^k as the single bottom-right element (thus identical to Σ_{22}), and $\Sigma_{12}^k, \Sigma_{21}^k$ accordingly. Then we have

$$Y_{n+k}|Y_{(-m):n} \sim \sum_{j \in \mathcal{J}_{n+k}} \alpha_j^k \, \mathrm{N}(\mu_j^k, \sigma_k^2), \tag{3.15}$$

where

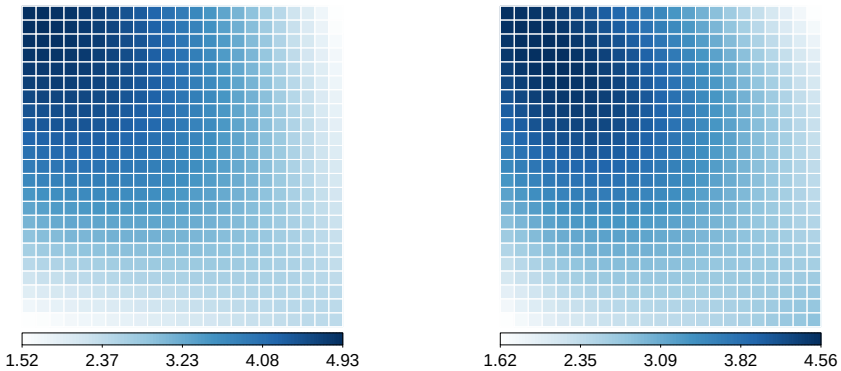
- $\alpha_j^k = \mathbb{P}(M_{n+k} = j|\tilde{Z}_n^k) \propto \exp(-(\tilde{W}_{nj}^k)'[\Sigma_{11}^k]^{-1}\tilde{W}_{nj}^k/2)$,
- $\mu_j^k = Y_{n+k-j} + \Sigma_{21}^k[\Sigma_{11}^k]^{-1}\tilde{W}_{nj}^k$,
- $\sigma_k^2 = \Sigma_{22}^k - \Sigma_{21}^k[\Sigma_{11}^k]^{-1}\Sigma_{12}^k$.

In terms of motifs, these predictive distributions result from considering the most recent subseries of the data of length $p-k+1$ as a partially observed motif instance, Z_{n+k}, which includes the future observation we wish to predict, Y_{n+k}. Using the implied motif relation in (3.14), we infer both the motif for which Z_{n+k} is an instance, and derive predictive distributions using simple multivariate normal properties (3.15).

Of course, we use $\hat{\Sigma}$, the solution to (3.12), in place of Σ in the above expressions, acknowledging that the resulting predictive distributions fail to account for the uncertainty in our estimate of Σ.

3.5. *Interpreting* $\hat{\Sigma}$. Figure 6 shows estimates $\hat{\Sigma}$ from two of the time series in our data. Interpreting these as covariance matrices, we see relatively high correlations across components, favoring instantiating motifs where the difference between the motif instance and the original motif is roughly linear

with a slope near 0. Also, the diagonal terms are decreasing from top to bottom, implying that more weight is given to the most recent components of the observed time series when inferring the latent motif instance and making predictions.



(A) $\hat{\Sigma}$ *for patient 10, day 1, beam 1.* (B) $\hat{\Sigma}$ *for patient 9, day 1, beam 2.*

FIG 6. *Illustration of* $\hat{\Sigma}$ *for two of the time series in our data, using* $p = 22$. *Note that the color scale differs slightly for each figure.*

4. Evaluating out-of-sample prediction error with competing methods. We compare out-of-sample prediction performance for tumor tracking using the LMAR(Σ) model with three methods that are straightforward to implement, provide real-time forecasts. Neural networks (4.1) and ridge regression (4.2) both compare favorably to alternative methods with regards to prediction accuracy [Sharp et al. (2004), Krauss, Nill and Oelfke (2011)]. LICORS (4.3) is a nonparametric and non-regression forecasting method based on predictive state representations of the time series [Goerg and Shalizi (2012; 2013)]. For each method, Sections 4.4–4.6 discuss data preprocessing and computational considerations relevant for real-time tumor tracking.

4.1. *Feedforward neural networks.* Multilayer feedforward neural networks with at least one hidden layer have been used to forecast lung tumor motion by Murphy, Isaakson and Jalden (2002) and Murphy and Dieterich (2006), as well as in simultaneous comparisons of several methods [Sharp et al. (2004), Krauss, Nill and Oelfke (2011), Ernst et al. (2013)]. Using $p \times h \times 1$ neural networks, we can predict Y_{i+k} as a function of $Y_{(-m):i}$. Let $X_i = Y_{(i-p)+1:p}$, then

$$\hat{Y}_{i+k} = \beta_0 + \beta' G(X_i), \tag{4.1}$$

where $G(X_i) = (g(w_{01} + w_1' X_i)\ g(w_{02} + w_2' X_i)\ \ldots\ g(w_{0h} + w_h' X_i))'$ with activation function g; here we assume $g(x) = 1/(1 + \exp(-x))$. Hyperparameters p and h are set by the user (as is the form of the activation function). Unknown parameters $\beta_0, \beta, w_{01}, \ldots, w_{0h}, w_1, \ldots, w_h$ are estimated by minimizing sum of squares using the `R` package `nnet` [Venables and Ripley (2002)]. Because the number of unknown parameters is large ($w_1, \ldots, w_h$ are p-vectors), to prevent overfitting, a regularization term is often used in the sum of squares minimization. Then, the model is fit by minimizing

$$C(Y, \theta) = \sum_{i=0}^{n-k} (\hat{Y}_{i+k} - Y_{i+k})^2 + \lambda \theta' \theta, \tag{4.2}$$

where θ represents a vector of all unknown parameters stacked together, and λ is a penalty hyperparameter that is supplied by the user, with higher values providing more shrinkage.

4.2. *Ridge regression.* The second competing method considered is a linear predictor of the form

$$\hat{Y}_{i+k} = \beta_0 + \beta' X_i, \tag{4.3}$$

with $X_i = Y_{(i-p)+1:p}$ and where β_0, β are found by minimizing

$$C(Y, \beta_0, \beta) = \sum_{i=0}^{n-k} (\hat{Y}_{i+k} - Y_{i+k})^2 + \lambda(\beta_0^2 + \beta' \beta). \tag{4.4}$$

Nearly all studies involving forecasting lung tumor motion consider predictors of this form, usually referred to as ridge regression. However, since ridge regression assumes $\{Y_i\}$ to be independent [Hoerl and Kennard (1970)], the model implied by (4.3)–(4.4) is better described as fitting an autoregressive model of order $p + k - 1$ (the first $k - 1$ coefficients being 0) using conditional least squares, with an L_2 penalty on the vector of autoregressive coefficients (yet we shall refer to this prediction method as ridge regression). Linear models lack many features that seem appropriate for this forecasting example, such as multimodal and/or heteroskedastic conditional distributions, yet still perform reasonably well and are commonly used as a baseline for comparing tumor prediction methods.

4.3. *Light Cone Reconstruction of States (LICORS).* Mixed LICORS [Goerg and Shalizi (2013)] is a recent nonparametric forecasting method based on predictive state representations of spatiotemporal fields [Shalizi

(2003), Goerg and Shalizi (2012)]. In the context of our forecasting example, mixed LICORS models $Y_{i+k}|Y_{(-m):i}$ as depending only on the *past light cone* (with horizon p) $X_i = Y_{(i-p)+1:p}$; furthermore, $\epsilon(X_i)$ is a minimal sufficient statistic for the predictive distribution of Y_{i+k}, so that

$$Y_{i+k}|Y_{(-m):i} \sim Y_{i+k}|X_i \sim Y_{i+k}|\epsilon(X_i), \tag{4.5}$$

and if $\epsilon(X_i) = \epsilon(X_j)$, then $Y_{i+k}|\epsilon(X_i) \sim Y_{j+k}|\epsilon(X_j)$. Without loss of generality, we may assume ϵ takes values in $\mathcal{S} = \{s_1, \ldots, s_K\}$, and for simpler notation let $S_i = \epsilon(X_i)$ and denote $\mathbb{P}_j(Y_{i+k}) = \mathbb{P}(Y_{i+k}|S_i = s_j)$. The unknown parameters of this model are the mapping ϵ, the number of predictive states K, and the predictive distributions of the predictive states $\{\mathbb{P}_j, 1 \leq j \leq K\}$. For fixed K, the remaining parameters are estimated by maximizing

$$C(Y, \epsilon, \mathbb{P}_1, \ldots, \mathbb{P}_K) = \prod_{i=0}^{n-k} \sum_{j=1}^{K} \mathbb{P}_j(Y_{i+k})\mathbb{P}(S_i = j|X_i), \tag{4.6}$$

which acts as a likelihood, except for $\mathbb{P}_j$ being unknown. Goerg and Shalizi (2013) maximized (4.6) with a nonparametric variant of the EM algorithm, using weighted kernel density estimators to approximate the unknown densities of the predictive distributions $\{\mathbb{P}_j, 1 \leq j \leq K\}$; they also advocated data-driven procedures for choosing the number of predictive states K.

It is possible to embed the LMAR model in a parametric (Gaussian) mixed LICORS framework, treating $\{V_{ij}, j \in \mathcal{J}_i\}$ as the past light cone ℓ_i^- and $\{V_{ij}$ where $M_i = j\}$ as the predictive state $S_i = \epsilon(\ell_i)$. While this choice of ϵ does provide a minimal sufficient statistic for the predictive distribution of Y_i (or L_i^+) under the LMAR model, it will not provide any dimension reduction or parsimony since $\epsilon(\ell_i)$ will almost surely be unique for each i under our model assumptions.

Fitting the mixed LICORS model to the time series in our data and using it for forecasting was accomplished using the `R` package `LICORS` [Goerg (2013b)]. Note that point forecasts using the inferred model (4.5) will be a weighted average of the means of the predictive states $s_i \in \mathcal{S}$.

4.4. *Data preprocessing.* Similar to Krauss, Nill and Oelfke (2011), we use a total of 80 seconds of data (2400 observations) from each time series, 40 seconds for model fitting, and 40 seconds for out-of-sample prediction given the model fit to the first 40 seconds of data. This necessitates removing time series for which we have fewer than $2400 + k$ observations, where k is the forecast window. This eliminates 61 of the 171 time series in our data base, unfortunately including all time series from patients 1, 2, and

Method	Hyperparameter	Description
LMAR	p	Motif length (3.14)
Neural Networks	p	Length of input vector X_i (4.1)
	h	Number of neurons in hidden layer (4.1)
	λ	Shrinkage; L2 penalty (4.2)
Ridge Regression	p	Length of input vector X_i (4.3)
	λ	Shrinkage; L2 penalty (4.4)
Mixed LICORS	p	Length of input vector X_i (4.5)

TABLE 2

List of global, patient-independent hyperparameters to be tuned for each prediction method

3. An additional 15 time series were eliminated because there were several gaps in the observation sequence. This leaves us with 95 total time series; patient 8 has only one time series, and patient 6 has the next fewest series with 9. Patient 11 has the most time series with 21. While each time series is three-dimensional, we predict using only the first principal component (the principal component transformation is estimated from the initial 40s of training data) as discussed in Section 2.1.

4.5. *Tuning hyperparameters.* Because of the need for real-time model fitting and prediction, all tuning and hyperparameters for the methods we consider must be specified prior to the administration of radiotherapy—before any data is observed. This suggests finding specifications for each model that perform reasonably well for all patients, though perhaps sub-optimally for each patient individually. Indeed, this is the approach usually taken in the literature [Sharp et al. (2004), Krauss, Nill and Oelfke (2011), Ernst et al. (2013)]. Because patients are typically given several or many instances of radiotherapy during different sessions, there seems to be potential for more patient-specific tuning of hyperparameters, though this is left as a separate problem for now.

Table 2 lists the hyperparameters and/or tuning parameters for each of the prediction methods we consider. As described in Section 4.4, since the first 40 seconds of each time series will not be used to evaluate out-of-sample prediction, we may use these subseries to find sensible, patient-independent values for all hyperparameters. Each 40 second subseries is further divided, where for a given set of hyperparameters each prediction method is fit to the first 30 seconds of data (900 observations), and then the remaining 10 seconds are used to generate out-of-sample predictions, for which we store the vector of errors.

Using a course grid search over the parameter space given in Table 2,

predictive error (both root mean squared error (RMSE), as well as median absolute error (MAE), which is more robust to heavy-tailed error distrbutions), is averaged across patients, allowing us to choose the best set of patient-independent hyperparameter values [Krauss, Nill and Oelfke (2011)]. Note that different hyperparameter values are chosen for different forecast windows.

4.6. *Computational considerations.* In addition to providing real-time forecasts, tumor tracking models require parameters that can be estimated very quickly so that accurate (forecast-assisted) radiotherapy can begin as soon as possible after observing a short window of training data.

Ridge regression yields almost instantaneous estimates of parameters necessary for prediction (β in (4.3)), since (4.4) can be minimized in closed form. Fitting neural networks (4.1), however, requires numerical optimization of (4.2). This was carried out using the `nnet` package in `R`, which implements the BFGS algorithm [Venables and Ripley (2002)]. Because (4.2) is not convex, we recommend several random starting points for initiating the optimization, insomuch as time allows; the dimension of the parameter space, as well as convergence criteria for the numerical optimization, are both extremely important considerations in addition to the length of the time series being fit. For example, on a Lenovo X220 laptop with an Intel Core i5-2520M 2.50 Ghz processor, a $45 \times 6 \times 1$ neural network required about 10 seconds to fit on 1200 observations when using `nnet`'s default convergence criteria, with 10 randomly initialized starting points.

The computation time in fitting the LMAR(Σ) depends critically on both the convergence criteria for the EM algorithm, as well as the initial value of Σ used. Typically, the likelihood (3.7) or log-likelihood is used; however, the EM updates given in (3.12) are only approximate, meaning the likelihood is not guaranteed to increase at every iteration. We found that using the approximate log-likelihood (3.10) to check convergence yielded convergence in the exact log-likelihood. This being the case, other metrics could possibly be used to check convergence that are quicker to calculate than (3.10), such as the Frobenius norm of differences in the updates of $\hat{\Sigma}$. To obtain good starting values, the algorithm can be run before having observed the entire training sequence, using a simple starting value of a diagonal matrix. Using a relative tolerance of 0.0001 for the approximate log-likelihood, we were able to compute $\hat{\Sigma}$ in no more than four seconds for each of the time series considered.

The value of m for the LMAR model may also trade off estimation speed and accuracy; we used $m = 400$, though found essentially identical results

for $m = 200$ and $m = 300$ (higher values of m favor faster, but less precise, estimation of Σ).

Parameter estimation for mixed LICORS took several minutes on our machine. However, much of this computational cost is accrued in inferring K, the number of predictive states. The procedure described in Goerg and Shalizi (2013) and implemented in the `LICORS R` package is to start at an upper bound for the number of predictive states, optimize the likelihood approximation (4.6), and then merge the two states whose predictive distributions are closest (measured by some distance or a hypothesis test). The optimizing and merging steps are repeated until we either have 1 state remaining, or alternatively all pairwise tests for equality among predictive distributions are rejected. Then, cross validation is used to choose among these candidate models indexed by different values of K.

While there may be some loss in prediction accuracy, estimation speed can be improved by fixing K (perhaps tuning it as in Section 4.5). Furthermore, initializing the nonparametric EM algorithm with informative starting values (learned from previously observed respiratory trace curves) and relaxing the convergence criteria may substantially increase estimation speed with little loss in predictive performance.

5. Prediction results for tumor tracking data. The results of out-of-sample predictions using the LMAR model, as well as the methods discussed in Section 4, are provided in this section. Point forecasts are discussed in Sections 5.1–5.3 and interval/distributional forecasts in Section 5.4.

5.1. *Results for point forecasts.* The measures of predictive performance we consider are root mean squared error (RMSE) and median absolute error (MAE), as well as the proportion of time each forecasting method obtains the minimum prediction error among the methods compared. We report these quantities averaged over each of the 8 patients, at forecast windows of 0.2s (6 observations), 0.4s (12 observations), and 0.8s (18 observations) in Table 3.

We stress that RMSE may not be the most useful summary of predictive performance since the error distributions are heavy-tailed, and in the application of radiotherapy, we are more concerned with whether or not the treatment beam was localized to the tumor than with the squared distance of the treatment beam to the tumor[3]. For this reason, we feel that the median (more generally, quantiles of the distribution function for absolute

[3]However, the loss function implied in the model fitting and point prediction is squared error loss, which is the simplest for many computation reasons.

errors) is the best summary of predictive performance for this data context. Ultimately, the dosimetric effects of these errors are of most interest, but their determination is complicated, and beyond the scope of this work.

Two further points of emphasis regarding the accuracy summaries are that while we eliminated time series with unevenly spaced observations from consideration, we still have quite a few time series with unusual motion in our data base. Without actually observing the patient, we are not sure whether observed deviations from normal breathing are caused by exogenous factors or are instances of relevant components of the data-generating process, such as coughs, yawns, deep breaths, etc. The other point is that there is a lot of disparity in the measures of predictive performance within the literature on this subject; in addition to working with different data sets, obtained from differing equipment, some authors account for the between-patient variation in respiratory dynamics by scaling or normalizing all curves, or by comparing errors from a prediction method against errors from making no prediction and just using the lagged value of the series. When using evaluation procedures of Krauss, Nill and Oelfke (2011) and Murphy and Dieterich (2006), we produced very similar results with ridge regression and linear models. However, the error summaries we present here, in comparison with the LMAR model, are not directly comparable to these results.

5.2. *Quantitative summaries of point forecasts.* Summarizing Table 3, we see that ridge regression is actually sub-optimal in all accuracy measures for all patients and forecast windows. The LMAR model strongly outperforms the other three methods for all forecast windows for patients 6, 7, 9, 10, and 11; neither neural networks nor LICORS appear to be optimal for any patient across all forecast windows, although neural networks perform well for patients 4, 5, and 8, while LICORS predicts well for patients 4, 8, and 11. Between-patient differences prevent any particular forecasting method from dominating other methods across patients, but the LMAR model seems to offer the most accurate overall point forecasts given these results.

5.3. *Qualitative summaries of point forecasts.* When looking at the predicted time series for each method used, the general pattern we observe is that LMAR outperforms the other three methods when the data undergo changes in shape, period, or amplitude—or more generally, when the test data do not resemble the training data. Figure 7 shows one (atypically dramatic) instance of such behavior. The top curve is the first 40 seconds of the time series, on which all prediction methods were trained. The next four curves give the predicted time series at a window of 0.2s for LMAR (red), NN (blue), ridge regression (green), and LICORS (purple). It is clear from

		0.2s forecast			0.4s forecast			0.6s forecast		
Pat.	Method	RMSE	MAE	Best	RMSE	MAE	Best	RMSE	MAE	Best
4	LMAR	0.52	0.24	0.27	0.99	0.39	0.27	**1.18**	**0.44**	**0.31**
	NNs	**0.46**	**0.22**	**0.28**	**0.90**	0.39	**0.28**	1.20	0.48	0.27
	Ridge	0.53	0.31	0.20	1.08	0.62	0.17	1.50	0.86	0.18
	LICORS	0.58	0.25	0.25	1.05	**0.37**	**0.28**	1.43	0.52	0.24
5	LMAR	0.56	**0.25**	**0.30**	0.96	0.42	0.29	**1.15**	**0.51**	**0.30**
	NNs	**0.55**	0.27	0.27	**0.89**	**0.40**	**0.30**	**1.15**	**0.51**	**0.30**
	Ridge	0.58	0.31	0.25	1.01	0.56	0.23	1.39	0.78	0.23
	LICORS	0.79	0.35	0.19	1.33	0.63	0.18	1.79	0.89	0.17
6	LMAR	**0.77**	**0.40**	**0.29**	**1.54**	**0.82**	**0.30**	**2.00**	**1.06**	**0.34**
	NNs	1.01	0.46	0.24	1.74	0.93	0.24	2.43	1.38	0.22
	Ridge	0.83	0.42	0.28	1.59	0.88	0.28	2.14	1.28	0.28
	LICORS	1.37	0.57	0.19	2.17	1.19	0.18	2.92	1.75	0.15
7	LMAR	**0.40**	**0.15**	**0.35**	**0.85**	**0.27**	**0.37**	**1.23**	**0.41**	**0.36**
	NNs	0.43	0.19	0.26	0.88	0.36	0.25	1.35	0.51	0.25
	Ridge	0.44	0.26	0.20	1.00	0.59	0.16	1.56	0.96	0.17
	LICORS	0.62	0.25	0.20	1.05	0.41	0.21	1.56	0.56	0.23
8	LMAR	1.27	**0.62**	**0.27**	**2.63**	1.46	0.26	3.57	2.00	0.24
	NNs	**1.26**	0.68	**0.27**	2.71	**1.27**	**0.28**	**3.46**	**1.76**	**0.29**
	Ridge	1.44	0.69	0.20	2.86	1.54	0.19	4.11	2.26	0.19
	LICORS	1.50	0.64	0.26	2.89	1.33	**0.28**	3.70	**1.76**	0.28
9	LMAR	**0.58**	**0.22**	**0.39**	**1.29**	**0.52**	**0.35**	**2.03**	**0.90**	**0.30**
	NNs	0.73	0.32	0.24	1.69	0.64	0.26	2.45	0.92	0.24
	Ridge	0.81	0.34	0.22	1.68	0.73	0.22	2.42	0.98	0.25
	LICORS	1.35	0.53	0.15	2.20	0.98	0.17	2.64	1.19	0.20
10	LMAR	**0.88**	**0.36**	**0.34**	**1.73**	**0.77**	**0.33**	**2.55**	**1.19**	**0.30**
	NNs	1.09	0.44	0.25	2.16	0.93	0.24	2.98	1.35	0.24
	Ridge	0.95	0.45	0.24	1.84	0.94	0.24	2.67	1.41	0.26
	LICORS	1.62	0.61	0.17	2.20	1.10	0.19	3.25	1.56	0.20
11	LMAR	**1.13**	**0.44**	**0.32**	**2.59**	**1.06**	**0.29**	**3.70**	**1.49**	**0.31**
	NNs	1.24	0.50	0.25	2.95	1.19	0.24	3.99	1.70	0.23
	Ridge	1.19	0.63	0.22	2.69	1.51	0.21	3.99	2.40	0.21
	LICORS	1.64	0.57	0.21	3.04	1.09	0.26	4.21	1.65	0.25

TABLE 3

Summary of errors in point forecasts for all four methods and all three forecast windows considered. RMSE is root mean squared error, MAE is median absolute error, and Best *refers to the proportion of time for which the absolute prediction error is smallest among the methods considered. For each metric, the most desirable value among the four methods for each patient/forecast window combination is in* **bold.**

the figure that the end of the training period for this time series coincided with a dramatic change in the patient's respiration.

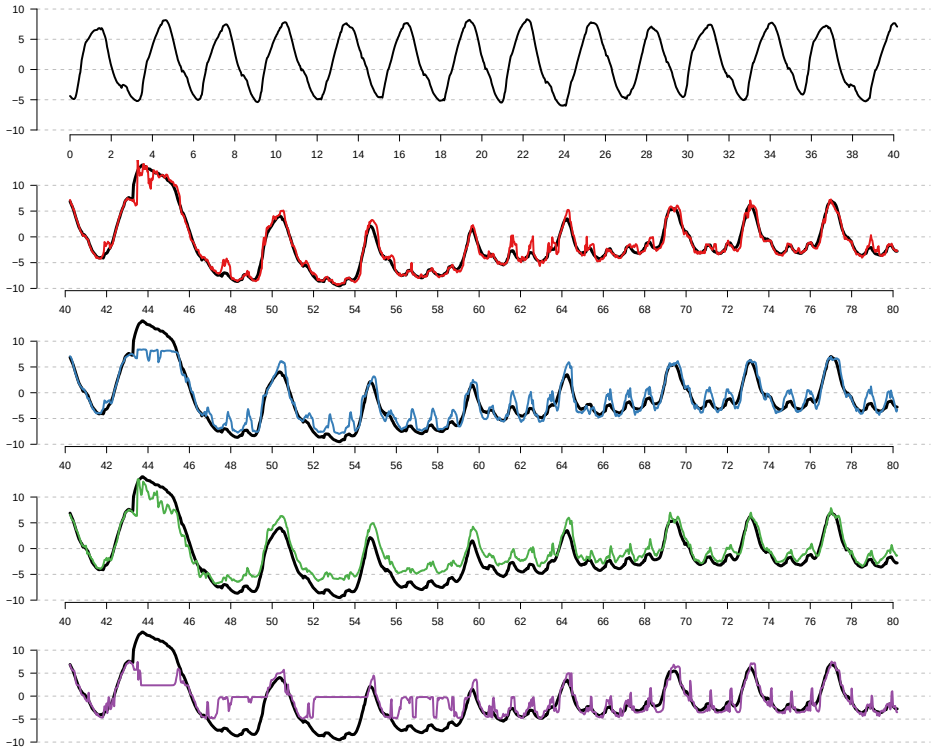


FIG 7. *Predictions for patient 9, day 3 beam 6 with a forecast window of 0.2s. Location (mm) is the y axis and time (s) the x axis. The 40s training sequence is top, with predictions for the next 40s from LMAR in red, NN in blue, ridge regression in green, and LICORS in purple.*

Both neural networks and LICORS suffer from the range of the curve being larger (dropping below -5mm and exceeding 10mm) after the training period; for both methods, the training data bounds the range of point forecasts, regardless of the input vector for future test cases. For LICORS, when the test data is below the minimum of the training data (-5mm), the single predictive state associated with the minimal values of the training data will dominate, leading to brief periods of static forecasts. With this time series, this particular predictive state represents an abrupt transition between sharp exhale and sharp inhale. Thus the forecasts for the test data are dramatic over-estimates throughout the "U" shaped motifs starting around $t = 47$, where the patient does not actually fully inhale.

Ridge regression seems to accurately predict the magnitudes of increases and decreases, yet the predictions are off by a nearly constant factor for $t \in (48, 68)$. In the context of the ridge regression model (4.3), this suggests that β is correctly specified, but perhaps β_0 is time-varying. The LMAR

model includes an autoregressive term for the most recent p observations in its forecast, and thus, like ridge regression, accurately predicts rates of change in the time series. Moreover, the stochastic location-mixture component in the LMAR prediction adjusts predictions for gradual magnitude shifts in the data.

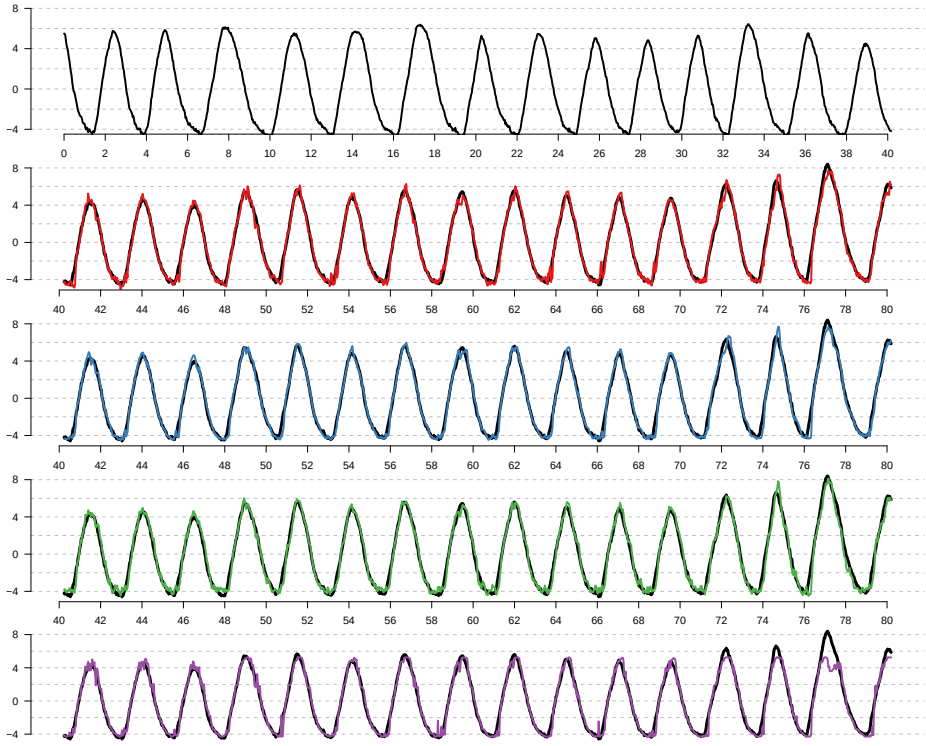


FIG 8. *Predictions for patient 4, day 6 beam 1 with a forecast window of 0.2s. Location (mm) is the y axis and time (s) the x axis. The 40s training sequence is top, with predictions for the next 40s from LMAR in red, NN in blue, ridge regression in green, and LICORS in purple.*

Another reason why the LMAR model works relatively well when the test data differ from the training data is that the form of the dependence of forecasts on the most recent p observations evolves, whereas it remains static for the other three methods. While the parameters of the model are not re-estimated during real-time prediction, LMAR uses the entire history of the time series in making forecasts, not just the first 40 seconds alongside the most recent p observations, as is the case with the other three methods. With appropriate parallel computating resources, all methods could theoretically update parameters continuously (or periodically) throughout

treatment. Murphy and Dieterich (2006) continuously retrained neural networks using the updated history of the respiratory trace. While they did not compare this to the alternative of not actively updating the forecast model, Krauss, Nill and Oelfke (2011) did so and found a small improvement in RMSE of about 1–3%.

When the time series are more well-behaved, all four methods perform quite well; in fact, neural networks tend to have the lowest errors when all four curves are accurate. Figure 8 shows the training and prediction test series for a strongly periodic respiratory trace. We should expect the performance of neural networks to be superior when the dynamics of the tumor motion are stable, as the parameter space for neural networks is far larger; in theory, feedforward neural networks with at least one hidden layer can approximate any continuous function arbitrarily well [Hornik, Stinchcombe and White (1989)], including time series prediction.

5.4. *Interval and distributional forecasts.* Unlike commonly used time series models in the tumor-tracking literature, the LMAR model provides multimodal, heteroskedastic predictive distributions, which are theoretically appropriate for forecasting respiratory motion. Despite this, our analysis of predictive performance has focused exclusively on the accuracy of point forecasts because in current implementations of tumor-tracking systems, there is no clinical value in obtaining interval or distributional forecasts. The treatment beam has a fixed width and is always on, meaning an interval or distributional forecast does not alter the optimal course of action of a tumor-tracking system already supplied with a point forecast. However, interval/distributional forecasts would prove valuable if we could, for instance, suspend the treatment beam instantaneously if the predicted probability of the tumor location being enclosed by the treatment beam fell below a certain threshold.

Table 4 gives a summary of the performance of out-of-sample interval and distributional forecasts to complement the summaries of point forecasts. The LMAR model, by specifying a data-generating process, naturally provides full predictive distributions as a by-product of point prediction. The same is true for ridge regression (assuming the typical homoskedastic Gaussian structure for the residuals) and LICORS. Neural networks do not naturally provide predictive distributions; following Tibshirani (1996) we obtain them by bootstrapping, while assuming prediction errors are (heteroskedastic) independent Gaussians, with mean 0 and variance estimated by bootstrapping.

We expect LMAR prediction intervals to undercover, since uncertainty in the estimation of Σ is omitted from our forecasts. While this is indeed the

		0.2s Forecast		0.4s Forecast		0.6s Forecast	
Patient	Method	Coverage	Log PS	Coverage	Log PS	Coverage	Log PS
4	LMAR	0.84	0.72	**0.86**	1.30	**0.93**	1.37
	NNs	0.88	**0.57**	0.83	1.34	0.85	1.58
	Ridge	0.85	0.80	0.84	1.53	0.84	1.86
	LICORS	**0.89**	0.70	0.84	**1.03**	0.84	**1.32**
5	LMAR	**0.87**	**0.71**	**0.88**	**1.20**	**0.93**	**1.30**
	NNs	0.85	0.72	0.78	1.52	0.80	1.75
	Ridge	0.85	0.91	0.84	1.53	0.82	1.91
	LICORS	0.84	1.04	0.82	1.46	0.79	1.78
6	LMAR	**0.87**	1.25	**0.88**	**1.85**	**0.93**	**2.07**
	NNs	0.79	1.31	0.74	2.16	0.76	2.53
	Ridge	**0.87**	**1.22**	0.85	1.91	0.83	2.26
	LICORS	0.79	1.58	0.70	2.57	0.66	2.82
7	LMAR	0.85	**0.30**	**0.85**	**0.87**	**0.89**	**1.09**
	NNs	**0.88**	0.48	0.84	1.35	0.84	1.82
	Ridge	0.86	0.63	0.83	1.49	0.82	1.95
	LICORS	0.84	0.78	0.77	1.16	0.76	1.59
8	LMAR	**0.89**	1.67	0.91	2.30	0.94	2.60
	NNs	0.94	**1.53**	0.82	2.36	**0.90**	2.59
	Ridge	0.88	1.82	0.85	2.51	0.82	2.90
	LICORS	0.94	1.71	**0.90**	**2.11**	0.88	**2.39**
9	LMAR	**0.89**	**0.87**	**0.90**	**1.65**	**0.92**	**2.07**
	NNs	0.86	1.02	0.78	2.20	0.80	2.77
	Ridge	0.81	1.54	0.81	2.21	0.81	2.64
	LICORS	0.86	1.62	0.81	1.98	0.79	2.31
10	LMAR	**0.86**	**1.18**	**0.88**	**1.94**	**0.91**	2.33
	NNs	0.84	1.23	0.76	2.25	0.79	2.65
	Ridge	0.83	1.35	0.84	2.03	0.84	2.44
	LICORS	**0.86**	1.61	0.82	2.02	0.81	**2.31**
11	LMAR	0.85	**1.38**	**0.87**	2.13	**0.91**	2.36
	NNs	0.87	1.50	0.80	2.70	0.83	2.91
	Ridge	0.86	1.63	0.85	2.44	0.85	2.84
	LICORS	**0.88**	1.56	0.83	**1.99**	0.82	**2.25**

TABLE 4

Summary of interval and distributional forecasts for all four methods at all three forecast windows. The interval coverage considered is 90% *confidence intervals.* Log PS *refers to the log probability score of the predictive distribution. For each metric, the most desirable value among the four methods for each patient/forecast window combination is in* **bold.**

case, for all patients and forecast windows, 90% prediction intervals have between 84% and 94% coverage—a more appropriate range than any other method can claim.

The logarithmic score in Table 4 refers to the negative logarithm of the predictive density evaluated at the true observation, averaged over each out-of-sample prediction (the result in Table 4 then averages each of these scores over all beams from the same patient). The logarithmic score is a *proper* scoring rule—its expected value is minimized by the oracle (or true) predictive distribution—thus lower values indicate a better fit between the predictive distributions and realized values of a patient's time series [Gneiting, Balabdaoui and Raftery (2007)].

Generalizing across patients and forecast windows, in comparison to the other methods considered, the LMAR model seems to most accurately characterize prediction uncertainty.

6. Discussion. The location-mixture autoregressive (LMAR) model introduced in this paper provides accurate, real-time forecasts of lung tumor motion. Our method achieves better performance on out-of-sample prediction for forecasts windows of 0.2s, 0.4s, and 0.6s for the majority of the patients considered than existing methods such as neural networks (which performed best in a prediction comparison study of Krauss, Nill and Oelfke (2011)) and penalized linear models (a common baseline for judging predictive performance). We also note that uncertainty quantification is quite straightforward using our model, where as it is hard to do using neural networks.

The LMAR model is similar to other autoregressive models that yield multimodal conditional distributions, such as the class of threshold autoregressive models [Tong (1978)], yet the parameter space consists of just a single, low-dimensional covariance matrix, and the model admits accurate closed-form approximations of multiple-step ahead predictive distributions. The LMAR model also has a useful interpretation in the context of time series motifs, which can describe the data-generating process and the form of forecasts.

While the predictive performance of our method on this data set is very encouraging, the parameter inference for the LMAR model presented here is approximate, and the assumptions of both the model and its inference may not be appropriate for some other non-linear time series. Formalizing and generalizing the LMAR model is thus a fruitful area for future work.

Real-time prediction of lung tumor motion presents additional challenges to those presented in this work. It is preferable to have as short a training window as possible, since during this time the patient may be irradiated without actually receiving the benefit of tumor tracking. While some training is actually necessary to estimate the system latency in some cases (we have

treated it as fixed throughout this work), the 40 seconds used for training in this paper (while typical in the literature on the subject) could ideally be reduced.

Also, one can consider patient-specific hyperparameter values and/or tuning parameters or modify the model to borrow information across the patients. Due to the need for real-time model fitting before we can forecast, it is most likely infeasible to apply any model selection criteria (either within-model, such as for hyperparameters, or between-model) after having begun to observe data. More study of between-patient and within-patient variability in model fits could help researchers use more patient-optimal prediction methods (as well as begin prediction after a shorter training sequence, as they wouldn't need to rely solely on the observed data for parameter estimation).

The parametric simplicity of the LMAR model, as well as its formalization as a statistical model as opposed to a prediction algorithm, enable generalizations of our procedure to include hierarchical models and other statistical structures that address the challenges of delivering accurate external beam radiotherapy. Combined with its excellent predictive performance on real data, the LMAR model represents a promising new contribution to this area of research.

Acknowledgements. The authors would like to thank Dr. Seiko Nishioka of the Department of Radiology, NTT Hospital, Sapporo, Japan and Dr. Hiroki Shirato of the Department of Radiation Medicine, Hokkaido University School of Medicine, Sapporo, Japan for sharing the patient tumor motion dataset with us. The project described was supported by Award Numbers RSCH1206 (JHL) from the Radiological Society of North America and R21CA156068 (RIB) from the National Cancer Institute. NSP is partially supported by the NSF grant DMS 1107070. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, National Science Foundation or the National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplement A: Data
(doi: COMPLETED BY THE TYPESETTER; .zip). Tumor tracking data used in this paper.

Supplement B: Code
(doi: COMPLETED BY THE TYPESETTER; .zip). `R` Code used for fitting and forecasting with the LMAR model.

References.

Berbeco, R. I., Nishioka, S., Shirato, H., Chen, G. T. and Jiang, S. B. (2005). Residual motion of lung tumours in gated radiotherapy with external respiratory surrogates. *Physics in Medicine and Biology* **50** 3655–3667.

Boots, B. and Gordon, G. J. (2011). An online spectral learning algorithm for partially observable nonlinear dynamical systems. In *Proceedings of the 25th International Conference on Artificial Intelligence* 293–300.

Boshnakov, G. N. (2009). Analytic expressions for predictive distributions in mixture autoregressive models. *Statistics and Probability Letters* **79** 1704–1709.

Carvalho, A. X. and Tanner, M. A. (2005). Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification. *IEEE Transactions on Neural Networks* **16** 39–56.

De Gooijer, J. G. and Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing, and forecasting. *International Journal of Forecasting* **8** 135–156.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1–38.

D'Souza, W., Naqvi, S. A. and Cedric, X. Y. (2005). Real-time intra-fraction-motion tracking using the treatment couch: a feasibility study. *Physics in Medicine and Biology* **50** 4021–4033.

Ernst, F., Schlaefer, A. and Schweikard, A. (2007). Prediction of respiratory motion with wavelet-based multiscale autoregression. In *Medical Image Computing and Computer-Assisted Intervention* (N. Ayache, S. Ourselin and A. Maeder, eds.) 668–675. Springer, New York, NY.

Ernst, F. and Schweikard, A. (2009). Forecasting respiratory motion with accurate online support vector regression (SVRpred). *International Journal of Computer Assisted Radiology and Surgery* **4** 439–447.

Ernst, F., Dürichen, R., Schlaefer, A. and Schweikard, A. (2013). Evaluating and comparing algorithms for respiratory motion prediction. *Physics in Medicine and Biology* **58** 3911–3929.

Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence* **24** 164–181.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 243–268.

Goerg, G. M. (2013a). Forecastable component analysis. In *Proceedings of the 30th International Conference on Machine Learning* 64–72.

Goerg, G. M. (2013b). LICORS: Light cone reconstruction of states—predictive state estimation from spatio-temporal data R package version 0.2.0.

Goerg, G. M. and Shalizi, C. R. (2012). LICORS: Light cone reconstruction of states for non-parametric forecasting of spatio-temporal systems Technical Report, Carnegie Mellon University, Department of Statistics.

Goerg, G. M. and Shalizi, C. R. (2013). Mixed LICORS: A nonparametric algorithm for predictive state reconstruction. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics* 289–297.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society* **57** 357–384.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2** 359–366.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society: Series B (Methodological)* **32** 175–208.

Kalet, A., Sandison, G., Wu, H. and Schmitz, R. (2010). A state-based probabilistic model for tumor respiratory motion prediction. *Physics in Medicine and Biology* **55** 7615–7631.

Krauss, A., Nill, S. and Oelfke, U. (2011). The comparative performance of four respiratory motion predictors for real-time tumour tracking. *Physics in Medicine and Biology* **56** 5303–5317.

Krolzig, H.-M. (2000). Predicting Markov-switching vector autoregressive processes. Technical Report, University of Oxford, Department of Economics.

Lau, J. W. and So, M. K. (2008). Bayesian mixture of autoregressive models. *Computational Statistics and Data Analysis* **53** 38–60.

Le, N. D., Martin, R. D. and Raftery, A. E. (1996). Modeling flat stretches, bursts, outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association* **91** 1504–1515.

Lin, J., Lonardi, S., Keogh, E. and Patel, P. (2002). Finding motifs in time series. In *Proceedings of the 2nd Workshop on Temporal Data Mining* 53–68.

Littman, M. L., Sutton, R. S. and Singh, S. P. (2002). Predictive representations of state. In *Proceedings of Advances in Neural Information Processing Systems 14* 1555–1561.

McCulloch, R. E. and Tsay, R. S. (1994). Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis* **15** 523–539.

Murphy, M. J. and Dieterich, S. (2006). Comparative performance of linear and nonlinear neural networks to predict irregular breathing. *Physics in Medicine and Biology* **51** 5903–5914.

Murphy, M. J., Isaakson, M. and Jalden, J. (2002). Adaptive filtering to predict lung tumor motion during free breathing. In *Computer Assisted Radiology and Surgery* (H. U. Lemke, M. W. Vannier, K. Inamura, A. G. Farman, K. Doi and J. H. C. Reiber, eds.) 539–544. Springer, New York, NY.

Riaz, N., Shanker, P., Wiersma, R., Gudmundsson, O., Mao, W., Widrow, B. and Xing, L. (2009). Predicting respiratory tumor motion with multi-dimensional adaptive filters and support vector regression. *Physics in Medicine and Biology* **54** 5735–5748.

Rottmann, J., Keall, P. and Berbeco, R. (2013). Markerless EPID image guided dynamic multi-leaf collimator tracking for lung tumors. *Physics in Medicine and Biology* **58** 4195–4204.

Ruan, D. and Keall, P. (2010). Online prediction of respiratory motion: multidimensional processing with low-dimensional feature learning. *Physics in Medicine and Biology* **55** 3011–3025.

Schweikard, A., Glosser, G., Bodduluri, M., Murphy, M. J. and Adler, J. R. (2000). Robotic motion compensation for respiratory movement during radiosurgery. *Computer Aided Surgery* **5** 263–277.

Shalizi, C. R. (2003). Optimal nonlinear prediction of random fields on networks. In *Discrete Mathematics and Theoretical Computer Science Proceedings (volume AB)* 11–30.

Sharp, G. C., Jiang, S. B., Shimizu, S. and Shirato, H. (2004). Prediction of respira-

tory tumour motion for real-time image-guided radiotherapy. *Physics in Medicine and Biology* **49** 425–440.

Tanaka, Y., Iwamoto, K. and Uehara, K. (2005). Discovery of time-series motif from multi-dimensional data based on MDL principle. *Machine Learning* **58** 269–300.

Tibshirani, R. (1996). A comparison of some error estimates for neural network models. *Neural Computation* **8** 152–163.

Tong, H. (1978). On a threshold model. In *Pattern Recognition and Signal Processing* (C. H. Chen, ed.) 575–586. Sijthoff & Noordhoff, Amsterdam.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford, UK.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)* **42** 245–292.

Tong, H. and Moeanaddin, R. (1988). On multi-step non-linear least squares prediction. *Journal of the Royal Statistical Society: Series D (The Statistician)* **37** 101–110.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York, NY.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62** 95–115.

Ye, L. and Keogh, E. (2009). Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 947–956.