Inference and Prediction Problems for Spatial and Spatiotemporal Data

Inference and Prediction Problems for Spatial and Spatiotemporal Data

a dissertation presented
by
Daniel Leonard Cervone
to
The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Statistics

Harvard University
Cambridge, Massachusetts
April 2015

Inference and Prediction Problems for Spatial and Spatiotemporal Data

Abstract

This dissertation focuses on prediction and inference problems for complex spatiotemporal systems. I explore three specific problems in this area—motivated by real data examples—and discuss the theoretical motivations for the proposed methodology, implementation details, and inference/performance on data of interest.

Chapter 1 introduces a novel time series model that improves the accuracy of lung tumor tracking for radiotherapy. Tumor tracking requires real-time, multiple-step ahead forecasting of a quasi-periodic time series recording instantaneous tumor locations. Our proposed model is 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. Compared with other families of mixture autoregressive models, LMAR is easier to fit (with a smaller parameter space) and better suited to online inference and multiple-step ahead forecasting as there is no need for Monte Carlo. Against other candidate models in statistics and machine learning, our model provides superior predictive performance for clinical data.

Chapter 2 develops a stochastic process model for the spatiotemporal evolution of a basketball possession based on tracking data that records each player's exact location at 25Hz. Our model comprises of multiresolution transition kernels that simultaneously describe players' continuous motion dynamics along with their decisions, ball movements, and other discrete actions. Many such actions occur very sparsely in player $\times$ location space, so we use hierarchical models to share information across different players in the league and disjoint regions on the basketball court—a challenging problem given the scale of our data (over 400 players and 1 billion space-time observations) and the computational cost of inferential methods in spatial statistics. Our

framework, in addition to offering valuable insight into individual players' behavior and decision-making, allows us to estimate the instantaneous expected point value of an NBA possession by averaging over all possible future possession paths.

In Chapter 3, we investigate Gaussian process regression where inputs are subject to measurement error. For instance, in spatial statistics, input measurement errors occur when the geographical locations of observed data are not known exactly. Such sources of error are not special cases of "nugget" or microscale variation, and require alternative methods for both interpolation and parameter estimation. We discuss some theory for Kriging in this regime, as well as using Hybrid Monte Carlo to provide predictive distributions (and parameter estimates, if necessary). Through simulation study and analysis of northern hemipshere temperature data from the summer of 2011, we show that appropriate methods for incorporating location measurement error are essential to reliable inference in this regime.

Contents

Listing of figures

Acknowledgments

First and foremost, I want to thank the members of my dissertation committee, Natesh Pillai, Carl Morris, and Luke Bornn. Their wisdom, energy, mentorship, and friendship have guided my research and built the foundation for my career in Statistics. I am also indebted to the Harvard Statistics faculty at large, for admitting me into the program and through lectures, seminars, and further conversation, teaching new ideas and challenging me. I am furthermore grateful for the work of the Harvard Statistics staff in helping me navigate graduate school and making the department feel like a home.

My classmates have been a constant source of inspiration and new ideas. Through conversations about everything from new paper topics to debugging code, I've realized that Statistics research is done best and most enjoyably as part of a team. In particular I'd like to acknowledge Alex Franks and Alex D'Amour, whose friendship has enriched both work and life during the last five years.

Teaching has been one of the most rewarding and enjoyable experiences of graduate school. My experience as a Teaching Fellow has made me a better communicator and collaborator, and I value the impact of good teaching on the field of Statistics equally as I do good research. I am especially thankful for the guidance of Virginia Mauer, Xiao-Li Meng, Dave Harrington, Joe Blitzstein, and Cassandra Pattanayak in developing my skill and confidence in teaching.

Lastly, I am especially grateful for the love and support of my family and fiancée, Einor.

I AM THE FIRST AUTHOR on the work comprising each section of this dissertation, though I also want to acknowledge those who made significant contributions to this material. Natesh Pillai and Debdeep Pati (Florida State University) collaborated on the methods discussed in Chapter 1, while Ross Berbeco and John Henry Lewis (Harvard Medical School) helped get us started on this problem and refine many of our early ideas. We are grateful for Dr. Seiko Nishioka (NTT Hospital, Sapporo, Japan) and Dr. Hiroki Shirato (Hokkaido University School of Medicine, Sapporo, Japan) for sharing the patient tumor motion dataset with us.

Alex D'Amour and Luke Bornn worked together with me on the material in Chapter 2. The research topic of Chapter 2 was inspired by discussions with Kirk Goldsberry (Harvard Center for Geographic Analysis), whose basketball knowledge has guided this project throughout its course. Talking with Dr. Goldsberry has motivated our research, and his presence has been invaluable to the success of the project. We are furthermore grateful to STATS LCC and the NBA for providing the data used in Chapter 2.

Natesh Pillai contributed to Chapter 3.

0

Preface

The chapters in this dissertation are each self-contained research topics, motivated by distinct data sets and applications. However, these chapters share common methodological challenges and reflect a unified approach to statistical research. Time series, spatial, and spatiotemporal data are indexed by observations in time and/or space. As such, they generally share a unique, non-exchangeable dependence structure: the joint distribution of a collection of variables that are close together (either in space, time, or both) behaves differently from a collection of variables that are spaced far apart.

Futhermore, the data applications considered in this thesis, and their corresponding statistical

models, can be represented by a general two-level model:

$$Y_i|\theta_i \overset{iid}{\sim} f(y;\theta_i) \tag{1}$$

$$\boldsymbol{\theta} \sim g(\boldsymbol{\theta};\phi) \tag{2}$$

$$Y_i \sim m(y;\phi) = \int f(y;\theta_i)g(\boldsymbol{\theta};\phi)d\boldsymbol{\theta} \tag{3}$$

$$\boldsymbol{\theta}|\mathbf{Y} \sim h(\boldsymbol{\theta};\phi,\mathbf{Y}) \propto \prod_{i=1}^{n} f(y_i;\theta_i)g(\boldsymbol{\theta};\phi) \tag{4}$$

$$Y^*|\mathbf{Y} \sim p(y^*;\phi,\mathbf{Y}) = \int f(y^*;\boldsymbol{\theta})h(\boldsymbol{\theta};\phi,\mathbf{Y})d\boldsymbol{\theta}. \tag{5}$$

Morris (1995) refers to (1)–(2) as the "descriptive model", which describes the data conditional on a multi-dimensional parameter of interest $\boldsymbol{\theta}$ and the prior for $\boldsymbol{\theta}$, and to (3)–(4) as the "inferential model". The marginal likelihood m summarizes all information in the data for inferring hyperparameters ϕ, and the posterior h provides inference for $\boldsymbol{\theta}$ given the hyperparameters and observed data. (5) is the posterior predictive distribution. A fully Bayesian treatment of (1)–(4) might place a prior on ϕ, while an empirical Bayes approach estimates ϕ using m, and fixes ϕ in subsequent analysis.

Each chapter of this dissertation, taken in view of (1)–(5), features $\boldsymbol{\theta}$ as a process of interest. In Chapter 1, $\boldsymbol{\theta}$ represents a sequence of motifs (distinctive patterns) that describe the shape and future evolution of a time series (individual measurements of this time series are Y_i). In Chapter 2, $\boldsymbol{\theta}$ is a latent process describing basketball players' hazards (small-scale probabilities) to initiate various actions, such as passing or attempting a shot. $\mathbf{Y}$ can be thought of as spatiotemporal point process data, with $\boldsymbol{\theta}$ representing an underlying space-time intensity. The investigation in Chapter 3 treats $\boldsymbol{\theta}$ as a Gaussian process—an unknown real-valued function supported on a spatial domain, and Y_i as measurements of this function at various locations in the domain.

Much of the methodology discussed in each chapter focuses on inferring the latent process of interest $\boldsymbol{\theta}$, given the data $\mathbf{Y}$. Of course, the models in this dissertation feature additional parameters and structural assumptions specific to each problem. The methodological and computational techniques necessary to leverage the information in the data also further vary by problem, yet the general structure of (1)–(5) is a common framework for my research.

The work throughout this thesis is motivated by computational challenges and constraints. In the three chapters to follow, these challenges originate from different features of each problem. In Chapter 1, our data application demands real-time inference and prediction, which motivates models and techniques that are computationally simple and efficient. In Chapter 2, we analyze a massive data set, consisting of around a billion observed data points in space-time, and are thus forced to design models and implement inference that is feasible for this data. Chapter 3 treats Gaussian process regression, which is a computationally expensive model that scales poorly with data size; traditional implementations fail with any more than several thousand observations.

Even more than by shared methodology, this work is anchored by application and scientific utility. Each chapter is motivated by—and includes analysis of—data that has scientific and/or cultural value, and is of broad interest outside Statistics. The famous quote attributed to John Tukey, "The best thing about being a statistician is that you get to play in everyone's backyard", while clichéd, truly describes my own passion for Statistics, and this dissertation reflects that.

1

A Location-Mixture Autoregressive Model for Online Forecasting of Lung Tumor Motion

1.1 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 et al. (2013); D'Souza et al. (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 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 et al. (2002)], state-space models [Kalet et al. (2010)], and wavelets [Ernst et al. (2007)]; machine learning methods include kernel density estimation [Ruan & Keall (2010)], support vector regression [Riaz et al. (2009); Ernst & Schweikard (2009)], and neural networks [Murphy et al. (2002); Murphy & 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 et al. (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,

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

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 & Lim (1980)], including Markov-switching autoregressive models [Hamilton (1989)] and the mixture autoregressive models of Wong & 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 & Keogh (2009); Tanaka et al. (2005); Fu (2011)]. For the purposes of forecasting, predictive state representations [Littman et al. (2002); Shalizi (2003); Boots & Gordon (2011)] categorize time series motifs not as subseries of the observed data, but as equivalence classes of conditional predictive distributions.

Section 1.2 of this paper discusses the important features of the data we use and graphically motivates our model. Section 1.3 formally introduces the LMAR model and describes parameter estimation and forecasting using principled methods that are feasible in real-time. Section 1.4 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 1.5, and Section 1.6 summarizes and points out future directions.

1.2 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.

1.2.1 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*. Figure 1.1 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 1.1 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

*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.

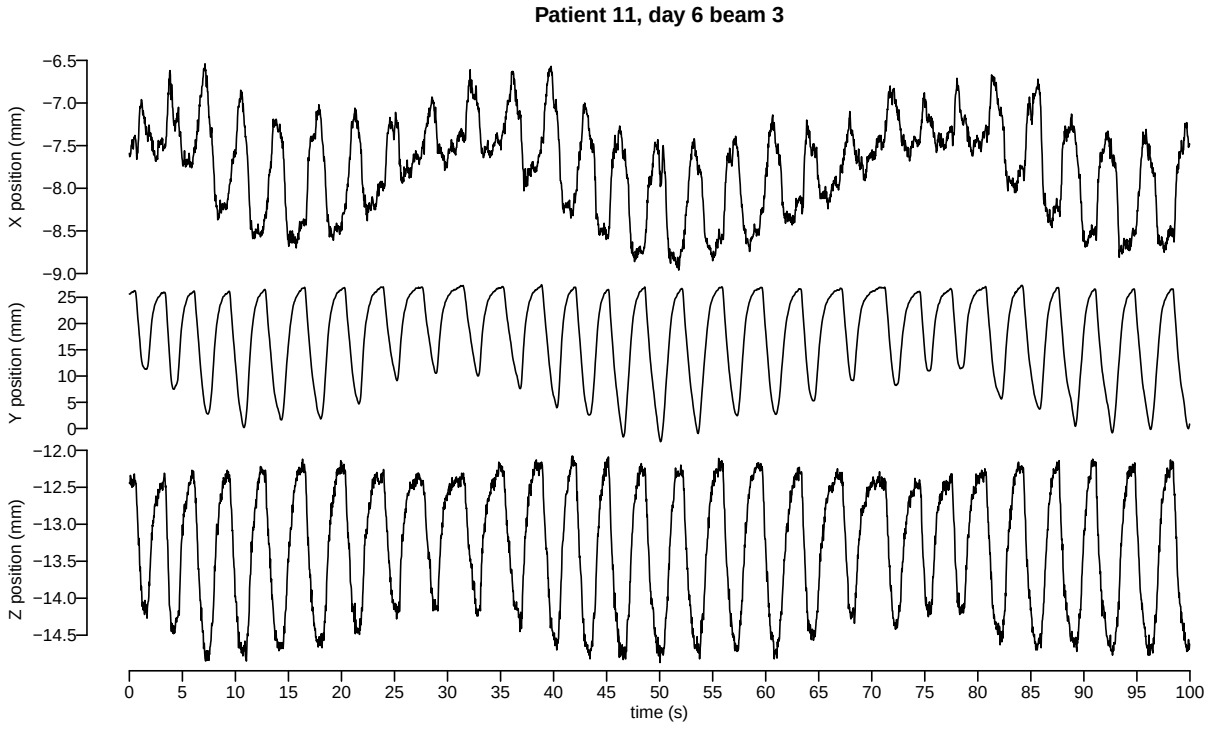


Figure 1.1: Sample time series of 3D locations of lung tumor. The X axis is the lateral-medial (left-right) direction, Y axis superior-inferior, and Z axis anterior-posterior

visually discerned from Figure 1.1, it is common for respiratory cycles to change periodicity, either sporadically or gradually over time. Table 1.1 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 1.1: 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 exhibit some periodic behavior (see Figure 1.2), but the signal is weak relative to the noise[†]. 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 & 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

[†]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.

Krauss et al. (2011) used the principal components, then transformed their predictions to the original linear basis of the data.

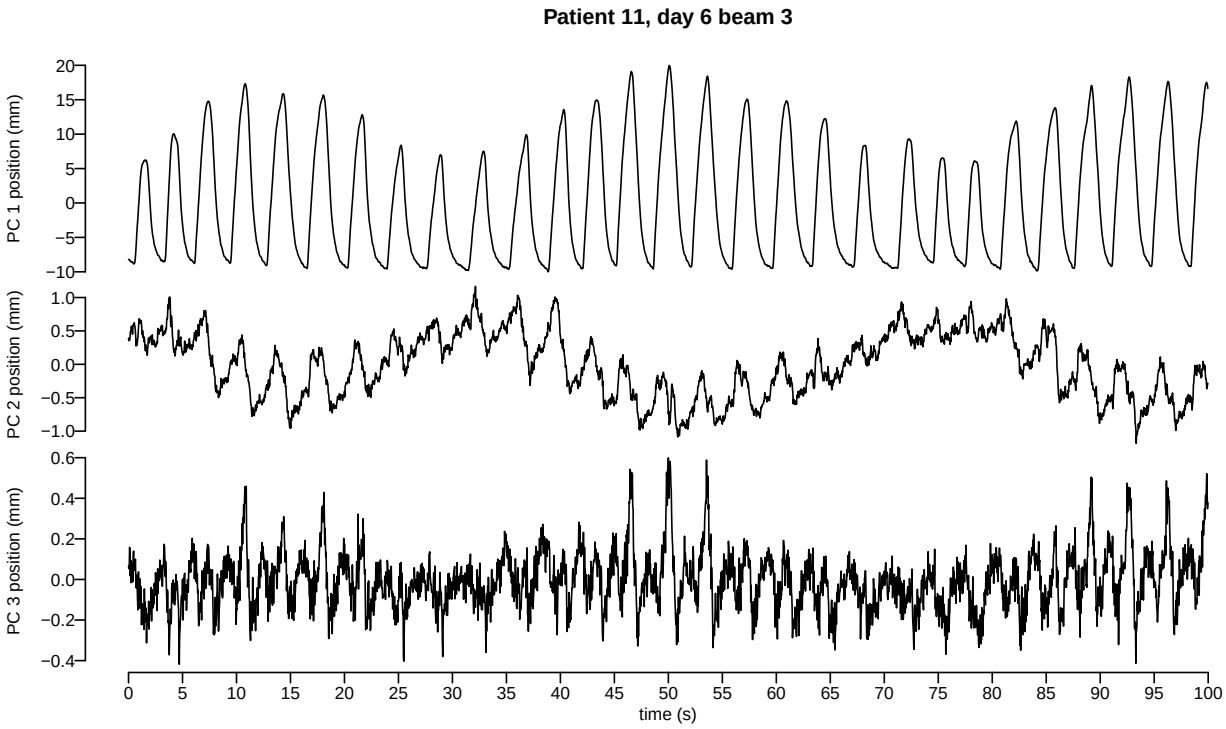


Figure 1.2: Time series of principal components. Components 2 and 3 exhibit periodic behavior, but with much smaller magnitude.

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 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.

1.2.2 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 1.3.2). Figure 1.3 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 et al. (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 1.2. 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 1.4. Colored 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 11^{th}–13^{th} points are decreasing).

To predict future observations, we can incorporate the points immediately succeeding the endpoints of black motifs. Figure 1.5 shows these trajectories (in gray), and the actual current trajectory of the process is shown in orange, with a point giving the value 0.4s in the future. The

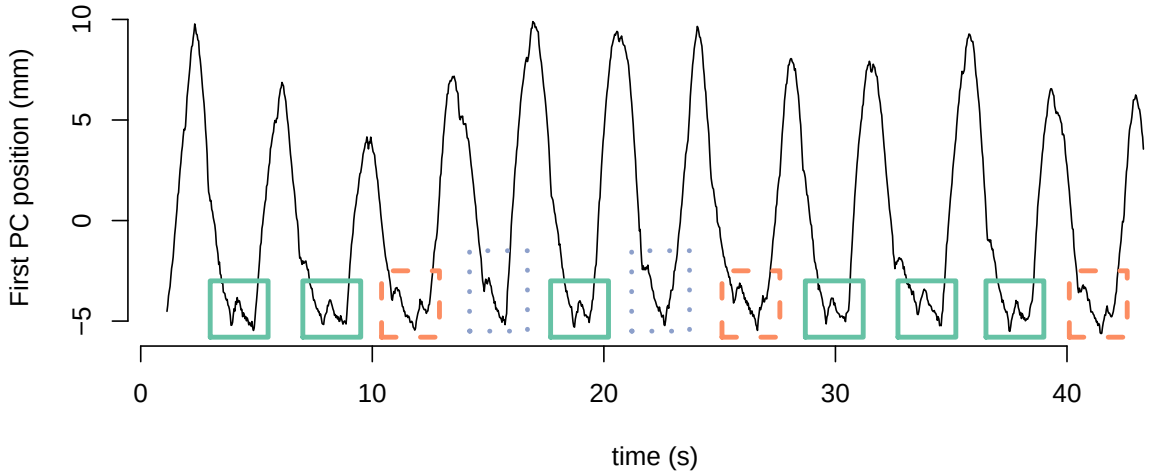


Figure 1.3: Recurring patterns (coded by color and line type) in the first principal component of patient 10, day 1, beam 3. Areas boxed by lines of the same color/line type resemble one another. The behavior highlighted in these motifs is most likely caused by the patient's heartbeat.

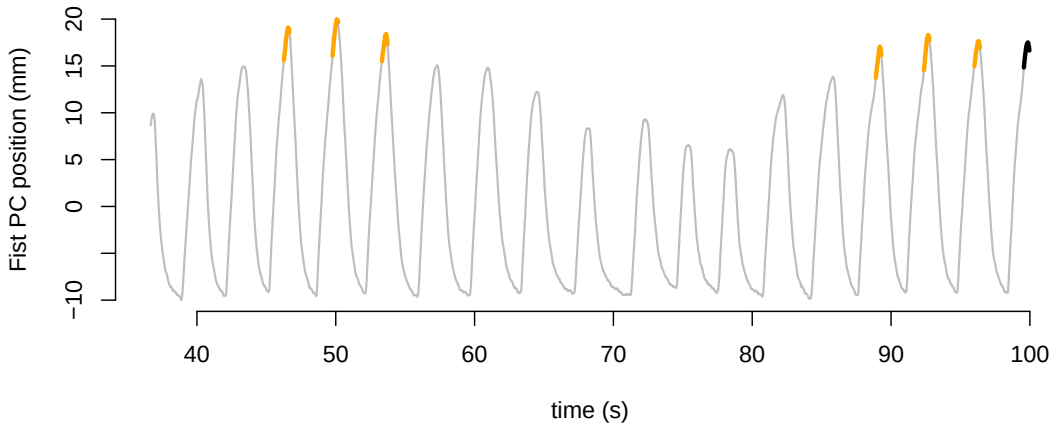


Figure 1.4: The most recent 0.43s (13 observations) are in black. The thicker, orange segments share similar local history.

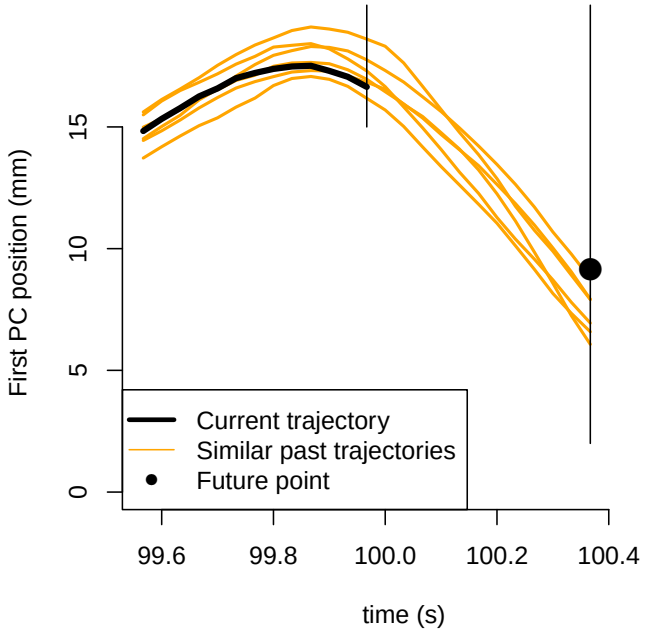


Figure 1.5: The recent history of the process (thick black line) instantiates a motif. Previous instances of this motif, and their subsequent evolutions, are in orange and provide reasonable predictions for future points (black dot).

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 1.3, implements the forecasting approach sketched in this subsection using an autoregressive model for the data-generating process.

1.3 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 $Y_{i+0:p}$.

1.3.1 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, non-negative 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:

$$
\begin{aligned}
Y_i | Y_{(-m):(i-1)} &\sim \sum_{j \in \mathcal{J}_i} \alpha_{i,j} \, \mathrm{N}(\mu_{i,j}, \sigma^2), \\
\text{where} \qquad \alpha_{i,j} &= \frac{\exp\left(-\frac{1}{2} V_{ij}' \Sigma_{11}^{-1} V_{ij}\right)}{\sum_{l \in \mathcal{J}_i} \exp\left(-\frac{1}{2} V_{il}' \Sigma_{11}^{-1} V_{il}\right)} \\
\text{and} \qquad \mu_{i,j} &= Y_{i-j} + \gamma' V_{ij}.
\end{aligned}
\tag{1.3}
$$

The model in (1.3) 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

$$
\tilde{\mu}_{i,j} = Y_{i-j} - \sum_{l=1}^{p} \gamma_l Y_{j-l} \tag{1.4}
$$

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 (1.3) 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 (1.3) specifies the role of time series motifs in the data-generating process, which was informally discussed in Section 1.2.2. To illustrate this, we introduce a latent variable M_i that takes values in $\mathcal{J}_i$, such that for all $j \in \mathcal{J}_i$,

$$\mathbb{P}(M_i = j | Y_{(-m):(i-1)}) \propto \exp\left(-\frac{1}{2} V_{ij}' \Sigma_{11}^{-1} V_{ij}\right). \tag{1.5}$$

Then, given $M_i = j$, we induce the same distribution for Y_i as in (1.3) by assuming

$$Y_i \mid [M_i = j, Y_{(-m):(i-1)}] \sim \mathcal{N}(Y_{i-j} + \gamma' V_{ij}, \sigma^2). \tag{1.6}$$

Expression (1.6) 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 (1.6)). 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 et al. (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 (1.2.2). From (1.5), 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.

1.3.2 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

$$F(y|Y_{(-m):(i-1)}) = \sum_{j=1}^{d} \alpha_{i,j} \Phi\left(\frac{y - (\beta_{0,j} + \sum_{l=1}^{p} \beta_{l,j} Y_{i-l})}{\sigma_j}\right), \tag{1.7}$$

where $\sum_{j=1}^{d} \alpha_{i,j} = 1$ for all i and Φ denotes the standard normal CDF. Models satisfying (1.7) can be represented in the framework of threshold autoregressive models [Tong (1978); Tong & Lim (1980); see Tong (1990) for a book-length treatment], which represent (1.7) using an indicator series $\{M_i\}$ taking values on $\{1, \dots, d\}$, such that

$$Y_i = \beta_{0,M_i} + \sum_{l=1}^{p} \beta_{l,M_i} Y_{i-l} + \sigma_{M_i} \epsilon_i, \tag{1.8}$$

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 & Lim (1980).

A canonical model of this form is the mixture autoregressive model of Le et al. (1996) and Wong & Li (2000), which assumes $\{M_i\}$ are i.i.d. and independent of Y. Another special case of (1.8) is when M is a Markov chain, such as in the Markov-switching autoregressive models of Hamilton (1989) and McCulloch & Tsay (1994). More general stochastic structure for M is considered by Lau & So (2008), as well as in mixture-of-experts models in the machine learning literature [Carvalho & Tanner (2005)]. These models seem favorable over the mixture autoregressive models of Wong & 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 (1.7) in that the mixture means, following (1.3)–(1.4), 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 (1.7). 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 (1.7), 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 (1.3). 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 (1.7) 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 (1.7), 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 1.3.4. Closed-form predictive distributions for $k > 1$ are not available for many models of the form (1.7) (the exception is the Markov-switching autoregressive models of Hamilton (1989); for a discussion see Krolzig (2000)). Wong & Li (2000) recommended Monte Carlo estimates of k-step ahead pre-

dictive 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 (1.7), De Gooijer & 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 & Moeanaddin (1988) for a discussion of the robustness of medium-to-long range forecasts using threshold autoregressive models).

1.3.3 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 & Sprott (1970)].

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

$$L_2(\Sigma) = \prod_{i=0}^{n} \frac{1}{\sigma} \left[\sum_{j \in \mathcal{J}_i} \exp\left(-\frac{1}{2\sigma^2}(Y_i - Y_{i-j} - \gamma' V_{ij})^2 \right) \times \left(\frac{\exp(-V_{ij}' \Sigma_{11}^{-1} V_{ij}/2)}{\sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2)} \right) \right]. \tag{1.9}$$

To maximize (1.9), we augment the data to $\{Y_{0:n}, M_{0:n}\}$, with M_i as in (1.5). This invites the use of the Expectation-Maximization (EM) algorithm [Dempster et al. (1977)] to estimate Σ. The

augmented-data (complete-data) conditional likelihood is

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

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

$$L_{2,\text{com}}(\Sigma) = \prod_{i=0}^{n} \frac{\exp\left(-\frac{1}{2} \sum_{j \in \mathcal{J}_i} \mathbf{1}[M_i = j] W_{ij}' \Sigma^{-1} W_{ij} \right)}{\sigma \sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2)}. \tag{1.10}$$

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 $V \sim \mathcal{N}(0, \Omega)$, then we have

$$\begin{aligned} \sum_{l \in \mathcal{J}_i} \exp(-V_{il}' \Sigma_{11}^{-1} V_{il}/2) &\approx |\mathcal{J}_i| \int \exp(-V' \Sigma_{11}^{-1} V/2) \frac{\exp(-V' \Omega^{-1} V/2)}{(2\pi)^{p/2} |\Omega|^{1/2}} dV \\ &= |\mathcal{J}_i| \left(\frac{|(\Sigma_{11}^{-1} + \Omega^{-1})^{-1}|}{|\Omega|} \right)^{1/2} \\ &= |\mathcal{J}_i| \left(\frac{|\Sigma_{11}|}{|\Sigma_{11} + \Omega|} \right)^{1/2}. \end{aligned} \tag{1.11}$$

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

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

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

$$\log(L_{2,\text{com}}(\Sigma)) \approx -\frac{n+1}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=0}^{n}\sum_{j\in\mathcal{J}_i}\mathbf{1}[M_i = j]W_{ij}'\Sigma^{-1}W_{ij}. \tag{1.12}$$

While (1.12) is much easier to work with than the logarithm of the exact conditional likelihood (1.10), the assumptions of this approximation are somewhat tenuous. Under this model (1.3), 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 (1.11) suggests. We nevertheless proceed with approximation (1.12) in place of (1.10), 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:

$$\begin{aligned} Q^{(t)}(\Sigma) &= \mathbb{E}_{\Sigma^{(t)}}[\log(L_{2,\text{com}}(\Sigma))|Y] \\ &\approx -\frac{n+1}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=0}^{n}\sum_{j\in\mathcal{J}_i}\omega_{ij}W_{ij}'\Sigma^{-1}W_{ij}, \end{aligned} \tag{1.13}$$

with $\Sigma^{(t)} = \text{argmax}(Q^{(t-1)}(\Sigma))$ 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 (1.13) can be found in closed form as a weighted sample covariance matrix,

$$\Sigma^{(t+1)} = \frac{1}{n+1}\sum_{i=0}^{n}\sum_{j\in\mathcal{J}_i}\omega_{ij}W_{ij}W_{ij}'. \tag{1.14}$$

Again, due to several different approximations used in maximizing the original conditional likelihood (1.9), 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.

1.3.4 A prediction model for fast implementation

Exact closed-form expressions for k-step ahead predictive distributions are not available for the model (1.3). 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 (1.12) 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 (1.12) from an exact model (1.3), here we treat (1.12) as exact and explore corresponding approximate models.

Let $Z_i = (Y_{i-p} \ \ldots \ Y_{i-1} \ \ Y_i)'$ for $0 \leq i \leq n$. Since $W_{ij} = Z_i - Z_j$, we may arrive at the likelihood expression (1.12) by assuming $Z_i \sim \mathcal{N}(Z_{i-M_i}, \Sigma)$ independently. This is obviously a misspecification, since for any $k \leq p$, 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 $L_{2,\text{com}}$:

$$L_a(\Sigma) = \prod_{i=0}^{n} \prod_{j \in \mathcal{J}_i} \left[|\Sigma|^{-1/2} \exp\left(-\frac{1}{2} W_{ij}' \Sigma^{-1} W_{ij} \right) \right]^{\mathbf{1}[M_i = j]}. \tag{1.15}$$

Indeed, we see that $L_a(\Sigma)$ is equal to the approximation of $L_{2,\text{com}}(\Sigma)$ given in (1.12). 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 (1.15) is that

$$Z_j \to Z_i \text{ if } Z_i | Z_j \sim \mathcal{N}(Z_j, \Sigma), \tag{1.16}$$

and indeed this relation is closely connected to the one defined in (1.6). They appear equivalent, as (1.6) is recovered by assuming $Z_i | Z_j \sim \mathcal{N}(Z_j, \Sigma)$, and then considering the conditional distribution $Y_i | Y_{(-m):(i-1)}$. However, for (1.16) to hold for all i requires the impossible assumption of Z_i being independent of Z_{i-1}, while the relation in (1.6) does not.

The corresponding Q function for this complete-data conditional likelihood (1.15) 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 (1.13), confirming that the "same" Σ parametrizes both the original data-generating process assumed in (1.3), and its degenerate approximation that we will use to make predictions (1.15). We may also think of maximizing Q as inferring motif instances given by the relation (1.16), 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

$$Y_{n+k}|Y_{(-m):n} \sim \sum_{j\in\mathcal{J}_{n+k}} \alpha_j^k \mathcal{N}(\mu_j^k, \sigma_k^2), \tag{1.17}$$

where

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

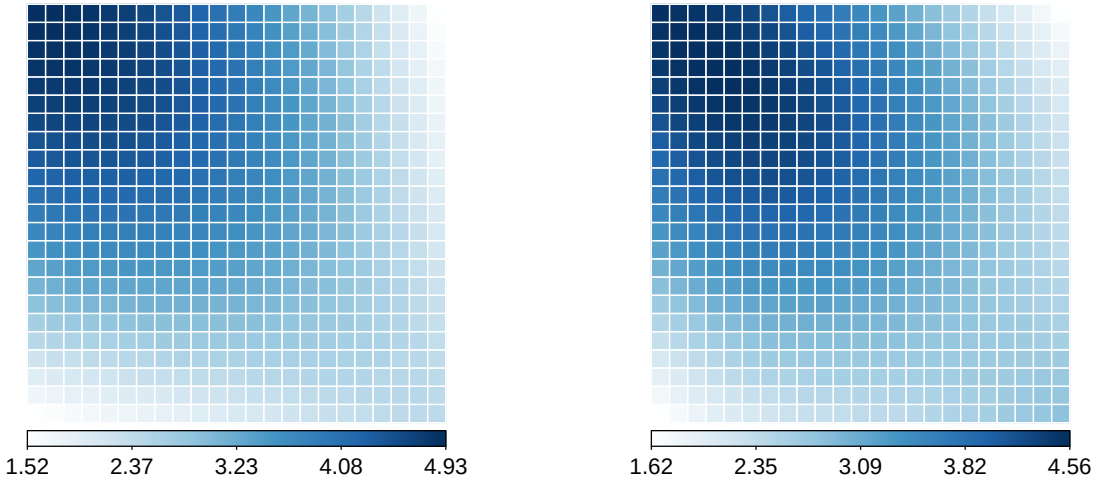
In terms of motifs, these predictive distributions result from considering the most recent subseries of the data of length $p-k+1$ as a partially observed motif instance, Z_{n+k}, which includes the future observation we wish to predict, Y_{n+k}. Using the implied motif relation in (1.16), we in-

fer both the motif for which Z_{n+k} is an instance, and derive predictive distributions using simple multivariate normal properties (1.17).

Of course, we use $\hat{\Sigma}$, the solution to (1.14), 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.

1.3.5 Interpreting $\hat{\Sigma}$

Figure 1.6 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.

Figure 1.6: 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.

1.4 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 (1.4.1) and ridge regression (1.4.2) both compare favorably to alternative methods with regards to prediction accuracy [Sharp et al. (2004); Krauss et al. (2011)]. LICORS (1.4.3) is a nonparametric and non-regression forecasting method based on predictive state representations of the time series [Goerg & Shalizi (2012, 2013)]. For each method, Sections 1.4.4–1.4.6 discuss data preprocessing and computational considerations relevant for real-time tumor tracking.

1.4.1 Feedforward neural networks

Multilayer feedforward neural networks with at least one hidden layer have been used to forecast lung tumor motion by Murphy et al. (2002) and Murphy & Dieterich (2006), as well as in simultaneous comparisons of several methods [Sharp et al. (2004); Krauss et al. (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

$$\hat{Y}_{i+k} = \beta_0 + \beta' G(X_i), \tag{1.18}$$

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 & 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

$$C(Y, \theta) = \sum_{i=0}^{n-k} (\hat{Y}_{i+k} - Y_{i+k})^2 + \lambda \theta' \theta, \tag{1.19}$$

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.

1.4.2 Ridge regression

The second competing method considered is a linear predictor of the form

$$\hat{Y}_{i+k} = \beta_0 + \beta' X_i, \tag{1.20}$$

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

$$C(Y, \beta_0, \beta) = \sum_{i=0}^{n-k} (\hat{Y}_{i+k} - Y_{i+k})^2 + \lambda(\beta_0^2 + \beta'\beta). \tag{1.21}$$

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 & Kennard (1970)], the model implied by (1.20)–(1.21) 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.

1.4.3 Light Cone Reconstruction of States (LICORS)

Mixed LICORS [Goerg & Shalizi (2013)] is a recent nonparametric forecasting method based on predictive state representations of spatiotemporal fields [Shalizi (2003); Goerg & 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

$$Y_{i+k}|Y_{(-m):i} \sim Y_{i+k}|X_i \sim Y_{i+k}|\epsilon(X_i), \tag{1.22}$$

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

$$C(Y, \epsilon, \mathbb{P}_1, \ldots, \mathbb{P}_K) = \prod_{i=0}^{n-k} \sum_{j=1}^{K} \mathbb{P}_j(Y_{i+k}) \mathbb{P}(S_i = j|X_i), \tag{1.23}$$

which acts as a likelihood, except for $\mathbb{P}_j$ being unknown. Goerg & Shalizi (2013) maximized (1.23) 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 $\{V_{ij} \text{ where } M_i = j\}$ 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 (1.22) will be a weighted average of the means of the predictive states $s_i \in \mathcal{S}$.

1.4.4 Data preprocessing

Similar to Krauss et al. (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 3. An additional 15 time series were eliminated because there were several gaps in the observation

Method	Hyperparameter	Description
LMAR	p	Motif length (1.16)
Neural Networks	p	Length of input vector X_i (1.18)
	h	Number of neurons in hidden layer (1.18)
	λ	Shrinkage; L2 penalty (1.19)
Ridge Regression	p	Length of input vector X_i (1.20)
	λ	Shrinkage; L2 penalty (1.21)
Mixed LICORS	p	Length of input vector X_i (1.22)

Table 1.2: List of global, patient-independent hyperparameters to be tuned for each prediction method

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 1.2.1.

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 et al. (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 1.2 lists the hyperparameters and/or tuning parameters for each of the prediction methods we consider. As described in Section 1.4.4, 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 1.2, 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 et al. (2011)]. Note that different hyperparameter values are chosen for different forecast windows.

1.4.6 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 (1.20)), since (1.21) can be minimized in closed form. Fitting neural networks (1.18), however, requires numerical optimization of (1.19). This was carried out using the `nnet` package in `R`, which implements the BFGS algorithm [Venables & Ripley (2002)]. Because (1.19) 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 (1.9) or log-likelihood is used; however, the EM updates given in (1.14) are only approximate, meaning the likelihood is not guaranteed to increase at every iteration. We found that using the approximate log-likelihood (1.12) 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 (1.12), such as the Frobenius norm of differences in the updates of $\hat{\Sigma}$. To ob-

tain 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. `R` code for fitting the LMAR model is included in this paper's supplementary materials [Cervone et al. (2014)].

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 & 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 (1.23), 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 1.4.5). 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.

1.5 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 1.4, are provided in this section. Point forecasts are discussed in Sections 1.5.1–1.5.3 and interval/distributional forecasts in Section 1.5.4.

1.5.1 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 fraction of time each forecasting method obtains the minimum prediction error among the methods compared. We report these quantities for each of the 8 patients, at forecast windows of 0.2s (6 observations), 0.4s (12 observations), and 0.8s (18 observations) in Table 1.3.

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[‡]. For this reason, we feel that the median (more generally, quantiles of the distribution function for absolute 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 et al. (2011) and Murphy & 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.

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

1.5.2 Quantitative summaries of point forecasts

Summarizing Table 1.3, 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.

1.5.3 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 1.7 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 the figure that the end of the training period for this time series coincided with a dramatic change in the patient's respiration.

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.

		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 1.3: 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.**

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 (1.20), 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,

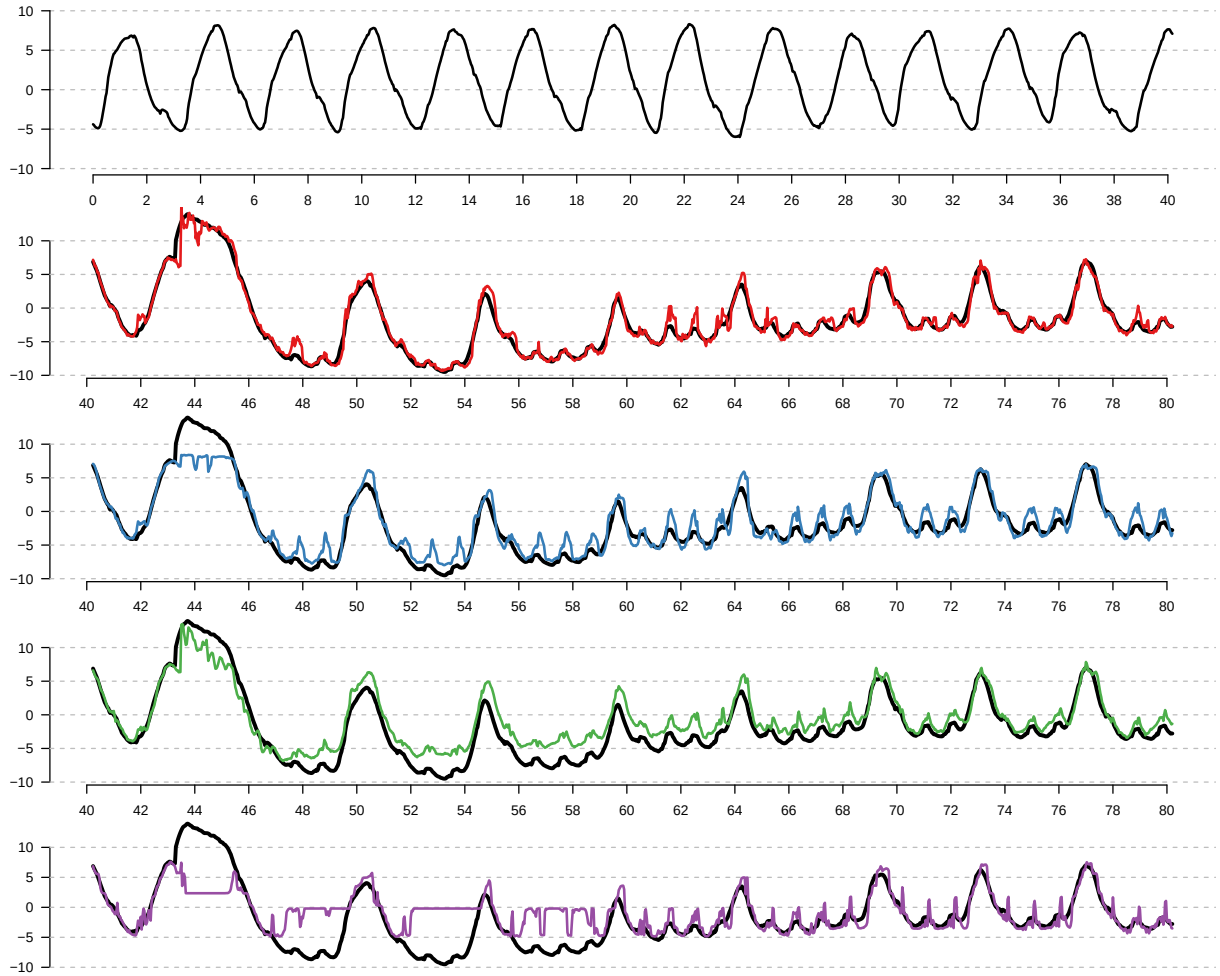


Figure 1.7: Predictions for patient 9, day 3 beam 6 with a forecast window of 0.2s. Location (mm) is the y axis and time (s) the x axis. The 40s training sequence is top, with predictions for the next 40s from LMAR in red, NN in blue, ridge regression in green, and LICORS in purple.

the stochastic location-mixture component in the LMAR prediction adjusts predictions for gradual magnitude shifts in the data.

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 treat-

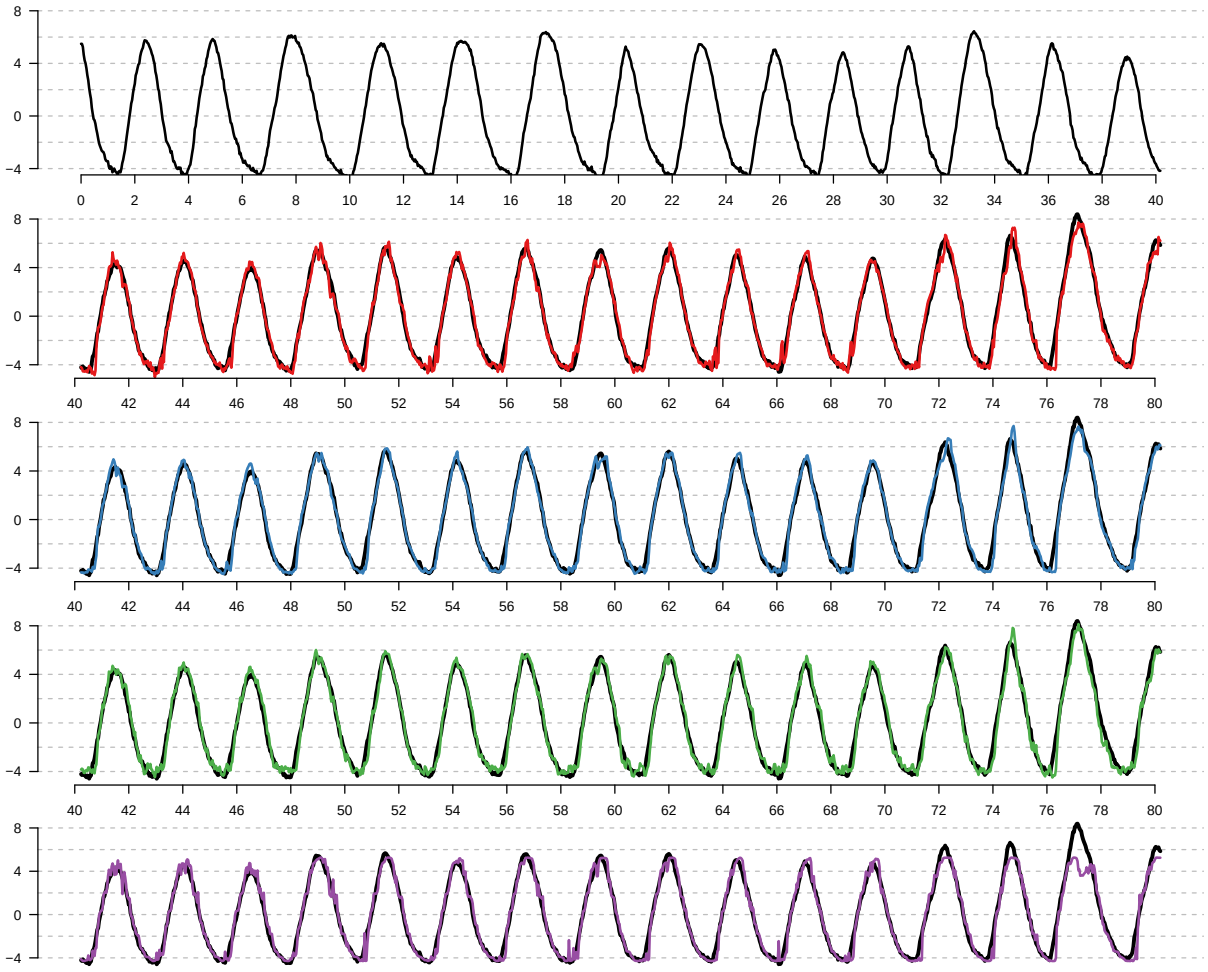


Figure 1.8: Predictions for patient 4, day 6 beam 1 with a forecast window of 0.2s. Location (mm) is the y axis and time (s) the x axis. The 40s training sequence is top, with predictions for the next 40s from LMAR in red, NN in blue, ridge regression in green, and LICORS in purple.

ment. Murphy & 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 et al. (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 1.8 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 net-

works with at least one hidden layer can approximate any continuous function arbitrarily well [Hornik et al. (1989)], including time series prediction.

1.5.4 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 1.4 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 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 1.4 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 1.4 then averages each of these scores over all beams from the same patient). The logarithmic

Patient	Method	0.2s Forecast		0.4s Forecast		0.6s Forecast	
		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 1.4: 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.**

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 et al. (2007)].

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

1.6 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 et al. (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.

2

A Mutiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes

2.1 Introduction

Basketball is a fast-paced sport, free-flowing in both space and time, in which players' actions and decisions continuously impact their teams' prospective game outcomes. Team owners, general managers, coaches, and fans all seek to quantify and evaluate players' contributions to their team's success. However, current statistical models for player evaluation such as "Player Efficiency Rating" [Hollinger (2005)] and "Adjusted Plus/Minus" [Omidiran (2011)] rely on highly reductive summary statistics of basketball games such as points scored, rebounds, steals, assists—

the so-called "box score" summary of the game. Such models reflect the fact that up until very recently, data on basketball games were only available in this low level of resolution. Thus previous statistical analyses of basketball performance have overlooked many of the high-resolution motifs—events not measurable by such aggregate statistics—that characterize basketball strategy. For instance, traditional analyses cannot estimate the value of a clever move that fools the defender, or the regret of skipping an open shot in favor of passing to a heavily defended teammate. The advent of player tracking data in the NBA, coupled with the appropriate inferential framework, has provided an opportunity to fill this gap.

In 2013 the National Basketball Association (NBA), in partnership with data provider STATS LLC, installed optical tracking systems in the arenas of all 30 teams in the league. The systems track the exact two-dimensional locations of every player on the court (as well as the three-dimensional location of the ball) at a resolution of 25Hz, yielding over 1 billion space-time observations over the course of a full season.

In this paper, we present a framework for modeling NBA tracking data that targets inferences at the resolution of this exciting new data. Specifically, we estimate the expected number of points the offense will score on a particular possession conditional on that possession's evolution up to time t. We term this quantity *expected possession value* (EPV). EPV acts like the stock ticker of an NBA possession in providing an instantaneous summary of the possession's value given all available information. Ideally, EPV mirrors the intuition of coaches and basketball strategists by attaching value to specific spatial positionings of players and personnel configurations. For instance, EPV may be high when a good shooter has an open look at the basket, but low when the ballcarrier is heavily defended far from the basket without any clear passing options. We may also, for example, see EPV as roughly constant while players are passing the ball around the perimeter, far from the basket, but then spike upwards as a player drives through an opening in the defense towards the basket. By monitoring how the EPV curves for various possessions respond to player decisions, analysts can evaluate players and strategies in real time, continuously throughout the entire course of any possession.

Our paper focuses specifically on modeling and calculating EPV curves by viewing basketball possessions as realizations of endpoint-valued stochastic processes; that is, we assume possessions

evolve probabilistically over space and time until reaching some endpoint (e.g., a made basket, or a turnover) with an observable point value. Such frameworks have been used for in-game points/run expectancy models in other sports—for instance, baseball [Bukiet et al. (1997)] and football [Burke (2010)]—yet ours is the first treatment of a process that is continuous in both space and time. To handle this additional complexity, we introduce the idea of multiresolution transitions, which distinguish between the continuous evolution of all players' positions (microtransitions) and motifs or events that unfold over longer time scales (macrotransitions), such as passes and shot attempts. We show that multiresolution transitions crucially allow conditioning that is both interpretable and computationally tractable.

While our methodology is motivated by basketball, we believe that this research can serve as an informative case study for analysts working in other application areas where continuous monitoring data are becoming widespread, including traffic monitoring [Ihler et al. (2006)], surveillance, and digital marketing [Shao & Li (2011)], as well as other sports such as soccer and hockey [Thomas et al. (2013)]. As such, we treat our particular approach to the player tracking problem in basketball with as much generality as possible, and hope that aspects of our methodology may also find use in these other contexts.

Section 2.2 formally defines EPV within the context of a stochastic process for basketball. Section 2.3 introduces multiresolution transitions and discusses the assumptions and conditioning statements that make EPV calculations tractable as averages over future paths of a stochastic process. The models for macro- and microtransitions are discussed in Sections 2.4 and 2.5, respectively, and represent the inferential component of our model, as transition probabilities rely on players' decision-making tendencies in various spatial and situational circumstances. Section 2.6 discusses Monte Carlo computation for EPV curves, given parameters for the multiresolution transition models, with some results from actual NBA possessions highlighted in Section 2.7. Directions for further work are discussed in Section 2.8.

The core of our paper is a high-level overview of the EPV estimation problem and our model, focusing on the stochastic process, multiresolution approach to EPV, and highlighting the types of inferences EPV and the associated multiresolution transition models provide basketball analysts. A substantial appendix follows the discussion section, in which we discuss specific details of

our implementation for EPV estimation, including choices of prior distributions for model parameters, and computational methods for obtaining approximate inferences from such a large, high dimensional data set. More than just a guide to reproducing our results, the appendix highlights novel contributions to estimation problems in spatiotemporal data. For instance, our macrotransition model employs a family of prior distributions that shares information across space as well as across players, representing hierarchical Gaussian processes in a computationally tractable parameterization.

2.2 Expected Possession Value

Let Ω represent the space of all possible basketball possessions in full detail, with $\omega \in \Omega$ describing the full path of a particular possession. Every basketball possession that we consider here results in 0, 2, or 3 points scored for the offense, denoted $X(\omega) \in \{0, 2, 3\}$. It is possible for the offense to score exactly 1 or 4 points as well if a foul occurs and free throws are made, but we exclude fouls and free throws from Ω due to limitations in our data. For any possession path ω, we denote by $Z(\omega)$ the optical tracking timeseries generated by this possession so that $Z_t(\omega) \in \mathcal{Z}$, $t > 0$, is a "snapshot" of the tracking data exactly t seconds from the start of the possession ($t = 0$). $\mathcal{Z}$ is a high dimensional space that includes (x, y) coordinates for all 10 players on the court, (x, y, z) coordinates for the ball, summary information such as which players are on the court and what the game situation is (game location, score, time remaining, etc.), and event annotations that are observable in real time, such as a turnover occurring, a pass, or a shot being attempted and the result of that attempt.

This intuitive view of Ω as a sample space of possession paths provides the formalism for defining EPV in probabilistic terms. We define $Z(\omega)$ to be a stochastic process, and likewise, define $Z_t(\omega)$ for each $t > 0$ as a random variable in $\mathcal{Z}$. $Z(\omega)$ provides the natural filtration $\mathcal{F}_t^{(Z)} = \sigma(\{Z_s^{-1} : 0 \leq s \leq t\})$, which intuitively represents all information available from the optical tracking data for the first t seconds of a possession. Because the point outcome of a possession (X) is apparent from observing $Z(\omega)$ for a sufficiently long time, X is $\mathcal{F}_\infty^{(Z)}$- measureable, and we can define EPV as the expected value of the number of points scored for the possession (X) given all available data up to time t ($\mathcal{F}_t^{(Z)}$):

Definition The *expected possession value*, or EPV, at time $t \geq 0$ during a possession is $\nu_t = \mathbb{E}[X|\mathcal{F}_t^{(Z)}]$.

Remark Except when introducing new summaries of the possession sample space Ω, we will omit the dependence on ω when writing function- or scalar-valued random variables, e.g., Z and Z_t instead of $Z(\omega)$ and $Z_t(\omega)$.

2.2.1 Possession Case Study

To illustrate the behavior of EPV, we consider the estimated EPV curve from a specific Miami Heat possession against the Brooklyn Nets from the second quarter of a game on November 1, 2013. This possession was chosen arbitrarily among those during which LeBron James (widely considered the best NBA player as of 2014) handles the ball. This is presented here in order to describe and motivate the object of our estimation; the methodology itself will be discussed in the sections that follow.

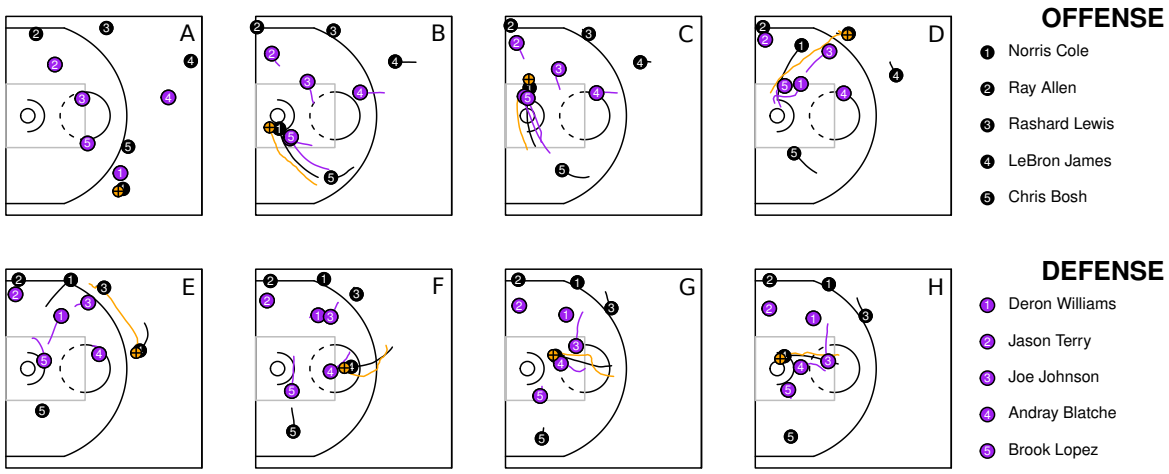


Figure 2.1: Visualization of Miami Heat possession against Brooklyn Nets. Norris Cole wanders into the perimeter (A) before driving toward the basket (B). Instead of taking the shot, he runs underneath the basket (C) and eventually passes to Rashard Lewis(D), who promptly passes to LeBron James (E). After entering the perimeter (F), James slips behind the defense (G) and scores an easy layup (H).

In this particular possession, diagrammed in Figure 2.1, point guard Norris Cole begins with possession of the ball crossing the halfcourt line (panel A). After waiting for his teammates to arrive in the offensive half of the court, Cole wanders gradually into the perimeter (inside the three

point line), before attacking the basket through the left post. He draws two defenders, and while he appears to beat them to the basket (B), instead of attempting a layup he runs underneath the basket through to the right post (C). He is still being double teamed and at this point passes to Rashard Lewis (D), who is standing in the right wing three position and being defended by Joe Johnson. As Johnson closes, Lewis passes to LeBron James, who is standing about 6 feet beyond the three point line and drawing the attention of Andray Blatche (E). James wanders slowly into the perimeter (F), until just behind the free throw line, at which point he breaks towards the basket. His rapid acceleration (G) splits the defense—Joe Johnson had also begun defending James as he entered the perimeter—and gains him a clear lane to the basket. He successfully finishes with a layup (H), providing the Heat two points.

Plotting the EPV curve for this possession (Figure 2.2), we see several moments when the expected point yield of the possession, given its history, changes dramatically. Beginning around 0.99, the EPV first rises as Cole drives toward the basket, starting around 5 seconds into the possession. It continues rising until peaking at around 1.34 when Cole is right in front of the basket. As Cole dribbles past the basket (and his defenders continue pursuit), however, EPV falls rapidly, bottoming out at 0.77 before "resetting" to 1.00 with the pass to Rashard Lewis. The EPV increases slightly to 1.03 when the ball is then passed to James. As EPV is sensitive to small changes in players' exact locations, we see EPV rise slightly as James approaches the three point line and then dip slightly as he crosses it. Shortly afterwards, EPV rises suddenly as James breaks towards the basket, eluding the defense, and continues rising until he is beneath the basket, when an attempted layup boosts the EPV from 1.52 to 1.62.

We will revisit this example possession in more detail in Section 2.7. For now, it serves to highlight the behavior and potential applications of EPV. Any point on the curve in Figure 2.2 provides an unbiased estimate of the possession's eventual point total as a function of its entire history, including but not limited to the identity and precise location of the ballcarrier, and those of his teammates and the defense. We see EPV rise and fall—at times dramatically—as players move the ball through the court, pass, attempt shots, and gain or lose separation from the defense.

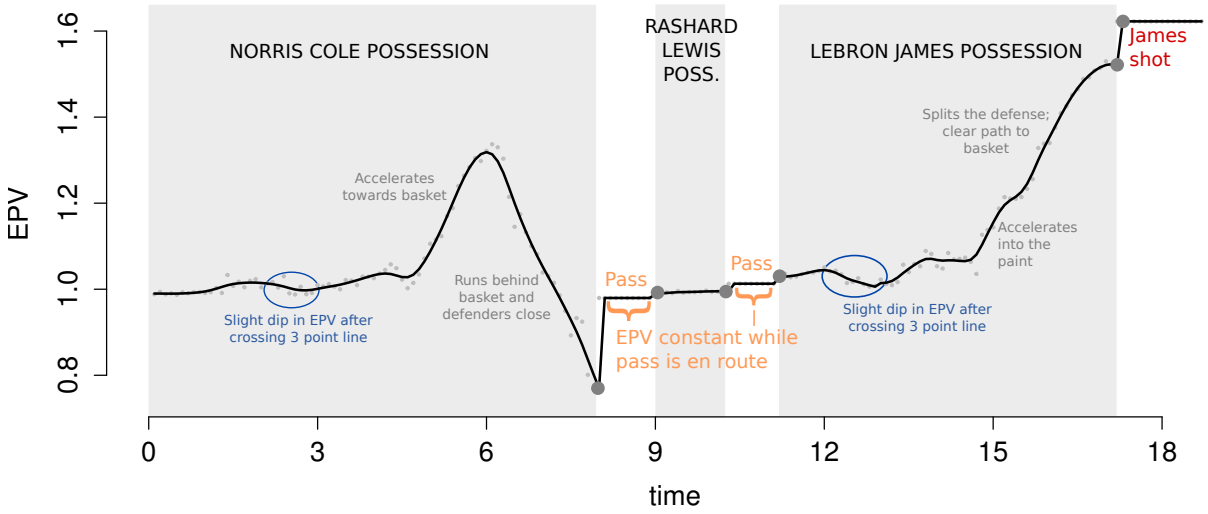


Figure 2.2: Estimated EPV over time for the possession shown in Figure 2.1. Changes in EPV are induced by changes in players' locations and dynamics of motion; macrotransitions such as passes and shot attempts produce immediate, sometimes rapid changes in EPV. The black line slightly smooths EPV evaluations at each time point (gray dots), which are subject to Monte Carlo error.

2.2.2 Stochastic Consistency

EPV is a theoretical quantity associated with the true distribution of possession paths Z. As with all statistical estimands, the definition of EPV does not restrict the method that an investigator can use to estimate it. Indeed, because EPV is simply a conditional expectation, a number of marginal regression or classification methods that map features from $\mathcal{F}_t^{(Z)}$ to the outcome space (either $[0, 3]$ or $\{0, 2, 3\}$) can be tempting options. While our data do not constitute the independent input/output pairs characteristic of regression (each possession outcome X has a series of inputs Z_t), a properly specified regression model would nevertheless consistently estimate ν_t as a function of features of $\mathcal{F}_t^{(Z)}$.

Regression estimates, however, lack the stochastic consistency inherent to the probabilistic formulation of EPV. ν is an instance of the "Doob martingale" with respect to the filtration $\mathcal{F}^{(Z)}$—that is, a sequence of conditional expectations of the same end quantity X taken with respect to increasing elements of the filtration $\mathcal{F}^{(Z)}$. Thus, EPV evaluated at a particular time t can be represented as the expected EPV evaluated at a later time $s > t$: $\mathbb{E}[\nu_s | \mathcal{F}_t^{(Z)}] = \nu_t$. An important consequence of this property is that no situation can systematically yield downstream events with consistently higher or lower EPV. This is not trivial, as regression and other marginal methods

that estimate each point of the EPV curve in isolation do not guarantee this coherence. For example, under a marginal estimation scheme yielding a sequence of estimates $\hat{\nu}_t$, it is possible for a Simpson's paradox to arise where for some t and Δ, $\hat{\nu}_{t+\Delta} < \hat{\nu}_t$ no matter what occurs at time t. On the other hand, our methodology, which explicitly computes the integral in (2.1) with respect to a model for the whole process Z, maintains stochastic consistency.

A simple model that does provide stochastic consistency is discretizing Z_t and modeling it as a homogeneous Markov chain. Markov chains have been commonly used for modeling final outcomes conditional on observed progress in other sports, such as in-game win probability in baseball [Bukiet et al. (1997); Yang & Swartz (2004)] and in-possession point totals in football [Goldner (2012)]. In both these examples, the data is naturally discrete in space and time; however, our data Z_t is essentially continuous in space and time (we do observe data only at regular intervals of 1/25 second). Discretizing this process forces the investigator to trade off between the smoothness and level of detail captured in the process ν_t (having a larger state space), and the ease of estimation and computation. While a state space with a huge number of states may in theory provide smooth, stochastically consistent estimated EPV curves, the associated transition probabily matrix would be very difficult to estimate since a large number of states induces sparsity in the observed transition matrix. Moreover, computing expected values of a homogeneous Markov chain requires solving a linear system of the same dimension as the number of states, which is a cubic-time operation; this would be computationally infeasible for a huge state space.

Our methodology for estimating EPV, leveraging the idea of multiresolution transitions, largely avoids this tradeoff and offers precise, stochastically consistent EPV curve estimates. Generating such estimates is computationally demanding, but feasible given modern computing infrastructure and inference techniques.

2.3 Multiresolution modeling

The stochastic process approach to estimating EPV requires that we integrate over the distribution of future paths the current possession can take. Letting $T(\omega)$ denote the time at which a

possession following path ω ends*, the possession's point total is a deterministic function of the full resolution data at this time, $X(\omega) = h(Z_{T(\omega)}(\omega))$. Thus, evaluating EPV amounts to integrating over the joint distribution of (T, Z_T):

$$\begin{aligned}\nu_t = \mathbb{E}[X|\mathcal{F}_t^{(Z)}] &= \int_\Omega X(\omega)\mathbb{P}(d\omega|\mathcal{F}_t^{(Z)}) \\ &= \int_t^\infty \int_{\mathcal{Z}} h(z)\mathbb{P}(Z_s = z|T = s, \mathcal{F}_t^{(Z)})\mathbb{P}(T = s|\mathcal{F}_t^{(Z)})dzds. \qquad (2.1)\end{aligned}$$

Note that we use probability notation $\mathbb{P}(\cdot)$ somewhat heuristically, as $\mathbb{P}(T = s|\mathcal{F}_t^{(Z)})$ is a density with respect to Lebesgue measure, while Z_s mixes both discrete (annotations) and continuous (locations) components. The best way to integrate (2.1) is by simulating future paths of the full resolution data with a transition kernel $\mathbb{P}(Z_{t+\epsilon}|\mathcal{F}_t^{(Z)})$ until the simulated possession ends by reaching an observed point outcome. Such simulations provide a Monte Carlo estimate of (2.1). A model for this transition kernel requires a novel mixture of components for both the continuous spatial evolution of players, as well as their discrete decisions and ball movements.

Our approach is to model the possession process Z at two separate levels of resolution. In addition to modeling the short-term evolution of Z at full resolution, we simultaneously model a coarsened view of the process Z that is discrete in space and continuous in time. We combine these models in a multiresolution conditioning scheme that yields EPV calculations that are both computationally tractable (using Monte Carlo) and interpretable in terms of relevant basketball motifs.

2.3.1 A Coarsened Process

A key component of our approach is a coarsening of the full-resolution data Z that preserves the characteristic dynamics of basketball play while shedding fine-resolution detail. For all time $0 < t \leq T$ during a possession, assume C_t summarizes the "state" of the possession, such that $C_t \in \mathcal{C}$ for some finite set $\mathcal{C}$. We populate the states $c \in \mathcal{C}$ with summaries of the full resolution data so that transitions between these states represent meaningful events in a basketball possession. We

*The time of a possession is bounded, even for pathological examples, by the 12-minute length of a quarter; yet we do not leverage this fact and simply assume that possession lenghts are almost surely finite.

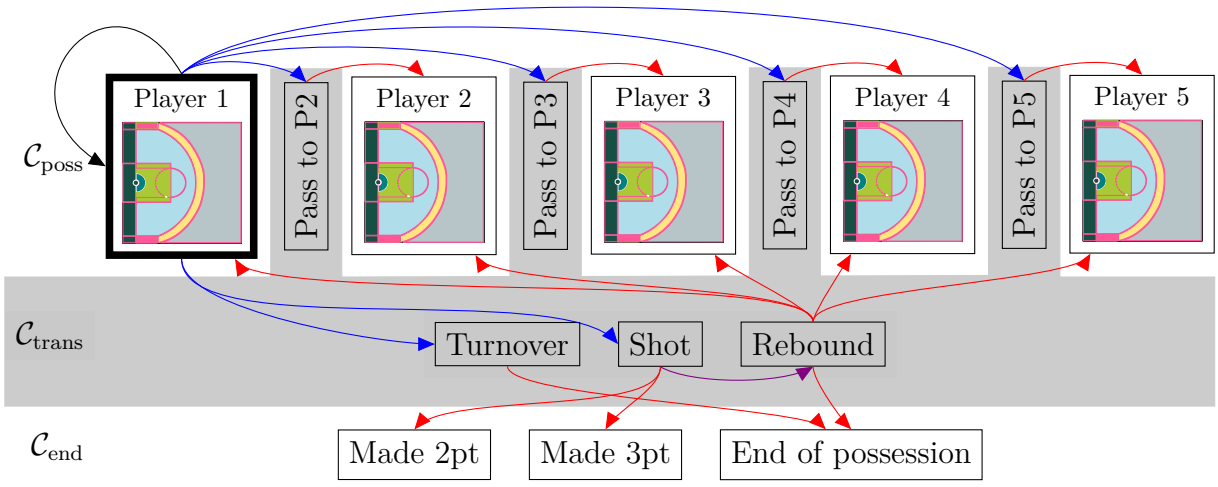


Figure 2.3: Schematic of the coarsened possession process C, with states (rectangles) and possible state transitions (arrows) shown. The unshaded states in the first row compose $\mathcal{C}_{poss}$. Here, states corresponding to distinct ballhandlers are grouped together (Player 1 through 5), and the discretized court in each group represents the player's coarsened position and defended state. The gray shaded rectangles are macrotransition states $\mathcal{C}_{trans}$, while the rectanbles in the third row represent the end states $\mathcal{C}_{end}$. Blue arrows are the beginnings of macrotransition that can result when a player (WLOG Player 1 in this figure) possesses the ball. Red arrows are macrotransition exits. The purple arrow (between Shot and Rebound) carries the same macrotransition (beginning with the shot attempt) into the rebound state (from which it will exit). The black arrow is a microtransition, as the ballcarrier is unchanged.

illustrate the coarsened process C in Figure 2.3.

First, there are 3 "bookkeeping" states, denoted $\mathcal{C}_{\text{end}}$, that categorize the end of the possession, so that $C_T \in \mathcal{C}_{\text{end}}$ and for all $t < T, C_t \notin \mathcal{C}_{\text{end}}$ (shown in the bottom row of Figure 2.3). These are $\mathcal{C}_{\text{end}}$ ={made 2 pt, made 3 pt, end of possession}. These three states have associated point values— 2 points for a made 2 point shot, 3 points for a made 3 point shot, and 0 points for the generic possession end state (which can be reached by turnovers, defensive rebounds, etc.). Thus, there is a map $h : \mathcal{C}_{\text{end}} \to \{0, 2, 3\}$ allowing us to rewrite the EPV equation in terms of the coarsened process: $\nu_t = \mathbb{E}[h(C_T)|\mathcal{F}_t^{(Z)}]$.

Next, whenever a player possesses the ball at time t, we assume C_t = (ballcarrier ID at t) × (court region at t) × (defended at t), having defined seven disjoint regions of the court and classifying a player as defended at time t by whether there is a defender within 5 feet of him. The possible values of C_t, if a player possesses the ball at time t, thus live in $\mathcal{C}_{\text{poss}}$ = {player ID} ×

$\{\text{region ID}\} \times \{\mathbf{1}[\text{defended}]\}$. These states are represented by the unshaded portion of the top row of Figure 2.3, where the differently colored regions of the court diagrams reveal the court space discretization.

Finally, if a player has initiated an annotated action currently in progress, we define C_t to take a "transition" state value. These states encapsulate constrained motifs in a possession, for example, when the ball is in the air traveling between players in a pass attempt. Explicitly, denote $\mathcal{C}_{\text{trans}} = \{\text{shot attempt from } c \in \mathcal{C}_{\text{poss}}, \text{ pass attempt toward } c' \in \mathcal{C}_{\text{poss}} \text{ from } c \in \mathcal{C}_{\text{poss}}, \text{ turnover in progress, rebound in progress}\}$ (listed in the gray shaded portions of Figure 2.3). These transition states carry information about the possession path, such as the most recent ballcarrier, or the target of the pass, while the ball is in the air during shot attempts and passes.

Note that due to limitations of the data, this construction of $\mathcal{C} = \mathcal{C}_{\text{poss}} \cup \mathcal{C}_{\text{trans}} \cup \mathcal{C}_{\text{end}}$ excludes several notable basketball events, such as fouls and violations, balls going out of bounds without a change of possession, and other stoppages of play. In the case of turnovers, the event labels do not discriminate among steals, intercepted passes, or lost balls out of bounds, thus we treat this as a single category.

2.3.2 Multiresolution Conditioning

When the coarsened process C_t transitions from a state in $\mathcal{C}_{\text{poss}}$ to one in $\mathcal{C}_{\text{trans}}$, we call this transition between coarsened states a *macrotransition*.

Definition If $C_t \in \mathcal{C}_{\text{poss}}$ and $C_{t+\epsilon} \in \mathcal{C}_{\text{trans}}$, then $C_t \to C_{t+\epsilon}$ is a *macrotransition*.

Macrotransitions, which include all ball movements (passes, shot attempts, turnovers), mark large-scale shifts that form the basis of offensive basketball play. The term carries a double meaning, as a macrotransition describes both a transition among states in our coarsened process, $C_t \to C_{t+\epsilon}$, as well as a transition of ballcarrier identity on the basketball court. By construction, for a possession that is in a state in $\mathcal{C}_{\text{poss}}$ to proceed to a state in $\mathcal{C}_{\text{end}}$ or a state in $\mathcal{C}_{\text{poss}}$ corresponding to a different ballhandler, a macrotransition must occur as possession passes through a transition state in $\mathcal{C}_{\text{trans}}$ (see possible transition paths illustrated in Figure 2.3).

This structure reveals that at any time t during a possession, we are guaranteed to observe the *exit state* of a future (or current, if $C_t \in \mathcal{C}_{\text{trans}}$) macrotransition. Specifically, let $\delta = \min\{s : s > t, C_{s-\epsilon} \in \mathcal{C}_{\text{trans}} \text{ and } C_s \notin \mathcal{C}_{\text{trans}}\}$ denote the time the possession reaches the state *after* the next (or current, if $C_t \in \mathcal{C}_{\text{trans}}$) macrotransition after time t. Thus, if the possession is currently in a macrotransition, δ is the first time at which a new possession or end state is occupied (ending the macrotransition), while if a player currently possesses the ball, δ is the time at which the possession reaches the exit state of a future macrotransition. δ is a bounded stopping time, so we can condition on C_δ to rewrite EPV (2.1) as

$$\nu_t = \sum_{c \in \mathcal{C}} \mathbb{E}[h(C_T)|C_\delta = c, \mathcal{F}_t^{(Z)}]\mathbb{P}(C_\delta = c|\mathcal{F}_t^{(Z)}). \tag{2.2}$$

It is helpful to expand the second term in (2.2), $\mathbb{P}(C_\delta = c|\mathcal{F}_t^{(Z)})$, by conditioning on the start of the macrotransition that corresponds to the exit state C_δ. Denote $M(t)$ as the event that a macrotransition begins in $(t, t+\epsilon]$, and let $\tau = \min\{s : s > t, M(s)\}$ be the time at which the macrotransition ending in C_δ begins. Thus, τ and δ bookend the times during which the possession is in the next (or current, but ongoing) macrotransition, with C_τ being the state in $\mathcal{C}$ immediately *prior* to the start of this macrotransition and C_δ the state immediately succeeding it. Like δ, at any time $t < T$, τ is a bounded stopping time; however, note that if a macrotransition is in progress at time t then $\tau < t$, and, having been observed, τ has a degenerate distribution. Defining τ allows us to write:

$$\begin{aligned}\mathbb{P}(C_\delta = c|\mathcal{F}_t^{(Z)}) = \sum_{c \in \mathcal{C}} \int_t^\infty \int_{\mathcal{Z}} &\mathbb{P}(C_\delta = c|M(\tau), Z_\tau = z, \tau = s, \mathcal{F}_t^{(Z)}) \\ &\times \mathbb{P}(M(\tau), Z_\tau = z, \tau = s|\mathcal{F}_t^{(Z)}) dz ds.\end{aligned} \tag{2.3}$$

We make one additional expansion to the terms we have introduced for calculating EPV. The second factor in (2.3), $\mathbb{P}(M(\tau), Z_\tau = z, \tau = s|\mathcal{F}_t^{(Z)})$, models the location and time of the next macrotransition—implicitly averaging over the intermediate path of the possession in the process. This is the critical piece of our multiresolution structure that connects the full-resolution process Z to the coarsened process C, and the component of our model that fully utilizes multiresolution

conditioning. We expand this term using our macro- and microtransition models.

Definition The *macrotransition model* is $\mathbb{P}(M(t)|\mathcal{F}_t^{(Z)})$.

Definition The *microtransition model* is $\mathbb{P}(Z_{t+\epsilon}|M(t)^c, \mathcal{F}_t^{(Z)})$, where $M(t)^c$ is the complement of $M(t)$. *Microtransitions* are instantaneous changes in the full resolution data $Z_t \to Z_{t+\epsilon}$ over time windows where a macrotransition is not observed; thus, only location components (and not event annotations) change from Z_t to $Z_{t+\epsilon}$.

Multiresolution transition models allow us to sample from $\mathbb{P}(\tau, Z_\tau|\mathcal{F}_t^{(Z)})$, enabling Monte Carlo evaluation of (2.3). The basic idea is that we use the macrotransition model to draw from $\mathbb{P}(M(t)|\mathcal{F}_t^{(Z)})$ and if $M(t)^c$ and no macrotransition occurs in $(t, t+\epsilon]$, we use the microtransition model to draw from $\mathbb{P}(Z_{t+\epsilon}|M(t)^c, \mathcal{F}_t^{(Z)})$. Iterating this process, we alternate draws from the macro- and microtransition models until observing (τ, Z_τ)—of course, this also yields $M(\tau)$ as a consequence of our definition of τ. Parametric forms for these macro- and microtransition models are dicussed expcility in Sections 2.4 and 2.5 respectively, while Section 2.6 provides additional details on the Monte Carlo integration scheme.

Expanding EPV by conditioning on intermediate values in principle does not ease the problem of its evaluation. However several of the components we have introduced motivate reasonable conditional independence assumptions that simplify their evaluation. Only by writing EPV as an average over additional random variables defined in the probability space of our possession can we articulate such assumptions and leverage them to compute EPV.

2.3.3 Conditional Independence Assumptions

Our expansions of $\nu_t = \mathbb{E}[h(C_T)|\mathcal{F}_t^{(Z)}]$ introduced in the previous subsection (2.2)–(2.3) express EPV in terms of three probability models:

$$\nu_t = \sum_{c \in \mathcal{C}} E[h(C_T)|C_\delta = c, \mathcal{F}_t^{(Z)}] \left(\int_t^\infty \int_{\mathcal{Z}} \mathbb{P}(C_\delta = c|M(\tau), Z_\tau = z, \tau = s, \mathcal{F}_t^{(Z)}) \right.$$
$$\left. \times \mathbb{P}(M(\tau), Z_\tau = z, \tau = s|\mathcal{F}_t^{(Z)}) dz ds \right). \tag{2.4}$$

The multiresolution transition models sample from $\mathbb{P}(M(\tau), Z_\tau, \tau | \mathcal{F}_t^{(Z)})$, eliminating the need to evaluate the third term in (2.4) explicitly when computing ν_t via Monte Carlo. The second term in (2.4) is actually quite easy to work with since C_δ is categorical, and given Z_τ the space of possible values it can take is relatively small. This is due to the manner in which macrotransitions constrain the spatiotemporal evolution of the possession. Given Z_τ, we can obtain the location and separation from the defense of all four possible pass recipients given a pass in $(\tau, \tau + \epsilon]$, so only a subset of states in $\mathcal{C}_{\text{poss}}$ are possible for C_δ. Similarly, if a shot attempt occurs in this time window, Z_τ indicates whether a successful shot would yield 2 or 3 points, further subsetting the possible values of C_δ. Modeling C_δ thus reduces to predicting the type of macrotransition corresponding to $M(\tau)$—a pass, shot attempt, or turnover. We discuss this in Section 2.4 in the context of our macrotransition model.

The first term in (2.4), $E[h(C_T)|C_\delta = c, \mathcal{F}_t^{(Z)}]$ provides the expected point value of the possession given the (coarsened) result of the next macrotransition. Prima facie, this term seems as difficult to evaluate as it has the same essential structure as EPV itself, requiring integration over the future trajectory of the possession after time δ. However, we make a key assumption that frees subsequent evolution of the possession, after time δ, from dependence on the full-resolution history $\mathcal{F}_t^{(Z)}$:

$$\mathbb{E}[h(C_T)|C_\delta, \mathcal{F}_t^{(Z)}] = \mathbb{E}[h(C_T)|C_\delta]. \tag{2.5}$$

This assumption is intuitive for two reasons. First, by constraining the possession to follow a restricted spatiotemporal path, it is reasonable to assume that the macrotransition exit state itself contains sufficient information to characterize the future evolution of the system. Secondly, because macrotransitions play out over much longer timescales than the resolution of the data (i.e., several seconds, as opposed to 1/25th of a second), it is reasonable to assume that fine-scale spatial detail before the start of the macrotransition has been "mixed out" by the time the macrotransition ends.

An additional, reasonable conditional independence assumption is that the coarsened state sequence $C_t, t > 0$ is marginally a semi-Markov process; that is, denoting $\mathcal{F}_t^{(C)} = \sigma(\{C_s^{-1}, 0 \leq s \leq t\})$ as the history of the coarsened process, for all $t' > t$ and $c \in \mathcal{C}$, we assume $\mathbb{P}(C_{t'} =$

$c|\mathcal{F}_t^{(C)}) = \mathbb{P}(C_{t'} = c|C_t)$. A semi-Markov process generalizes a continuous time Markov Chain in that sojourn times need not be exponentially distributed. We associate with this semi-Markov process an embedded discrete, homogeneous Markov Chain: denote $C^{(0)}, C^{(1)}, \ldots, C^{(K)}$ as the sequence of consecutive states $c \in \mathcal{C}$ visited by C_t during the possession $0 < t \leq T$. Thus, $C^{(K)} = C_T$, and K records the length of the possession in terms of the number of transitions between states in $\mathcal{C}$, which like T is random.

Combining these assumptions, the first term in (2.4), $\mathbb{E}[h(C_T)|C_\delta, \mathcal{F}_t^{(Z)}]$, can be computed easily from the transition probability matrix of the homogeneous Markov chain embedded in C_t. As $C^{(K)}$ is an absorbing state, ending the possession, we can rewrite (2.5) as $\mathbb{E}[h(C^{(K)})|C_\delta]$. This is easily obtained by solving a linear system of equations deriving from the transition probability matrix of $C^{(0)}, C^{(1)}, \ldots, C^{(K)}$. Estimating this transition probability matrix is also discussed in Section 2.4, where we show that it actually derives from the macrotransition model.

Compared to using discrete, homogeneous Markov Chains alone to calculate EPV, the multiresolution approach we take ultimately leverages much of the same computational advantages while remaining attenuated to the full-resolution data, responding smoothly as the possession evolves over space and time.

2.4 Macrotransition model

Macrotransitions play a fundamental role in our EPV framework. Intuitively, they represent the strategies, schemes, and decision-making that characterize basketball offense. Mathematically, macrotransitions are part of our multiresolution conditioning scheme used to evaluate EPV at any time during a possession given its history. Introduced in the previous section, the macrotransition model is $\mathbb{P}(M(t)|\mathcal{F}_t^{(Z)})$. More generally, we consider a family of macrotransition models $\mathbb{P}(M_j(t)|\mathcal{F}_t^{(Z)})$, where j indexes the type of macrotransition corresponding to $M(t)$. At any given moment when a player possesses the ball, there are six possible categories of macrotransition, corresponding to 4 pass options, a shot attempt, or a turnover, which we index by $j \in \{1, \ldots, 6\}$. Without loss of generality, assume $j \leq 4$ correspond to pass events, $j = 5$ is a shot attempt and $j = 6$ a turnover. Thus, $M_j(t)$ is the event that a macrotransition of type j begins in the time window $(t, t + \epsilon]$, and $M(t) = \bigcup_{j=1}^{6} M_j(t)$.

We now introduce the parameterization of the macrotransition models $\mathbb{P}(M_j(t)|\mathcal{F}_t^{(Z)})$, and also discusses how other components of our EPV equation (2.4) derive from these models.

2.4.1 Macrotransition Entry Model

As $M_j(t)$ denotes the start of a macrotransition in $(t, t+\epsilon]$, we refer to $\mathbb{P}(M_j(t)|\mathcal{F}_t^{(Z)})$, for all j, as the macrotransition entry model. This is specified using competing risks [Prentice et al. (1978)]: assuming player ℓ possesses the ball at time $t > 0$ during a possession, then denote

$$\lambda_j^\ell(t) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(M_j(t)|\mathcal{F}_t^{(Z)})}{\epsilon} \tag{2.6}$$

as the hazard for macrotransition j at time t, or the cause-specific hazard. As events $M_1(t), \ldots, M_6(t)$ are disjoint, it follows that the total macrotransition hazard is the sum of the cause-specific hazards,

$$\lim_{\epsilon \to 0} \frac{\mathbb{P}(M(t)|\mathcal{F}_t^{(Z)})}{\epsilon} = \sum_j \lambda_j(t)$$

We assume the cause-specific hazards are log-linear,

$$\log(\lambda_j^\ell(t)) = [\mathbf{W}_j^\ell(t)]'\boldsymbol{\beta}_j^\ell + \xi_j^\ell\left(\mathbf{z}_\ell(t)\right) + \left(\tilde{\xi}_j^\ell\left(\mathbf{z}_j(t)\right)\mathbf{1}[j \leq 4]\right), \tag{2.7}$$

where $\mathbf{W}_j^\ell(t)$ is a $p_j \times 1$ vector of time-varying covariates, $\boldsymbol{\beta}_j^\ell$ a $p_j \times 1$ vector of coefficients, $\mathbf{z}_\ell(t)$ is the ballcarrier's 2D location on the court (denote the court space $\mathbb{S}$) at time t, and $\xi_j^\ell : \mathbb{S} \to \mathbb{R}$ is a mapping of the player's court location to an additive effect on the log-hazard, providing spatial variation. The last term in (2.7) only appears for pass events ($j \leq 4$) to incorporate the location of the receiving player for the corresponding pass: $\mathbf{z}_j(t)$ (which slightly abuses notation) provides his location on the court at time t, and $\tilde{\xi}_j^\ell$, analogously to ξ_j^ℓ, maps this location to an additive effect on the log-hazard. All spatial effects ξ are assumed to be realizations of Gaussian processes; a detailed discussion of the structure and estimation of these spatial effects is included in Appendix A.

The macrotransition model (2.6)–(2.7) represents the ballcarrier's decision-making process as an interpretable function of the unique basketball predicaments he faces. For example, in consid-

ering the hazard of a shot attempt, the time-varying covariates ($\mathbf{W}_j^\ell(t)$) we use are the distance between the ballcarrier and his nearest defender (transformed as $\log(1+d)$ to moderate the influence of extremely large or small observed distances), an indicator for whether the ballcarrier has dribbled since gaining possession, and a constant representing a baseline shooting rate (this is not time-varying)[†]. The spatial effects ξ_j^ℓ reveal locations where player ℓ is more/less likely to attempt a shot in a small time window, holding fixed the time-varying covariates $\mathbf{W}_j^\ell(t)$. Such spatial effects (illustrated in Figure 2.4) are well-known to be nonlinear in distance from the basket and asymmetric about the angle to the basket [Miller et al. (2013)].

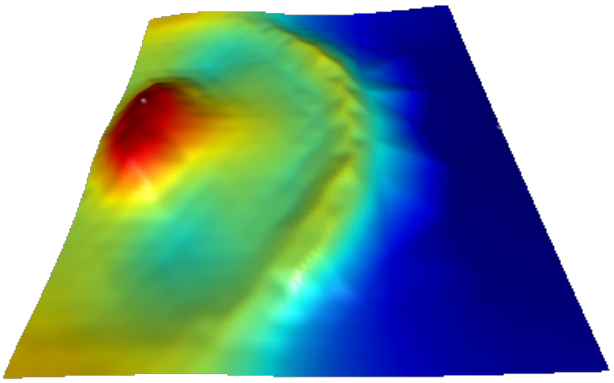


Figure 2.4: Estimated ξ_j^ℓ for LeBron James' shot-taking hazard ($j = 5$). For each location on the court, the color and height of the surface are proportional to the additive effect on the log hazard of attempting a shot at that location.

For pass events, the time-varying covariates, their coefficients, and the spatial effect ξ_j^ℓ vary for the same ballcarrier (ℓ) across his different passing options $j = 1, \ldots, 4$. This reflects the fact that pass events between two players depend on those two players' positions and roles on the team. For instance, a point guard will pass to a center in different situations than those in which he passes to a shooting guard. Thus in some sense, we are modeling the pass events independently for every passer-receiver pair; main effects from the passer's and receiver's identities are not explicitly modelled, though hierarchical models allow information sharing among the different pass models associated with player ℓ (Appendix A introduces such hierarchical structure, which we use in our parameter estimation). It is conceptually useful to think of $(\xi_j^\ell, \tilde{\xi}_j^\ell) : \mathbb{R}^4 \to \mathbb{R}$ as jointly providing a single spatial effect for the 4D location of the passer/receiver pair, and of

[†]Full details on all covariates used for all macrotransition types are included in Appendix A.2

the factorization

$$(\xi_j^\ell, \tilde{\xi}_j^\ell)(\mathbf{z}_\ell(t), \mathbf{z}_j(t)) = \xi_j^\ell\left(\mathbf{z}_\ell(t)\right) + \tilde{\xi}_j^\ell\left(\mathbf{z}_j(t)\right)$$

as an assumption designed to reduce the computational complexity of fitting such a model.

2.4.2 Macrotransition Exit Model

In Section 2.3, we noted that the model for the macrotransition exit state conditional on a macrotransition occurring, $\mathbb{P}(C_\delta | M(\tau), Z_\tau, \tau, \mathcal{F}_t^{(Z)})$, derives from the macrotransition model. We show this by noting that

$$\mathbb{P}(C_\delta | M(\tau), Z_\tau, \tau, \mathcal{F}_t^{(Z)}) = \sum_j \mathbb{P}(C_\delta | M_j(\tau), Z_\tau, \tau, \mathcal{F}_t^{(Z)}) \mathbb{P}(M_j(\tau) | M(\tau), Z_\tau, \tau, \mathcal{F}_t^{(Z)}). \tag{2.8}$$

Probabilities constituting the second term in (2.8) are proportional to $\lambda_j(t)$, thus they derive from our family of competing risks macrotransition models. For $j \neq 5$ (not a shot attempt), the first term in (2.8) is actually degenerate. For each pass (corresponding to $j \leq 4$), the location and position relative to the defense of the pass recipient are given by Z_τ, thus yielding only one possible macrotransition exit state C_δ for each pass option. Note that while players' positions change in the time window (τ, δ), while a pass is airborn, we do not expect the pass recipient's location in the coarsened space $\mathcal{C}$ to change (though our model could be augmented to incorporate this). Similarly, if $j = 6$ and a turnover occurs at τ, then C_δ is the turnover state in $\mathcal{C}_{\text{end}}$ with probability 1.

However, if $j = 5$ and a shot is attempted in $(\tau, \tau + \epsilon]$, then C_δ has two possible values depending on the shooter's location in Z_τ: a made or missed 2 point shot, or made or missed 3 point shot. This motivates a shot probability model, predicting the probability of success given a shot attempt at time t and the associated full resolution data. The parametric form of our shot probability model is exactly the same as our macrotransition model, though we use a logit link function as we are modeling a probability instead of a hazard. Specifically, for player ℓ possessing attempting a shot at time t, let $p^\ell(t)$ represent the probability of the shot attempt being success-

ful (resulting in a basket). We assume

$$\text{logit}(p^{\ell}(t)) = [\mathbf{W}^{\ell}(t)]'\boldsymbol{\beta}^{\ell} + \xi^{\ell}(\mathbf{z}_{\ell}(t)) \tag{2.9}$$

with components in (2.9) having the same interpretation as their j-indexed counterparts in the competing risks model (2.7); that is, $\mathbf{W}^{\ell}$ is a vector of time-varying covariates (we use distance to the nearest defender—transformed as $\log(1+d)$—an indicator for whether the player has dribbled, and a constant to capture baseline shooting efficiency) with $\boldsymbol{\beta}^{\ell}$ a corresponding vector of coefficients, and ξ^{ℓ} is a smooth spatial effect, assumed to be a realization of a Gaussian process.

2.4.3 Transition Probability Matrix for Coarsened Process

The last component of the EPV calculation supplied by the macrotransition model is the transition probability matrix for the embedded Markov chain corresponding to the coarsened process $C^{(0)}, C^{(2)}, \ldots, C^{(K)}$. This transition probability matrix is used to compute terms $\mathbb{E}[h(C_T)|C_\delta]$ that appear in EPV equations (2.4)–(2.5). We shall denote the transition probability matrix as $\mathbf{P}$, where $P_{qr} = \mathbb{P}(C^{(i+1)} = c_r | C^{(i)} = c_q)$ for any $c_q, c_r \in \mathcal{C}$.

Without any other probabilistic structure assumed for $C^{(i)}$ other than Markov, for all i, j, the maximum likelihood estimator of P_{qr} is the observed transition frequency $\frac{\#\{c_q \to c_r\}}{\#\{\text{visits to } c_q\}}$. Of course, this estimator has undesirable performance if the number of visits to any particular state c_q is small, as the estimated transition probabilities from that state may be degenerate. One common approach is to model transition probability matrices hierarchically, possibly in a Bayesian fashion [Lee et al. (1968); Meshkani & Billard (1992)].

Under our multiresolution model for basketball possessions, however, transition probabilities between many coarsened states $C^{(i)}$ can be computed as summaries of the macrotransition model. To show this, for any arbitrary $t > 0$ let $M_j^r(t)$ be the indicator

$$M_j^r(t) = \mathbf{1}[\mathbb{P}(M_j(t) \text{ and } C_{t+\epsilon} = c_r | \mathcal{F}_t^{(Z)}) > 0].$$

Thus $M_j^r(t) = 1$ if it is possible for a macrotransition of type j into state c_q to occur in $(t + t + \epsilon]$. Now, for any c_q such that $c_q \to c_r$ is a macrotransition, we can write

$$
\begin{aligned}
P_{qr} &= \mathbb{P}(C_{t+\epsilon} = c_r | C_t = c_q) \\
&= \mathbb{P}(M_j(t) | C_t = c_q, M_j^r(t) = 1) \\
&= \epsilon \mathbb{E}[\lambda_j(t) | C_t = c_q, M_j^r(t) = 1], \qquad (2.10)
\end{aligned}
$$

where the last equality follows simply from iterated expectation, noting that C_t and $M_j^r(t)$ are both $\mathcal{F}_t^{(Z)}$-measureable. Since we assume C_t is semi-Markov, (2.10) holds for any t.

The integrating measure in (2.10), which conditions on C_t and $M_j^r(t)$, is not immediately available from the multiresolution models without an onerous set of assumptions, so we substitute the empirical distribution of possession paths that occupy c_q at some time point. This yields a simple (unnormalized) estimator $\tilde{P}_{qr} = \sum_{t \in \mathcal{T}^q} \epsilon \lambda_j(t)$ for each r such that $c_q \to c_r$ is a macrotransition for some j, where $\mathcal{T}^q$ is the set of (discretized at resolution ϵ) times for which $C_t = c_q$. Thus, we estimate the transition probability by accumulating the appropriate transition hazard $\lambda_j(t)$. This method leverages the parametric structure of our macrotransition model, and by propagating the shrinkage and temporal smoothing in the macrotransition model to the estimates of P_{qr}, we achieve greater precision than with the naïve MLE.

Transition probabilities corresponding to macrotransition exits are often degenerate (either 0 or 1); this is because for passes and turnovers, the exit state is encoded in the definition of the transition state—for instance, each pass state in $\mathcal{C}_{\text{trans}}$ transitions to a single possession state in $\mathcal{C}_{\text{poss}}$ with probability 1. The exception to this is the exit state of a shot attempt. Recalling that shot attempt states c_q are indexed by the state from which the shot originated, denoted $c_{q'}$, then the next state c_r is either a made or missed 2 point shot, or a made or missed 3 point shot, depending on whether the location corresponding to $c_{q'}$ is behind the three point line. Given the potential point value of the shot, we determine its success probability following a similar line of reasoning: $\tilde{P}_{qr} = \sum_{t \in \mathcal{T}^{q'}} \epsilon \lambda_5(t) p(t)$ when r represents a successful shot, and $\tilde{P}_{qr} = \sum_{t \in \mathcal{T}^{q'}} \epsilon \lambda_5(t)(1 - p(t))$ when r represents a missed shot.

For all other transitions, where $c_q \to c_r$ is not a macrotransition entry or exit, we simply use observed transitions $\tilde{P}_{qr} = \sum_{t \in \mathcal{T}^q} \mathbf{1}[C_{t+\epsilon} = c_r]$. Then the unnormalized transition rates yield

estimated transition probabilities for all q, r:

$$\hat{P}_{qr} = \frac{\tilde{P}_{qr}}{\sum_{r'} \tilde{P}_{qr'}}. \tag{2.11}$$

For transitions from a state not associated with a particular player, we group observed transitions separately by team. The only example of this is the rebound state in $\mathcal{C}_{\text{trans}}$, which with some probability transitions to a defensive rebound (thus ending the possession) and in the case of an offensive rebound transitions into one of the possession states.

2.5 Microtransition model

While the macrotransition model in Section 2.4 models ball movements, the microtransition model describes player movement with the ballcarrier held constant. In the periods between transfers of ball possession (including shots), all players on the court move in order to influence the character of the next ball movement (macrotransition). For instance, the ballcarrier might drive toward the basket to attempt a shot, or move laterally to gain separation from a defender, while his teammates move to position themselves for passes or rebounds, or to set screens and picks. The defense moves correspondingly, attempting to deter easy shot attempts or passes to certain players while simultaneously anticipating a possible turnover.

As defined in Section 2.3, the microtransition model supplies $\mathbb{P}(Z_{t+\epsilon}|M(t)^c, \mathcal{F}_t^{(Z)})$, giving the small-scale evolution of the current possession conditional on the ballcarrier staying the same in the next ϵ time. Separate models are assumed for offensive and defensive players, as we shall describe.

2.5.1 Offensive Movement

Predicting the motion of offensive players over a relatively short time window is driven by the players' dynamics (velocity, acceleration, etc.). Let the location of an offensive player (the ballcarrier, for instance) at time t be $\mathbf{z}(t) = (x(t), y(t))$. Assuming the player's position is differentiable, a Taylor series expansion shows $x(t+\epsilon) = x(t) + \dot{x}(t)\epsilon + e_x(t)$ where the innovations $e_x(t)$ depend on higher derivatives of position (acceleration, jerk, etc.) and possibly involve white

noise, as there is small measurement error associated with the position tracking in our data. The velocity $\dot{x}(t)$ is unobserved, yet it is natural to replace this with $(x(t) - x(t-\epsilon))/\epsilon$, acknowledging that doing so adds additional structure to the residual $e_x(t)$ term. We are now left with

$$x(t+\epsilon) = x(t) + a_x[x(t) - x(t-\epsilon)] + e_x(t), \tag{2.12}$$

where $a_x = 1$ in theory, but this is relaxed because $a_x < 1$ in practice provides some predictive stability, as the sequence of differences $x(t) - x(t+\epsilon)$ becomes a stationary AR(1) process if $e_x(t)$ is white noise. Note that (2.12) defines $x(t)$ as an ARI(1,1) process (or equivalently, as an ARIMA(1,1,0) process).

We also assume spatial structure for the innovations, $e_x(t) \sim \mathcal{N}(\mu_x(\mathbf{z}(t)), \sigma_x^2)$, where μ_x maps the player's two-dimensional location on the court to an additive effect in (2.12), which has the interpretation of an acceleration effect. Players' future motion is informed not only by their current dynamics, but also their position on the court. Players within the perimeter, for instance, may be more likely to accelerate towards the basket as they get closer, eventually decelerating to attempt a shot. Also, players will accelerate away from the edges of the court as they approach these, in order to stay in bounds (see Figure 2.5 for an illustration). These behaviors motivate the inclusion of μ_x in the model (2.12). For the $y(t)$, we construct (2.12) analogously in terms of a_y and $e_y(t) \sim \mathcal{N}(\mu_y(\mathbf{z}(t)), \sigma_y^2)$.

2.5.2 Defensive Movement

The defensive components of $P(Z_{t+\epsilon}|M(t)^c, \mathcal{F}_t^{(Z)})$, corresponding to the positions of the five defenders, are easier to model conditional on the evolution of the offense's positions. Following Franks et al. (2014), we assume each defender's position is centered on a linear combination of the basket's location, the ball's location, and the location of the offensive player he is guarding. Franks et al. (2014) use a hidden Markov model (HMM), based on this assumption, to learn which offensive players each defender is guarding, as well as the coefficients of this linear combination. The location coefficients they estimate are 0.62 for the offensive player being guarded, 0.11 for the ball ball, and 0.27 for the basket; that is, conditional on defender i guarding offender

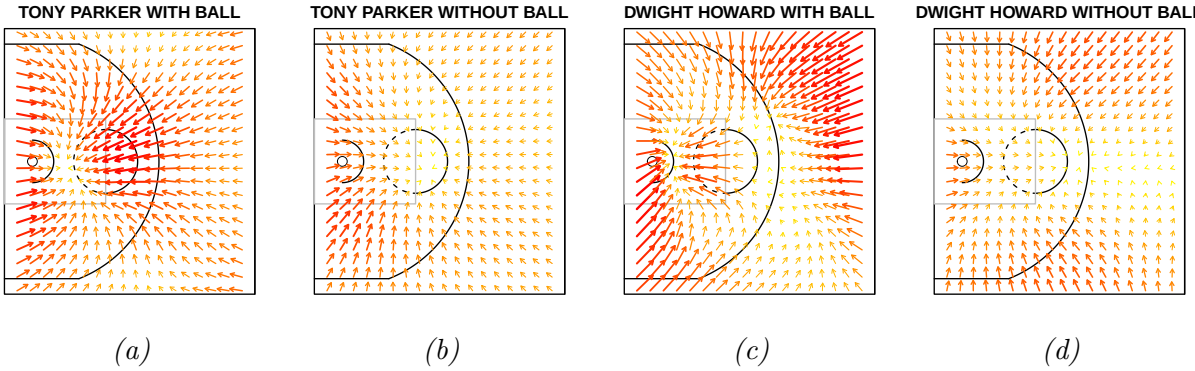


Figure 2.5: Acceleration fields $(\mu_x(\mathbf{z}(t)), \mu_y(\mathbf{z}(t)))$ for Tony Paker (a)–(b) and Dwight Howard (c)–(d) with and without ball possession. The arrows point in the direction of the acceleration at each point on the court's surface, and the size and color of the arrows are proportional to the magnitude of the acceleration. Comparing (a) and (c) for instance, we see that when both players possess the ball, Parker more frequently attacks the basket from outside the perimeter. Howard does not accelerate to the basket from beyond the perimeter, and only tends to attack the basket inside the paint.

j his location $\mathbf{z}_i(t)$ should be normally distributed with mean $\mathbf{m}_j^{\text{opt}}(t) = 0.62\mathbf{z}_j(t) + 0.11\mathbf{z}_{\text{bask}} + 0.27\mathbf{z}_{\text{ball}}(t)$.

Of course, the dynamics (velocity, etc.) of defensive players' are still hugely informative for predicting their locations within a small time window. Thus our microtransition model for defenders balances these dynamics with the mean path induced by the player each is guarding:

$$x(t+\epsilon)|m_{j,x}^{\text{opt}}(t) = x(t) + a_x[x(t) - x(t-\epsilon)] + b_x[m_{j,x}^{\text{opt}}(t+\epsilon) - m_{j,x}^{\text{opt}}(t)] + \mathcal{N}(0, \tau_x^2). \tag{2.13}$$

Rather than implement the HMM procedure used in Franks et al. (2014), we simply assume each defender is guarding at time t whichever offensive player j yields the smallest residual $||\mathbf{z}(t) - \mathbf{m}_j^{\text{opt}}(t)||$. Note that more than one defender may be guarding the same offender (as in a "double team"). Thus, conditional on the locations of the offense at time $t+\epsilon$, (2.13) provides a distribution over the locations of the defense at time $t+\epsilon$. As in estimating the microtransition components for the offense, we fit (2.13) separately for all defenders, for both the x and y component of the position.

2.6 Inference

To this point, we have expressed EPV using multiresolution conditioning, introduced parameterizations for our macro- and microtransition models, and derived all necessary probability components for evaluating EPV (such as the shot probability model and the transition probability matrix for C_t) from parameters of these models. This section outlines the pipeline for estimating these model parameters and computing the derived EPV estimates that yield actual results, such as those highlighted in Section 2.2.1. Our estimates are based on all games from the 2013-14 season up until February 7, 2014, though only 90% of these games were used in model fitting (the remaining 10% were used to assess out of sample predictive performance or our model components).

2.6.1 Likelihood Inference

We estimate multiresolution transition models by Bayesian inference. Note that the multiresolution transition framework helps us rewrite the data generating process by conditioning on macrotransition events. The likelihood of a possession $\{Z_t, 0 \leq t \leq T\}$ observed at a temporal resolution of ϵ can be written

$$\prod_{t=0}^{T-\epsilon} \mathbb{P}(Z_{t+\epsilon}|\mathcal{F}_t^{(Z)}) = \left(\prod_{t=0}^{T-\epsilon} \mathbb{P}(Z_{t+\epsilon}|M(t)^c, \mathcal{F}_t^{(Z)})^{\mathbf{1}[M(t)^c]} \prod_{j=1}^{6} \mathbb{P}(Z_{t+\epsilon}|M_j(t), \mathcal{F}_t^{(Z)})^{\mathbf{1}[M_j(t)]} \right) \times \left(\prod_{t=0}^{T-\epsilon} \mathbb{P}(M(t)^c|\mathcal{F}_t^{(Z)})^{\mathbf{1}[M(t)^c]} \prod_{j=1}^{6} \mathbb{P}(M_j(t)|\mathcal{F}_t^{(Z)})^{\mathbf{1}[M_j(t)]} \right), \quad (2.14)$$

where the first term models $Z_{t+\epsilon}$ conditional on a macrotransition (or lack thereof) in the window $(t, t+\epsilon]$, and the second term models such macrotransition events. Following this factorization, we consider two separate models for macrotransitions and microtransitions, decomposiing (2.14) into partial likelihoods [Cox (1975b)] that inform the parameters of the macro- and microtransitions independently. The macrotransition model is estimated using the second term in the likelihood (2.14), whereas the microtransition model is fit using the term $\left(\prod_{t=0}^{T-\epsilon} \mathbb{P}(Z_{t+\epsilon}|M(t)^c, \mathcal{F}_t^{(Z)})^{\mathbf{1}[M(t)^c]}\right)$. Under mild conditions, this inferential procedure leads to consistent and asymptotically well-

behaved estimators [Wong (1986)].

The reader is referred to Appendix A for explicit expressions of the partial likelihood terms for the macro and microtransition model in terms of the corresponding model parameters. Prior distributions are also given for these parameters, such that the inference is partially Bayesian [Cox (1975a)]. All spatial effects in our models (ξ in the macrotransition model and μ in the microtransition model) are represented using functional bases whose loadings are given a unique prior structure that shares information across space and across players. This not only provides more precise inference with better out-of-sample predictive performance (see Table 2.1), but it also offers substantial computational advantages. Appendix A also outlines the computational requirements for parameter inference using (2.14).

2.6.2 EPV Algorithm

Given all parameter values for our multiresolution transition models, denoted $\boldsymbol{\Theta}$, then EPV at any time during a possession ν_t can be evaluated deterministically—though this may require Monte Carlo integration depending on the current state of the possession. Algorithm 1 illustrates this process explicitly. EPVdraw obtains a draw from the distribution of $X = h(C_T)$ given $\mathcal{F}_t^{(Z)}$, and repeated draws yield an arbitrarily accurate estimate of ν_t.

Inside the function EPVdraw, we repeatedly iterate draws from the macro- and microtransition models in order to simulate a future possession path up until a macrotransition occurs. From this sample path, we then draw the macrotransition exit state C_δ, and compute its value $\mathbb{E}[h(C_T)|C_\delta]$ using the transition probability matrix $\mathbf{P}$. This procedure mirrors the multiresolution conditioning equations (2.4) given in Section 2.3. Note that it is computationally necessary to work with a compressed version of $\mathbf{P}$ corresponding to only the states accessible from C_t. This dramatically reduces the dimension of $\mathbf{P}$, as most states in $\mathcal{C}$ are possession or transition states belonging to players not on the same team as the ballcarrier, meaning they are not accessible from C_t.

Algorithm 1 Calculating EPV (ν_t).

Require: Player ℓ possess the ball at time t

function MACRO($\mathcal{F}_s^{(Z)}, \boldsymbol{\Theta}$) ▷ Simulates a possible macrotransition in $(s, s+\epsilon]$
  for j in $1, \ldots, 6$ **do**
    Set $M_j(s) = 1$ with probability $\min\{1, \lambda_j^\ell(s)\}$
  end for
  if $\sum_j M_j(s) > 1$ **then**
    Keep only one j such that $M_j(s) = 1$, choosing it proportional to $\lambda_j^\ell(s)$
  end if
  return $\{M_j(s), j = 1, \ldots, 6\}$
end function

function EPVDRAW($\mathcal{F}_t^{(Z)}, \boldsymbol{\Theta}$) ▷ Gets EPV from single simulation of next macro
  Initialize $s \leftarrow t$
  Initialize $M_j(s) \leftarrow$ MACRO($\mathcal{F}_s^{(Z)}, \boldsymbol{\Theta}$)
  while $M_j(s) = 0$ for all j **do**
    Draw $Z_{s+\epsilon} \sim \mathbb{P}(Z_{s+\epsilon} | M(s)^c, \mathcal{F}_s^{(Z)})$
    $\mathcal{F}_{s+\epsilon}^{(Z)} \leftarrow \{\mathcal{F}_t^{(Z)}, Z_{s+\epsilon}\}$
    $s \leftarrow s + \epsilon$
    $M_j(s) \leftarrow$ MACRO($\mathcal{F}_s^{(Z)}, \boldsymbol{\Theta}$)
  end while
  Draw $C_\delta \sim \mathbb{P}(C_\delta | M_j(s), \mathcal{F}_s^{(Z)})$
  $\nu_t \leftarrow \mathbb{E}[h(C_T) | C_\delta]$
  return ν_t
end function

function EPV($N, \mathcal{F}_t^{(Z)}, \boldsymbol{\Theta}$) ▷ Averages over simulations of next macrotransition
  Initialize $\nu_t \leftarrow 0$
  for i in $1, \ldots, N$ **do**
    $\nu_t \leftarrow \nu_t +$ EPVDRAW($\mathcal{F}_t^{(Z)}, \boldsymbol{\Theta}$)
  end for
  return ν_t / N
end function

2.7 Results

Applying Algorithm 1 using our parameter estimates for the multresolution transition model, we can plot EPV (ν_t) throughout the course of any possession in our data. We view EPV curves as the main contribution of our work, and their behavior and potential inferential value has been introduced in Section 2.2.1. Analysts may also find meaningful aggregations of EPV curves that

summarize players' behavior over a possession, game, or season in terms of EPV—we offer two such aggregations in Appendix B.

2.7.1 Possession Inference from Multiresolution Transitions

Understanding the calculation of EPV in terms of multiresolution transitions is also a valuable exercise for a basketball analyst, as these model components reveal precisely how the EPV estimate derives from the spatiotemporal circumstances of the time point considered. Figure 2.6 diagrams four moments during our example possession (introduced originally in Figures 2.1 and 2.2) in terms of multiresolution transition probabilities. These diagrams illustrate equation (2.4) by showing EPV as a weighted average of the value of the next macrotransition. Potential ball movements representing macrotransitions are shown as arrows, with their respective values and probabilities graphically illustrated by color and line thickness (this information is also annotated explicitly). Microtransition distributions are also shown, indicating distributions of players' movement over the next two seconds. Note that the possession diagrammed here was not used in our model fitting.

Analyzing Figure 2.6, we see that our model estimates largely agree with basketball intuition. For example, players are quite likely to take a shot when they are near to and/or moving towards the basket, as shown in panels A and D. Additionally, because LeBron James is a better shooter than Norris Cole, the value of his shot attempt is higher, even though in the snapshot in panel D he is much farther from the basket than Cole is in panel A. While the value of the shot attempt averages over future microtransitions, which may move the player closer to the basket, when macrotransition hazards are high this average is dominated by microtransitions on very short time scales.

We also see Ray Allen, in the right corner 3, as consistently one of the most valuable pass options during this possession, particularly when he is being less closely defended as in panels A and D. In these panels, though, we never see an estimated probability of him receiving a pass above 0.05, most likely because he is being fairly closely defended for someone so far from the ball, and because there are always closer passing options for the ballcarrier. Similarly, while Chris Bosh does not move much during this possession, he is most valuable as a passing option

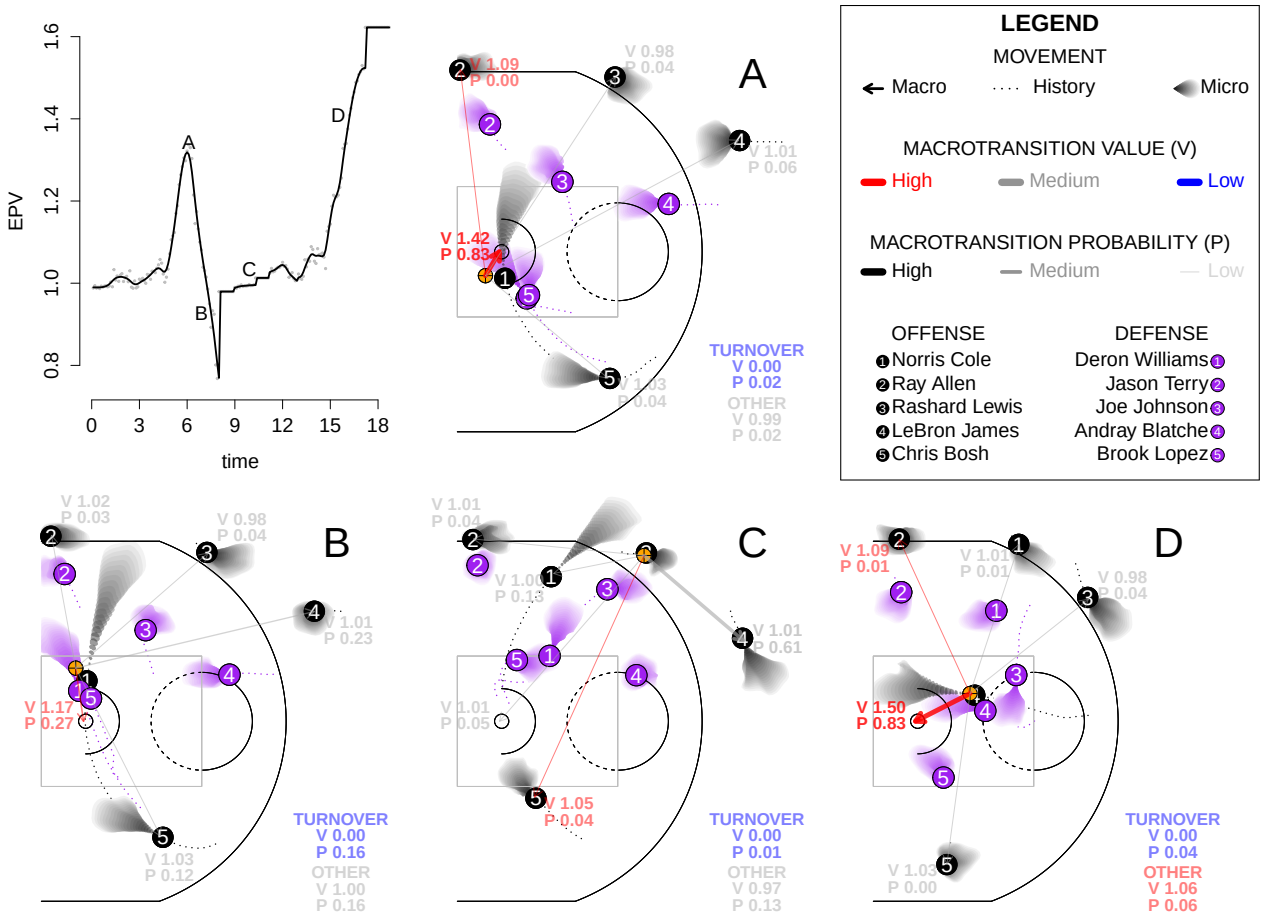


Figure 2.6: Detailed diagram of EPV as a function of multiresolution transition probabilities for four time points (labeled A,B,C,D) of the possession featured in Figures 2.1–2.2. Two seconds of microtransitions are shaded (with forecasted positions for short time horizons darker) while macrotransitions are represented by arrows, using color and line thickness to encode the value and probability of such macrotransitions. The value and probability of the "other" category represents the case that no macrotransition occurs during the next two seconds.

in panel C where he is closest to the basket and without any defenders in his lane. Lastly, while we estimated the probability of Lewis passing to James in panel C at 0.61 (by far Rashard Lewis' most likely passing option), we only estimated the probability of the pass from Cole to Lewis (panel B) at 0.04 (the pass actually happens a fraction of a second after the situation in panel B). More generally, our model anticipated a layup from Cole, instead of his path underneath the basket and eventual pass to Lewis. These issues aside, the estimated probabilities and values of the macrotransitions highlighted in Figure 2.6 match well with basketball intuition.

The analysis presented here could be repeated on any of tens (hundreds) of thousands of pos-

sessions available in a season of optical tracking data. EPV plots as in Figure 2.2 and diagrams as in Figure 2.6 provide powerful insight as to how players' movements and decisions contribute value to their team's offense. With this insight, coaches and analysts can formulate strategies and offensive schemes that make optimal use of their players' ability—or, defensive strategies that best suppress the motifs and situations that generate value for the opposing offense.

2.7.2 Predictive Performance of EPV

Our paper introduces EPV, and as such there are no existing results to benchmark the predictive performance of our estimates. We can, however, compare the proposed implementation for estimating EPV with simpler models, based on lower resolution information, to verify whether our model captures meaningful features of our data. Assessing the predictive performance of an EPV estimator is difficult because the estimand is a curve whose length varies by possession. Moreover, we never observe any portion of this curve; we only know its endpoint. Therefore, rather than comparing estimated EPV curves between our method and alternative methods, we compare estimated transition probabilities. For any EPV estimator method that is stochastically consistent, if the predicted transitions are properly calibrated, then the derived EPV estimates should be as well.

As mentioned in Section 2.6, we use only 90% of our data set for parameter inference, with the remaining 10% used to evaluate the out-of-sample performance of our model. We also evaluated out-of-sample performance of alternative macrotransition models, which use varying amounts of information from the data. Table 2.1 provides the out-of-sample log-likelihood for the macrotransition model applied to the 10% of the data not used in model fitting for various hazard parameterizations. The most basic parameterization assumes constant hazards for each ballcarrier/macrotransition type. We also consider a hazard mode that is unique for each ballcarrier and macrotransition, yet includes only the time-referenced covariates (situational effects) used in our full model (and no spatial effect). Finally, we consider our full model, with unique situational and spatial effects for each ballcarrier/macrotransition, both with and without the hierarchical model we use for information sharing across players. This hierarchical model is discussed in Appendix A.4. Without any shrinkage, our full model performs in some cases worse than a model

with no spatial effects included, but with shrinkage, it consistently performs the best of the configurations compared—this behavior motivates the novel hierarchical structure we discuss in Appendix A, which incorporates both spatial and between-player structure in a computationally efficient manner.

	Macrotransition Model			
Macrotransition	Player	Covariates	Covariates + Spatial	Full
Pass1	-29399.68	-27659.72	-27251.40	-26433.76
Pass2	-24884.99	-23691.81	-23294.98	-22226.61
Pass3	-26326.99	-25199.52	-25335.92	-23909.71
Pass4	-20426.53	-20266.06	-24487.17	-18879.47
Shot Attempt	-48885.21	-46471.49	-40914.66	-40711.55
Made Basket	-6579.31	-6626.55	-5601.75	-5284.31
Turnover	-9311.80	-9075.60	-8990.61	-8390.85

Table 2.1: Out of sample log-likelihood for macrotransition models (and shot probability model) under various model specifications. "Player" assumes constant hazards for each player/event type combination. "Covariates" augments this model with situational covariates, $\mathbf{W}(t)$ *as given in* (2.7)*. "Covariates + Spatial" adds a spatial effect, yielding* (2.7) *in its entirety. Lastly, "Full" implements this model with the hierchical model discussed in Appendix A.*

Our comparison of the predictive performance of the competing risks macrotransition model under several different parameterizations includes a notable case. Assuming constant hazards for each player/macrotransition type is equivalent to using only the discrete, homogeneous Markov chain $C^{(0)}, C^{(1)}, \ldots, C^{(K)}$ to compute EPV, using empirical transition frequencies to estimate the transition probability matrix. The superior predictive performance of our EPV model illustrates the value in modeling the full resolution data instead of simply relying on discrete summaries.

2.8 Discussion

This paper introduces a new quantity, EPV, which represents a paradigm shift in the possibilities for statistical inferences about basketball. Using high resolution, optical tracking data, EPV reveals the value in many of the schemes and motifs that characterize basketball offenses but are omitted in the box score. For instance, as diagrammed in Figures 2.2 and 2.6, we see that EPV may rise as a player attacks the basket (more so for a strong scorer like LeBron James than for a bench player like Norris Cole), passes to a well-positioned teammate, or gains separation from the defense. Aside from simply tracking changes in EPV, analysts can understand why EPV

changes by expressing its value as a weighted average of transition values (as done in Figure 2.6). Doing so reveals that the source of a high (or low) EPV estimate may come from alternate paths of the possession that were never realized, but were probable enough to have influenced the EPV estimate—an open teammate in a good shooting location, for instance. These insights, which can be reproduced for any valid NBA possession in our data set, have the potential to reshape the way we quantify players' actions and decisions.

We make a number of assumptions—mostly to streamline and simplify our modeling and analysis pipeline—that could be relaxed and yield a more precise model. The largest assumption is that the particular coarsened view of a basketball possession that we propose here is marginally semi-Markov. While this serves as a workable first-order approximation, there are cases that clearly violate this assumption, for example, pre-set plays that string together sequences of runs and passes. Future refinements of the model could define a wider set of macrotransitions that encapsulate these motifs, effectively encoding this additional possession structure from the coach's playbook. A number of smaller details could also be addressed. For instance, it seems desirable to model rebound outcomes conditional on high resolution information, such as the identities and motion dynamics of potential rebounders; we do not do this, however, and use a constant probability for each team of a rebound going to either the offense or defense. We also do not distinguish between different types of turnovers (steals, bad passes, ball out of bounds, etc.), though this is due to a technical feature of our data set. Indeed, regardless of the complexity and refinement of an EPV model, we stress that the full resolution data still omits key information, such as the positioning of players' hands and feet, their heights when jumping, and other variables which impact basketball outcomes. As such, analyses based on EPV are best accompanied by actual game film and the insight of a basketball expert.

The computational requirements of estimating EPV curves (and the parameters that generate them) likely limit EPV discussions to academic circles and professional basketball teams with access to the appropriate resources. Our model nevertheless offers a case study whose influence extends beyond basketball. High resolution spatiotemporal data sets are an emerging inferential topic in a number of scientific or business areas, such as climate, security and surveillance, advertising, and gesture recognition. Many of the core methodological approaches in our work, such

as using multiresolution transitions and hierarchical spatial models, provide insight beyond the scope of basketball to other spatiotemporal domains.

3

Gaussian Process Regression with Location Errors

3.1 Introduction

Gaussian process models assume an output variable of interest varies smoothly over an input space (e.g., percipitation totals across geographical coordinates, crop yield across factor levels of an experimental design). Such models appear frequently in areas as diverse as climate science [Mardia & Goodall (1993)], epidemiology [Lawson (1994)], and black-box problems such as computer experiments, and Bayesian optimization [Sacks et al. (1989); Srinivas et al. (2009)].

Noisy spatial input data are common in many applications; for example, geostatistical data

is often imprecisely spatially referenced, "binned" to the nearest latitude/longitude grid point, or referenced to maps with distorted coordinates [Veregin (1999); Barber et al. (2006)]. Previous research on such error sources has focused on demonstrating their existence and quantifying their magnitude [Bonner et al. (2003); Ward et al. (2005)]. Location (geocoding) errors have also been studied in the context of point process data [Zimmerman & Sun (2006); Zimmerman et al. (2010)].

Relatively little work has been done on interpolation or Gaussian process regression problems in the presence of location measurement error. Gaussian process models do not straightforwardly extend to incorporate input measurement error, and simply ignoring noise in the input space can lead to poor performance. Gabrosek & Cressie (2002) (and later Cressie & Kornak (2003)) adjust Kriging equations for the presence of location errors, and Fanshawe & Diggle (2011) further develop research for this regime to include problems where predictive locations are subject to error.

Through theory and simulation study, our paper provides guidelines on situations when location errors are most impactful for data analysis, and suggestions for incorporating this source of error into inference and prediction. We expand the research in Cressie & Kornak (2003) on best linear unbiased prediction (Kriging) to include methods for obtaining interval forecasts and for quantifying the cost of ignoring location errors. We also discuss MCMC methods for obtaining optimal (minimum mean squared error) predictions, which are averaged over the conditional distribution of (latent) location errors given the observed data.

To establish terminology, let $s_1, s_2, \ldots, s_n$ be locations (inputs) for which we observe the value of a smooth process $x(s_i) \in \mathbb{R}$. Locations are assumed to be p-dimensional, so that $s_i \in \mathbb{S} \subset \mathbb{R}^p$ for all i. The process $x : \mathbb{S} \to \mathbb{R}$ is called a *Gaussian process* if, for any $s_1, \ldots s_n \in \mathbb{S}$, $\mathbf{x}_n = (x(s_1) \;\; x(s_2) \;\; \ldots \;\; x(s_n))'$ is jointly Normally distributed. Typically, the form of this joint distribution is specified by a deterministic or parametric mean function (for now, taken without loss of generality to be 0) and a covariance function $c : \mathbb{S}^2 \to \mathbb{R}$, so that

$$\begin{pmatrix} x(s_1) \\ \vdots \\ x(s_n) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} c(s_1,s_1) & \cdots & c(s_1,s_n) \\ \vdots & \ddots & \\ c(s_n,s_1) & & c(s_n,s_n) \end{pmatrix}\right). \tag{3.1}$$

For c to be a valid covariance function, the covariance matrix in (A.3) must be positive semi-definite for all input vectors $\mathbf{s}_n = (s_1 \ s_2 \ \dots \ s_n)'$.

Gaussian process regression is primarily used as a method for interpolating (predicting) values $\mathbf{x}_k^*$ at unobserved points $\mathbf{s}_k^* = (s_1^* \dots s_k^*)'$ in the input space, given all available observations. Such conditional distributions are easily obtained by exploiting the joint normality of the response x at observed and unobserved locations:

$$\mathbf{x}_k^* | \mathbf{x}_n \sim \mathcal{N}\big(\mathbf{C}(\mathbf{s}_k^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{x}_n, \\ \mathbf{C}(\mathbf{s}_k^*, \mathbf{s}_k^*) - \mathbf{C}(\mathbf{s}_k^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{C}(\mathbf{s}_n, \mathbf{s}_k^*)\big). \tag{3.2}$$

For notation in this paper, we will use $\mathbf{s}_n = (s_1 \ s_2 \ \dots \ s_n)'$ to denote a n-vector of locations in the input space $\mathbb{S}$, and $\mathbf{x}_n = (x(s_1) \ x(s_2) \ \dots \ x(s_n))'$ as the associated vector of observations at $\mathbf{s}_n$, and similarly denoting $\mathbf{x}_k^* = (x(s_1^*) \ \dots \ x(s_k^*))'$, or $x^* = x(s^*)$. Furthermore, $\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)$ will be the covariance matrix of $\mathbf{x}_n$, $\mathbf{C}(\mathbf{s}_k^*, \mathbf{s}_n)$ the $k \times n$ covariance matrix between $\mathbf{x}_k^*$ and $\mathbf{x}_n$, and so forth.

This paper focuses on regimes where locations in the input space $\mathbb{S}$ are affected by error. We can describe this in terms of a surrogate process $y(s_i) = x(s_i + u_i)$ where s_i is a known location in $\mathbb{S}$ and $u_i \in \mathbb{S}$ is unobserved location error. Because of location errors, the analyst observes samples from y, $\mathbf{y}_n$, but is interested in predicting x at unobserved (exact) locations $x(s^*)$.

When x is assumed to be a Gaussian process, there is no nontrivial structure for u that results in y being a Gaussian process. Additionally, it is not possible to write y as a convolution of x and a white noise process as differences between the surfaces y and x will generally be correlated across space: $\mathbb{C}[y(s_1) - x(s_1), y(s_2) - x(s_2)] \neq 0$. Gaussian process regression with location errors therefore cannot be thought of as a classical or Berkson errors-in-variables problem [Carroll et al. (2006)].

Properly accounting for location errors is essential for optimal interpolation and uncertainty quantification, as well precise and efficient parameter estimation when parameters of the covariance function are unknown. Interestingly, in some cases, the process y may be more informative for prediction at a new location $x(s^*)$ than the process x is. Thus, appropriate methods can deliver lower MSE interpoloations in a location-error regime than can the usual methods in an

error-free regime.

In Section 3.2, we discuss Kriging using the covariance structure of the error-induced process y. Section 3.3 considers Markov Chain Monte Carlo methods that sample exactly from $\mathbf{s}_k^*|\mathbf{y}_n$ and obtain optimal forecasts. We compare these methods through simulation study in Section 3.4, and explore an application to interpolating northern hemisphere temperature anomolies in Section 3.5.

3.2 Kriging the Location Error Induced Process y

As Cressie & Kornak (2003) show, we can use second moment properties of y to perform Kriging (they term this "Kriging adjusting for location error" or KALE), noting that measurement errors u induce a new covariance function

$$\begin{aligned} k(s_1, s_2) = \mathbb{C}[y(s_1), y(s_2)] &= \mathbb{E}[c(s_1 + u_1, s_2 + u_2)] \text{ for } s_1 \neq s_2 \\ k(s, s) = \mathbb{C}[y(s), y(s)] \quad &= \mathbb{E}[c(s + u, s + u)] \\ k^*(s, s^*) = \mathbb{C}[y(s), x(s^*)] \quad &= \mathbb{E}[c(s + u, s^*)]. \end{aligned} \tag{3.3}$$

The expectation here is taken over the input errors u, which may be (but not need be) assumed i.i.d. from some distribution $g(u)$. It is important to note that if c is a valid covariance function, then so is k, regardless of the error structure $g(u)$.

Proposition 3.2.1 *Assume for all n and $\mathbf{s}_n \in \mathbb{S}$, $(u_1, u_2, \ldots, u_n) \sim g_{\mathbf{s}_n} \in \mathcal{G}$, where $\mathcal{G}$ is any family of probability measures on $\mathbb{S}$. Then k is a valid covariance function if and only if c is.*

Note that regardless of the form of c, k will always exhibit the "nugget" effect, or discontinuities in the covariance function $\lim_{s_2 \to s_1} k(s_2, s_1) \neq k(s_1, s_1)$ [Matheron (1962)]. In fact, several authors cite location/positional error as a justification for including a nugget term in an arbitrary covariance function c [Cressie & Cassie (1993); Stein (1999)], alongside independent measurement error in observing the resonse, $x(s) + \epsilon$. Location errors, however, cause k to differ from c throughout the spatial domain $\mathbb{S}^2$ (this is pictured in Figure 3.1), meaning that while they induce a nugget, a nugget term alone cannot capture the effect of location errors.

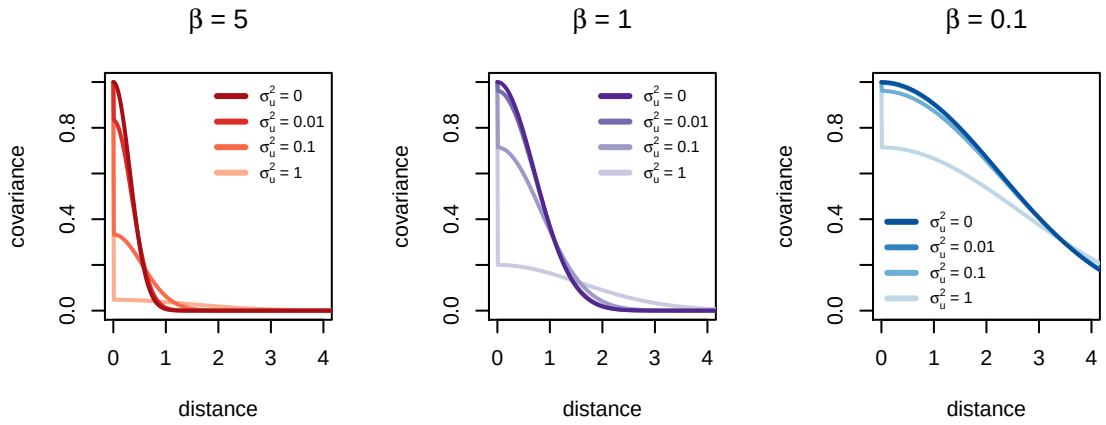


Figure 3.1: Comparison of c and k for $\mathbb{S} = \mathbb{R}^2$ and $c(s_1, s_2) = \exp(-\beta||s_1 - s_2||^2)$, with $u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2 \mathbf{I}_2)$. Location errors $\sigma_u^2 > 0$ cause c and k to differ as a function of distance, and induce a nugget discontinuity at 0.

Using k, we get the Kriging estimator (adjusting for location error) for $x(s^*)$, an unobserved location of x:

$$\hat{x}_{\text{KALE}}(s^*) = \mathbf{K}^*(s^*, \mathbf{s}_n)\mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{y}_n. \tag{3.4}$$

$\hat{x}_{\text{KALE}}(x^*)$ is the best linear unbiased predictor (in terms of MSE) and has all the usual Kriging properties. When there are no location errors, the Kriging estimator is equivalent to the conditional expectation of $x(s^*)$ given $\mathbf{x}_n$ (3.2).

In general, the covariance functions k and k^* can be evaluated using Monte Carlo integration, sampling independently from $g(u)$. For several common combinations of covariance function and location error models, however, it is possible to arrive at expressions for (3.3) in closed form. In particular, if $c(s_1, s_2)$ has the form $\tau^2 \exp(-\beta g(s_1, s_2))$, then we can define a random variable $Z = g(s_1 + u_1, s_2 + u_2)$ and find its moment generating function $M_Z(t)$. If we can evaluate $M_Z(t)$ at $t = -\beta$, then this yields $k(s_1, s_2)$. For instance, for the squared exponential covariance function $g(s_1, s_2) = ||s_1 - s_2||^2$ and Normal location errors $u \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$, Z has a scaled noncentral χ_p^2 distribution and

$$\begin{aligned} k(s_1, s_2) &= \frac{\tau^2}{(1 + 4\beta\sigma_u^2)^{p/2}} \exp\left(-\frac{\beta}{1 + 4\beta\sigma_u^2}||s_1 - s_2||^2\right) \text{ for } s_1 \neq s_2 \\ k(s, s) &= \tau^2 \end{aligned} \tag{3.5}$$

with a similar expression for $k^*(s, s^*)$. Thus the covariance function for y is also squared exponential (it is not generally true that covariance functions c and k will share the same functional form). Note, however, that not all parameters are identifiable—we must know at least one of $(\tau^2, \beta, \sigma_u^2)$ in order to estimate the other two parameters.

Interestingly, it is possible for the KALE to yield lower MSE predictions than those given from an error-free regime, where $\mathbf{u}_n \equiv 0$ and $x = y$. In other words, $\mathbf{y}_n$ can be more informative than $\mathbf{x}_n$ for predicting $x(s^*)$. Heuristically, this happens when $\mathbf{y}_n$ is more strongly correlated with $x(s^*)$ than is $\mathbf{x}_n$. Below we characterize the conditions for observing this phenomenon in a simple model with one observed data point (Figure 3.1 provides an illustration); it is difficult to generalize to larger observed location samples and covariance/error structures.

Proposition 3.2.2 *Assume $n = 1$, $||s - s^*||^2 = \Delta^2$, $c(s, s^*) = \tau^2 \exp(-\beta\Delta^2)$ for all $s, s^* \in \mathbb{S}$, and $u \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Without location error ($\sigma_u^2 = 0$), the MSE in predicting $x(s^*)$ from $x(s)$ is $c_0 = \tau^2(1 - \exp(-2\beta\Delta^2))$. We can find σ_u^2 such that $\mathbb{E}[(\hat{x}_{\text{KALE}}(s^*) - x(s^*))^2] < c_0$ if and only if $\beta\Delta^2 > p/2$.*

3.2.1 Interval predictions

For many applications of Gaussian process regression, particularly in geostastics and environmental modeling, both point and interval predictions are of interest. However, Kriging, being strictly a moment-based procedure, does not provide uncertainty quantification for predictions other than variance. In a location-error Gaussian process regime, KALE predictions will always be non-Gaussian, thus variance alone is not sufficient to provide distributional or interval predictions.

However, it is relatively straightforward to derive confidence intervals for predictions at unobserved locations $x(s^*)$ given measurements $\mathbf{y}_n$ at locations $\mathbf{s}_n$. The following proposition provides the exact distribution function (CDF) for prediction errors $x(s^*) - \hat{x}_{\text{KALE}}(s^*)$, which can be inverted to obtain a confidence interval for $x(s^*)$ based on $\hat{x}_{\text{KALE}}(s^*)$.

Proposition 3.2.3 *Let*

$$V(\mathbf{u}_n) = \sigma^2 + \gamma'\mathbf{C}(\mathbf{s}_n + \mathbf{u}_n, \mathbf{s}_n + \mathbf{u}_n)\gamma - 2\gamma'\mathbf{C}(\mathbf{s}_n + \mathbf{u}_n, s^*)$$
$$\textit{where } \gamma = \mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{K}^*(\mathbf{s}_n, s^*).$$

Then

$$\mathbb{P}(x(s^*) - \hat{x}_{\text{KALE}}(s^*) < z) = \mathbb{E}\left[\Phi\left(\frac{z}{\sqrt{V(\mathbf{u}_n)}}\right)\right], \tag{3.6}$$

where Φ *is the standard normal distribution function.*

The proof of Proposition 3.2.3 is provided in the appendix. It may be necessary to evaluate (3.6) using Monte Carlo; if so, it is practical to use the same draws of $\mathbf{u}_n$ when evaluating different quantiles z, as this guarantees a Monte Carlo estimate of the distribution function be non-decreasing.

While intervals based on (3.6) provide exact coverage (modulo Monte Carlo error), such coverage is achieved by averaging over all data, both observed ($\mathbf{y}_n$) and unobserved $x(s^*)$ as well as the location errors $\mathbf{u}_n$. This is contrast to usual interval estimates from Gaussian process regression without location error, which are conditional probability statements and have the correct coverage for any observed data $\mathbf{x}_n$. The reason this is an important distinction is because when the usual Gaussian process conditional probability intervals yield the proper coverage rate across multiple prediction intervals from the same data set, whereas the confidence intervals corresponding to KALE may not.

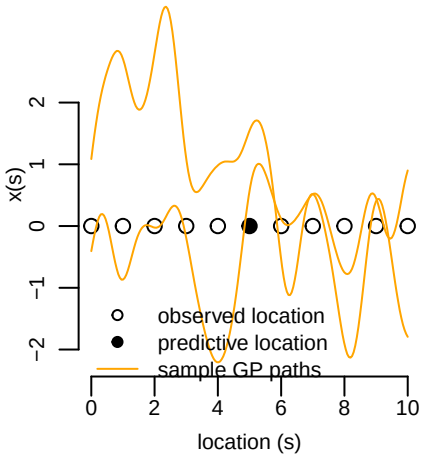
3.2.2 Advantages over Kriging while Ignoring Location Errors

Failing to adjust for location errors when Kriging (Cressie & Kornak (2003) call this "Kriging ignoring location errors" or KILE) can lead to poor performance. If the data analyst ignores location errors, he/she will use

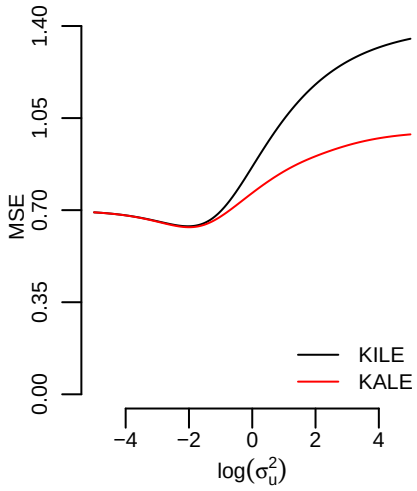
$$\hat{x}_{\text{KILE}}(s^*) = \mathbf{C}(s^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{y}_n. \tag{3.7}$$

Since $\hat{x}_{\text{KALE}}(s^*)$ is the best linear unbiased estimator for $x(s^*)$ and $\hat{x}_{\text{KILE}}$ is also an unbiased linear estimator, KALE dominates KILE and always yields a reduced MSE. Figure 3.2 illustrates the

disparity in MSE for a simple model; intuitively, the relative cost of ignoring location errors increases as the magnitude of the location errors increases. We also see, following Proposition 3.2.2, that for small values of σ_u^2, the MSE associated with both KALE and KILE is decreasing as a function of σ_u^2.



(a) Locations at which we observe $y(s)$, as well as the location at which we wish to predict $x(s)$. Sample paths of Gaussian processes with this covariance function are also shown.



(b) For both KALE and KILE, MSE actually decreases as the magnitude of location errors increases when this magnitude is quite small. Above a certain point, greater location error yields higher MSE and greater disparity between KALE and KILE.

Figure 3.2: Here we assume $x(s)$ is a Gaussian process with mean 0 and covariance function $c(s_1, s_2) = \exp(-(s_1 - s_2)^2)$, with $u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$. We compare MSE for predicting $x(5)$ using KALE and KILE.

Besides yielding suboptimal predictions relative to KALE, ignoring location errors also leads to an estimator for $x(s^*)$ that is not self-efficient [Meng (1994)]; that is, the estimator can be improved (in terms of MSE) by using only a portion of the observed data. The following theorem states that the KILE MSE is unbounded as a function of any single spatial location s_i for $i = 1, \ldots, n$, which is a stronger result than lack of self-consistency.

Theorem 3.2.4 *Assume c is continuous in $\mathbb{S}^2$ and location errors satisfy $\mathbb{P}(u_1 \neq u_2) < 1$ for all $s_1, s_2 \in \mathbb{S}$. Let $\hat{x}_{\text{KILE}}(s^*)$ be the KILE estimator for $x(s^*)$ given $\mathbf{y}_n$. Then for any $M > 0$ and $s_2, \ldots, s_n \in \mathbb{S}$, there exists s_1 such that $\mathbb{E}[(x(s^*) - \hat{x}_{\text{KILE}}(s^*))^2] > M$.*

The proof of Theorem 3.2.4 is given in the Appendix. Note that the condition that c is continuous excludes a nugget term from the distribution of x. The mechanism behind Theorem 3.2.4 is that when observed locations are very close together, this creates degeneracy in the covariance matrix. Without location errors, the usual Kriging error is unaffected by this due to the relationship between the Kriging estimator and the second moment properties of x. However, with the noise-corrupted process y having a different covariance structure, this is no longer the case and the MSE can become arbitrarily large.

Simulation results suggest that even when c contains a nugget term σ_x^2, KILE is still not self-efficient, and additional observations can increase MSE. Figure 3.3 illustrates the change in MSE as a function of the location of an additional observation of y. Following Theorem 3.2.4, we see the MSE is unbounded when $\sigma_x^2 = 0$. But even when $\sigma_x^2 = 1$, it is possible for an additional observation to (slightly) increase MSE.

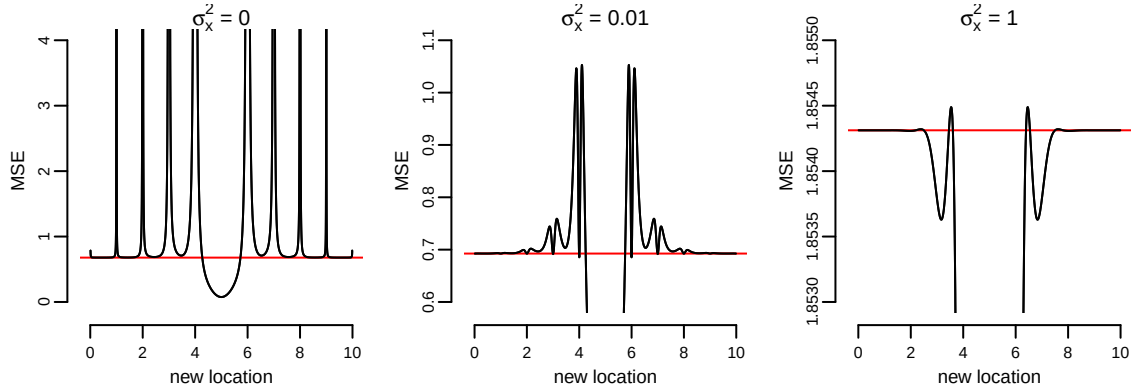


Figure 3.3: Here we assume $x(s)$ is a Gaussian process with mean 0 and covariance function $c(s, s^*) = \exp(-(s - s^*)^2) + \sigma_x^2 \mathbf{1}_{s=s^*}$. Location errors have the form $u_i \sim \mathcal{N}(0, 0.04)$. Despite the magnitude of the location errors being relatively small, observing another measurement of y at some locations can increase (possibly dramatically) the MSE when using KILE to predict $x(5)$ based on $\{y(0), \ldots, y(4), y(6), \ldots, y(10)\}$ (the MSE based on these observations is a red line).

3.2.3 Parameter Estimation for Kriging

In typical applied settings, some or all parameters of the covariance function are unknown and must be estimated by the analyst in order to obtain Kriging equations. For Gaussian process models without a location error component, parameter estimation can be accomplished using likelihood methods. This can be computationally challenging for large data sets, as each likeli-

hood evaluation requires a Cholesky factorization of the covariance matrix (or equivalent operations), which is $\mathcal{O}(n^3)$ except in special cases. An alternative is to choose parameters by maximizing goodness of fit between the empirical variogram and the theoretical (parametric) variogram, though this is less efficient for (parametric) Gaussian models.

Location errors present challenges for both such procedures as the covariance function for the observed provess y (3.3) may not be available in closed form, meaning neither the likelihood function or variogram can be evaluated exactly. While Monte Carlo methods surely offer effective approaches in theory [**??**], they muliply the computational expense of the problem, as each evaluation of the likelihood requires M matrix factorizations, where M is the number of Monte Carlo samples used to approximate the likelihood. Cressie & Kornak (2003) advocate a pseudo-likelihood procedure [Carroll et al. (2006)] that uses a Gaussian likelihood approximation based on the first two moments of y,

$$\tilde{L}(\theta; \mathbf{y}_n) \propto |\mathbf{K}_\theta(\mathbf{s}_n, \mathbf{s}_n)|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{y}_n' \mathbf{K}_\theta(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{y}_n\right), \tag{3.8}$$

where we write $\mathbf{K}_\theta$ to explicitly mark the dependence of the covariance function k on unknown parameters θ. This pseudo-likelihood requires inverting $\mathbf{K}$ only once per pseudo-likelihood evaluation, even when $\mathbf{K}_\theta$ is computed by Monte Carlo.

We can work out inferential properties of the maximum pseudo-likelihood estimator $\hat{\tilde{\theta}} = \text{argmax}_\theta \tilde{L}(\theta; \mathbf{y}_n)$. First, it is straightforward to check that the pseudo-score is an unbiased estimating equation:

$$\mathbb{E}[\tilde{S}(\theta; \mathbf{y}_n)] = \mathbb{E}[\nabla \log(\tilde{L}(\theta; \mathbf{y}_n))] = \mathbf{0}. \tag{3.9}$$

Moreover, one can show the covariance matrix of the pseudo-score is given by

$$\begin{aligned}
\tilde{G}(\theta) =& \mathbb{E}[\tilde{S}(\theta; \mathbf{y}_n)\tilde{S}(\theta; \mathbf{y}_n)'] \\
\tilde{G}(\theta)_{ij} =& \mathbb{E}\left[\frac{1}{2}\text{Tr}\{\Omega_i \mathbf{C}_\theta(\mathbf{u}_n)\Omega_j \mathbf{C}_\theta(\mathbf{u}_n)\}\right] \\
&+ \frac{1}{4}\Big(\mathbb{E}[\text{Tr}\{\Omega_i \mathbf{C}_\theta(\mathbf{u}_n)\}\text{Tr}\{\Omega_j \mathbf{C}_\theta(\mathbf{u}_n)\}] - \text{Tr}\{\Omega_i \mathbf{K}_\theta\}\text{Tr}\{\Omega_j \mathbf{K}_\theta\}\Big),
\end{aligned} \tag{3.10}$$

using the notational abbreviations $\mathbf{C}_\theta(\mathbf{u}_n) = \mathbf{C}_\theta(\mathbf{s}_n + \mathbf{u}_n, \mathbf{s}_n + \mathbf{u}_n)$, $\mathbf{K}_\theta = \mathbf{K}_\theta(\mathbf{s}_n, \mathbf{s}_n) = \mathbb{E}[\mathbf{C}_\theta(\mathbf{u}_n)]$, and $\Omega_i = \mathbf{K}_\theta^{-1}\left(\frac{\partial}{\partial\theta_i}\mathbf{K}_\theta\right)\mathbf{K}_\theta^{-1}$. Lastly, the expected negative Hessian of the log pseudo-likelihood is

$$\begin{aligned}\tilde{H}(\theta)_{ij} &= \mathbb{E}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(\tilde{L}(\theta;\mathbf{y}_n))\right] \\ &= \frac{1}{2}\mathrm{Tr}\{\Omega_i\mathbf{K}_\theta\Omega_j\mathbf{K}_\theta\}. \end{aligned} \tag{3.11}$$

If there are no location errors ($\mathbf{u}_n \equiv \mathbf{0}$), $\tilde{L}$ is an exact likelihood and the second term in (3.10) vanishes so that $\tilde{G}(\theta) = \tilde{H}(\theta)$, confirming the second Bartlett identity. For non-zero location errors, however, we construct the Godambe information matrix as an analog to the Fisher information matrix [Varin et al. (2011)], $\tilde{I}(\theta) = \tilde{H}(\theta)[\tilde{G}(\theta)]^{-1}\tilde{H}(\theta)$. Evaluating $\tilde{I}(\theta)$ for different location error models illustrates the information loss in estimating covariance function parameters θ relative to the error-free case, where $\tilde{I}(\theta) = \tilde{G}(\theta) = \tilde{H}(\theta)$ is equivalent to the Fisher information matrix.

General theory of unbiased estimation equations [Heyde (1997)] suggests the asymptotic behavior of the pseudo-likelihood procedure satisfies

$$\tilde{I}(\theta)^{1/2}(\hat{\tilde{\theta}} - \theta) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3.12}$$

However, (3.12) does not hold in general even in an error-free regime $\mathbf{u}_n \equiv 0$, as asymptotic results for Gaussian process covariance parameters depend on the spatial sampling scheme used and the specific form of the covariance function [Stein (1999)]. We nevertheless expect (3.12) to hold for suitably well-behaved processes under increasing-domain asymptotics (Guyon (1982) gives an applicable result when locations $\mathbf{s}_n$ are on a lattice).

3.3 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo methods offer an alternative to Kriging for prediction in a regime with noisy inputs. They allow us to compute the optimal prediction

$$\begin{aligned}\hat{x}(s^*) &= \mathbb{E}[x(s^*)|\mathbf{y}_n] \\ &= \int \left(\mathbf{C}(s^*, \mathbf{s}_n + \mathbf{u}_n)[\mathbf{C}(\mathbf{s}_n + \mathbf{u}_n, \mathbf{s}_n + \mathbf{u}_n)]^{-1}\mathbf{y}_n\right) \pi(\mathbf{u}_n|\mathbf{y}_n) d\mathbf{u}_n, \end{aligned} \tag{3.13}$$

which will dominate the KALE estimator (3.4) in terms of MSE for any model and set of observed and predicted locations. MCMC methods are necessary for evaluating (3.13) as the density for the conditional distribution $\pi(\mathbf{u}_n|\mathbf{y}_n)$ will not be available in closed form (no possible "conjugate" form for the distribution of $\mathbf{u}_n$ is known to the authors). When model parameters, such as in the covariance function c or the distribution of u are unknown, the distribution $\pi(\mathbf{u}_n|\mathbf{y}_n)$ implicitly averages over the posterior distributions of such parameters.

MCMC methods also allow us to compute prediction intervals $(z_{\text{low}}, z_{\text{high}})$ such that $\mathbb{P}(z_{\text{low}} < x(s^*) < z_{\text{high}}|\mathbf{y}_n) = 1 - \alpha$, which is a stronger coverage guarantee than achieved with the KALE procedure in Proposition 3.2.3, where coverage is achieved only by averaging over $\mathbf{y}_n$.

3.3.1 Distributional Assumptions

MCMC inference for (3.13) requires the assumption that $\mathbf{x}_n$ is Gaussian. While this is a common assumption in practice and has been assumed throughout this paper, it is not necessary to derive the KALE equations and their MSE (but it is necessary to produce coverage intervals as in Proposition 3.2.3). Thus, Kriging approaches, including KALE, are attractive when we do lack knowledge of the joint distribution of x beyond its first two moments.

In this scenario, however, we can still advocate—from a decision-theoretic perspective—a Gaussian assumption when the goal of the analysis minimum MSE prediction. Let $\pi \in \Pi_{\mathbf{0},\mathbf{C}}$ be a choice of joint distribution for $\mathbf{x}_n$ with the appropriate first two moments $\mathbf{0}$ and $\mathbf{C}$. Let $R_{\pi_1}(\pi_2)$ be the risk of the Bayes rule prediction (under squared error loss) at $x(s^*)$ assuming $\pi = \pi_2$ when

in fact $\pi = \pi_1$; that is,

$$R_{\pi_1}(\pi_2) = \mathbb{E}_{\pi_1}[(\mathbb{E}_{\pi_2}[x(s^*)|\mathbf{x}_n] - x(s^*))^2].$$

Note that risk under squared error loss is equivalent to MSE. Because all admissible predictors are Bayes rule predictors for some π, restricting the class of predictors for $x(s^*)$ to be Bayes rule predictors seems appropriate. We then have the following proposition, based on Morris (1983) (Theorem 5.5):

Proposition 3.3.1 *Let $\pi_0 \in \Pi_{\mathbf{0},\mathbf{C}}$ be Gaussian. Then for all $\pi \in \Pi_{\mathbf{0},\mathbf{C}}$ we have*

$$R_\pi(\pi) \leq R_\pi(\pi_0) = R_{\pi_0}(\pi_0) \leq R_{\pi_0}(\pi).$$

If an analyst has decided to use Kriging for predicting $x(s^*)$, then the risk in making an incorrect distributional assumption is $R_{\pi_0}(\pi_0) - R_\pi(\pi_0) = 0$. However, there is an "opportunity cost" in making any non-Gaussian assumption $R_\pi(\pi_0) - R_\pi(\pi) > 0$ for $\pi \neq \pi_0$, which represents the reduction in MSE under π that could be achieved by using a different estimator.

Obviously, if there is a strong reason to believe a non-Gaussian π is true, then analysis should proceed with this assumption, ideally leveraging an estimator that is optimal under these assumptions (instead of Kriging). However, without strong distributional knowledge, the analyst can assume Gaussianity without risking increased MSE or paying an opportunity cost for using an inefficient method.

3.3.2 Hybrid Monte Carlo

Hybrid Monte Carlo is well-suited for the problem of sampling $\pi(\mathbf{u}_n|\mathbf{y}_n) \propto \pi(\mathbf{y}_n|\mathbf{u}_n)\pi(\mathbf{u}_n)$ in order to evaluate (3.13). This is because while $\pi(\mathbf{y}_n|\mathbf{u}_n)$ is computationally expensive (requiring inversion of the covariance matrix $\mathbf{C}_\theta(\mathbf{u}_n) = \mathbf{C}_\theta(\mathbf{s}_n + \mathbf{u}_n, \mathbf{s}_n + \mathbf{u}_n)$), the gradient $\nabla \log(\pi(\mathbf{u}_n|\mathbf{y}_n))$ is a relatively cheap byproduct of this calculation. Often the conditional distribution $\mathbf{u}_n|\mathbf{y}_n$ is correlated across components, making gradient-based MCMC methods more efficient for generating samples. Other gradient-based MCMC sampling methods, such as the Metropolis-adjusted

Langevin algorithm [Roberts et al. (2001)] and variants, may also be well-suited to this problem.

Bayes rule provides $\pi(\theta, \mathbf{u}_n|\mathbf{y}_n) \propto \pi(\mathbf{y}_n|\theta, \mathbf{u}_n)\pi(\theta, \mathbf{u}_n)$, where θ here represents any unknown parameter(s) of the covariance function c. In most situations it will be reasonable to assume $\mathbf{u}_n$ and θ are independent a priori—this is trivially true in the case that θ is assumed known. Recognizing that $\pi(\mathbf{y}_n|\theta, \mathbf{u}_n)$ is Gaussian, we can write the log posterior and its gradient:

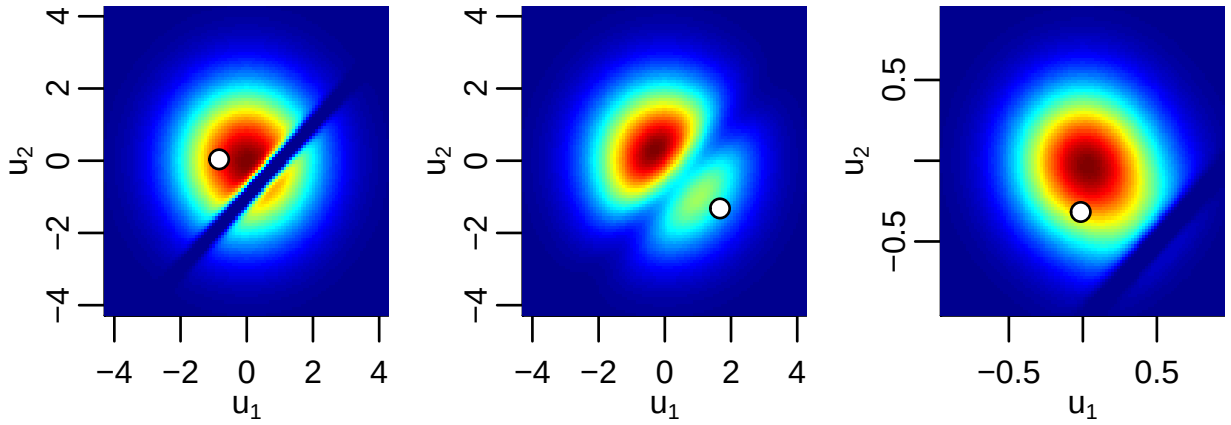
$$\begin{aligned}
\log(\pi(\theta, \mathbf{u}_n|\mathbf{y}_n)) &= -\frac{1}{2}\log(|\mathbf{C}_\theta(\mathbf{u}_n)|) - \frac{1}{2}\mathbf{y}_n'\mathbf{C}_\theta(\mathbf{u}_n)^{-1}\mathbf{y}_n + \text{const.} \\
\frac{\partial}{\partial u_i}\log(\pi(\theta, \mathbf{u}_n|\mathbf{y}_n)) &= \\
&\frac{1}{2}\text{Tr}\left(\mathbf{C}_\theta(\mathbf{u}_n)^{-1}\left[\frac{\partial}{\partial u_i}\mathbf{C}_\theta(\mathbf{u}_n)\right]\left(\mathbf{C}_\theta(\mathbf{u}_n)^{-1}\mathbf{y}_n\mathbf{y}_n' - \mathbf{I}_n\right)\right) + \frac{\partial}{\partial u_i}\log(\pi(\mathbf{u}_n)) \\
\frac{\partial}{\partial \theta_i}\log(\pi(\theta, \mathbf{u}_n|\mathbf{y}_n)) &= \\
&\frac{1}{2}\text{Tr}\left(\mathbf{C}_\theta(\mathbf{u}_n)^{-1}\left[\frac{\partial}{\partial \theta_i}\mathbf{C}_\theta(\mathbf{u}_n)\right]\left(\mathbf{C}_\theta(\mathbf{u}_n)^{-1}\mathbf{y}_n\mathbf{y}_n' - \mathbf{I}_n\right)\right) + \frac{\partial}{\partial \theta_i}\log(\pi(\theta))
\end{aligned}$$

The computational cost of both the likelihood and gradient are dominated by solving $\mathbf{C}_\theta(\mathbf{u}_n)$ (e.g., Cholesky factorization), which is $\mathcal{O}(n^3)$. Every likelihood evaluation computes this term, which can then be re-used in the gradient equations. Thus, the computational cost of computing both the likelihood and gradient remains $\mathcal{O}(n^3)$.

3.3.3 Multimodality

The posterior distribution $\pi(\theta, \mathbf{u}_n|\mathbf{y}_n)$ is often multimodal, more so if the distrubution $\pi(\mathbf{u}_n)$ is diffuse. This is because if there is a local mode at $(\hat{\theta}, \hat{\mathbf{u}}_n)$, there may be a local mode at any $(\theta, \mathbf{u}_n)$ such that $\mathbf{C}_\theta(\mathbf{u}_n) = C_{\hat{\theta}}(\hat{\mathbf{u}}_n)$, as the likelihood is constant for such $(\theta, \mathbf{u}_n)$. In particular, for isotropic covariance models, the likelihood is constant for additive shifts in $\mathbf{u}_n$ or rotations of $\mathbf{s}_n + \mathbf{u}_n$, as these operations preserve pairwise distances. Additionally, multimodality can be induced by many-to-one mapping of the set of true locations $\{s_i + u_i, i = 1, \ldots, n\}$ to the set of observed locations $\{s_i, = 1, \ldots, n\}$. For instance, with $n = 2$ and an isotropic covariance function, for any choice of u_1, u_2 we get the same likelihood with $\tilde{u}_1 = s_2 + u_2 - s_1$ and $\tilde{u}_2 = s_1 + u_1 - s_2$. Moreover, for fixed $\mathbf{u}_n$, for many common covariance functions it is possible for the posterior of θ to be multimodal [Warnes & Ripley (1987)].

HMC (and other gradient MCMC methods) can efficiently sample from multiple modes, however this becomes difficult when the modes are isolated by regions of extremely low likelihood [Neal (2011)]. Isolated modes can occur in the location-error GP regime. For example, assume one-dimensional locations ($p = 1$) and an isotroptic covariance model with known parameters θ and nugget σ_x^2. Marginally, as $||s_1 + u_1 - (s_2 + u_2)|| \to 0$, $y_1 - y_2 \xrightarrow{D} \mathcal{N}(0, 2\sigma_x^2)$; that is, the scaled difference $|y_1 - y_2|/(2\sigma_x)$ must be reasonably small. When this is not the case (e.g. $\sigma_x^2 = 0$), then the log-likelihood asymptotes at $s_1 + u_1 = s_2 + u_2$ almost surely. Thus, the Markov chain can only sample $\mathbf{u}_n$ such that the ordering of $\{s_i + u_i, i = 1, \ldots, n\}$ is preserved. Note that when $p > 1$, while the log-likelihood may still asymptote at $s_1 + u_1 = s_2 + u_2$, this no longer constrains the space of $\mathbf{u}_n$ (except on sets with posterior measure 0).



(a) $\sigma_u^2 = 2$, $\sigma_x^2 = 0.0001$ *(b)* $\sigma_u^2 = 2$, $\sigma_x^2 = 1$ *(c)* $\sigma_u^2 = 0.1$, $\sigma_x^2 = 0.0001$

Figure 3.4: Density of (u_1, u_2) *using covariance function* $c(s_1, s_2) = \exp(-(s_1 - s_2)^2) + \sigma_x^2 \mathbf{1}_{s_1 = s_2}$. *We simulate data* (y_1, y_2) *using* $s_1 = 0$, $s_1 = 1$, *and* $u_i \sim \mathcal{N}(0, \sigma_u^2)$, *and different values of* σ_x^2 *and* σ_u^2.

Figure 3.4 demonstrates the modal behavior for this simple example with $p = 1$ and $n = 2$. When location errors are large in magnitude and the nugget variance σ_x^2 is small, the posterior modes of (u_1, u_2) are separated by a contour of near 0 density (panel A). A higher nugget σ_x^2 increases the density between the modes, making it easier for the same MCMC chain to travel between them (panel B). Decreasing the magnitude of the (Gaussian) location errors, σ_u^2, puts more mass on a single mode, as the unimodal distribution $\pi(\mathbf{u}_n)$ has a greater influence on $\pi(\mathbf{u}_n|\mathbf{y}_n)$ (panel C).

Thus, as with any MCMC application, for the location-error GP problem it is advisible to run

separate chains in parallel, with different, diffuse starting points, and monitor mixing diagnostics [Gelman & Shirley (2011)]. Multiple chains failing to mix is likely a symptom of multiple isolated modes, in which case we should modify the HMC algorithm to include tempering [Salazar & Toral (1997)] or non-local proposals that allow for mode switching [Qin & Liu (2001); Lan et al. (2013)]. Another strategy to overcome multiple isolated modes is importance sampling: as Figure 3.4 shows, increasing the nugget variance σ_x^2 increases the density between modes. If we generate samples according to $\tilde{\pi}(\theta, \mathbf{u}_n|\mathbf{y}_n) \propto \tilde{\pi}(\mathbf{y}_n|\theta, \mathbf{u}_n)\pi(\theta)\pi(\mathbf{u}_n)$ where $\tilde{\pi}(\mathbf{y}_n|\theta, \mathbf{u}_n)$ is the density corresponding to $\mathcal{N}(\mathbf{0}, \mathbf{C}_\theta(\mathbf{u}_n) + \kappa\mathbf{I}_n)$ for some fixed κ, then it is straightforward to compute importance weights $\pi(\theta, \mathbf{u}_n|\mathbf{y}_n)/\tilde{\pi}(\theta, \mathbf{u}_n|\mathbf{y}_n)$. This is because $\mathbf{C}_\theta(\mathbf{u}_n)^{-1}$ is easy to compute from $(\mathbf{C}_\theta(\mathbf{u}_n) + \kappa\mathbf{I}_n)^{-1}$ (and vice versa) using the Woodbury formula. Either standard importance sampling, or Hamiltonian importance sampling [Neal (2005)], could be used to generate parameter estimates, point/interval predictions, and any other posterior estimates of interest.

3.4 Simulation study

We compare Kriging (both KALE and KILE) and HMC methods for point/interval forecasts for Gaussian process regression in a simulation study. For various combinations of parameter values for the covariance function $c(s_1, s_2)$ and location error model $g(u)$ we simulate observations $\mathbf{y}_n$ where $y_i = x(s_i + u_i)$ and make predictions for values of x at unobserved locations: $\mathbf{x}_k^* = (x(s_1^*) \ \ldots \ x(s_k^*))'$.

We simulate data using the squared exponential covariance function $c(s_1, s_2) = \tau^2 \exp(-\beta||s_1 - s_2||^2) + \sigma_x^2 \mathbf{1}_{s_1=s_2}$ and an i.i.d. Gaussian location error model $u_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2\mathbf{I}_p)$. The squared exponential covariance function and Gaussian location error model combine to form a convenient regime, as we can evaluate k in closed form (3.5). Without loss of generality, we can use $\tau^2 = 1$ for all simulations as it is simply a scale parameter. We consider a $p = 2$ dimensional location space, $s_i \in \mathbb{R}^2$. On a 8×8 grid, we randomly select 54 locations at which we observe y, and target the remaining 10 locations for interpolating x. Figure 3.5 illustrates a range of data samples for processes used in our simulations on this space, while Table 3.1 provides a full summary of all the parameter value combinations we consider. Data from each parameter combination is simulated 100 times.

Parameter	Values simulated	Prior support
τ^2	1	$(0, 10)$
β	0.001, 0.01, 0.1, 0.5, 1, 2	$(0.0005, 3)$
σ_x^2	0.0001, 0.01, 0.1, 0.5, 1	$(0, 10)$
σ_u^2	0.0001, 0.01, 0.1, 0.5, 1	$(0, 10)$

Table 3.1: Parameter values used in simulation study. The range $(0.0005, 3)$ *for* β *guarantees that at least one pair of points among our observed data has a correlation in the range* $(0.05, 0.95)$*. This eliminates modes corresponding to white noise processes from the likelihood surface.*

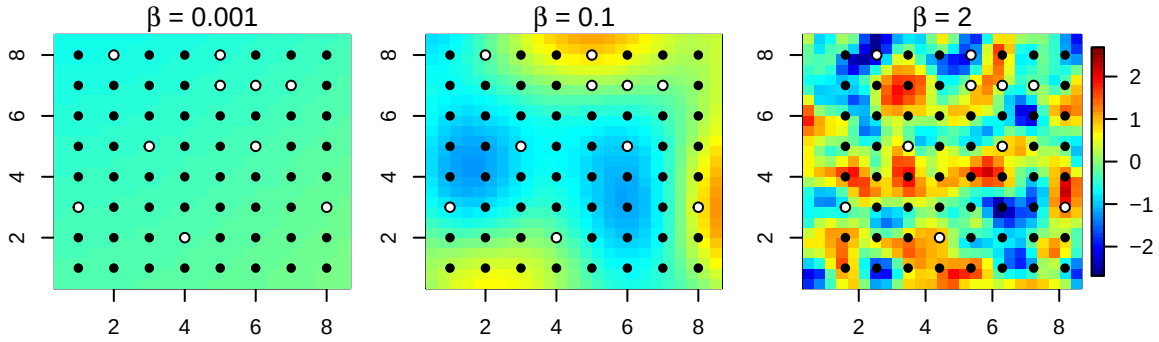


Figure 3.5: Samples of $x(s)$ *for different values of the length-scale parameter* β *with the squared exponential covariance function,* $c(s_1, s_2) = \exp(-\beta||s_1 - s_2||^2) + \sigma_x^2 \mathbf{1}_{s_1 = s_2}$*. Black points are where we have observed* $y(s)$ *and white points are where we wish to predict* $x(s)$*. Observed/predicted locations were randomly sampled from an* 8×8 *grid.*

We evaluate the three prediction methods—KALE, KILE, and HMC—using both adjusted root mean squared error (RMSE) and the coverage probability of a 95% interval. "Adjusted" RMSE is based on the MSE with σ_x^2 subtracted out, as this term appears in the MSE for any prediction method. For every parameter combination of interest used, these statistics are calculated first by averaging over each of the $k = 10$ prediction targets in each simulated draw of new data, and then over the $J = 100$ independent data draws.

Both evaluation statistics can be evaluated more precisely during simulation by utilizing a simple Rao-Blackwellization. For iteration j, instead of drawing $\mathbf{x}_k^*$ in addition to $\mathbf{y}_n$ and calculating $\text{rmse}_j = ||\mathbf{x}_k^* - \hat{\mathbf{x}}_k^*||/k$, we simply condition on the simulated location errors $\mathbf{u}_n$ to get $\text{rmse}_j = \mathbb{E}[||\mathbf{x}_k^* - \hat{\mathbf{x}}_k^*||/k \mid \mathbf{y}_n, \mathbf{u}_n]$. Similarly, to calculate coverage of an interval $(L_{s^*}(\mathbf{y}_n), U_{s^*}(\mathbf{y}_n))$

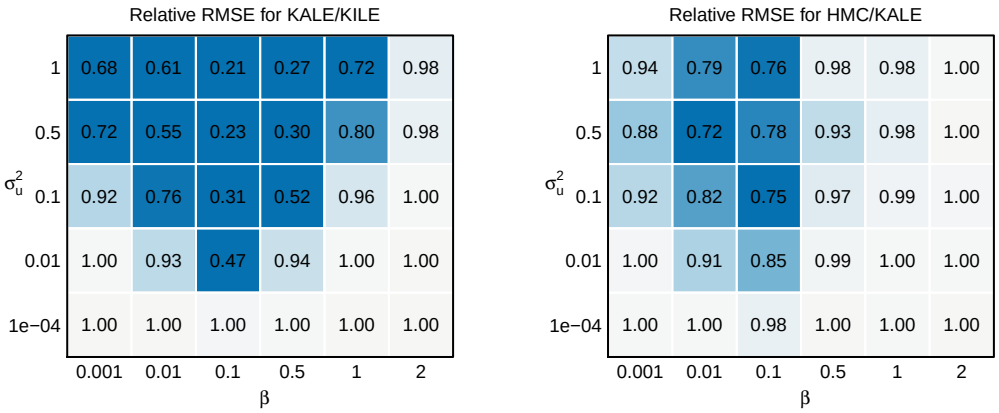
for $x(s^*)$, for iteration j we use

$$\text{cov}_j = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[\mathbf{1}[x(s_i^*) \in (L_{s_i^*}(\mathbf{y}_n), U_{s_i^*}(\mathbf{y}_n))] \mid \mathbf{y}_n, \mathbf{u}_n].$$

HMC is done using the software `RStan` [Stan Development Team (2014)], which implements the "no-U-turn" HMC sampler [Homan & Gelman (2014)]. 10000 samples were drawn during each simulation iteration, which (for most parameter values) takes a few minutes on a single 2.50Ghz processor.

3.4.1 Known covariance parameters

We first simulate point and interval prediction for KALE, KILE, and HMC using the same parameter values that generated the data. By doing so, we leave aside the issue of parameter inference and simply compare the extent to which different methods leverage the information in the location-error corrupted data $\mathbf{y}_n$ to infer $x(s^*)$. Figure 3.6 compares RMSE for the three methods when there is a very small nugget, $\sigma_x^2 = 0.0001$.



(a) RMSE ratio of KALE to KILE. *(b) RMSE ratio of HMC to KALE.*

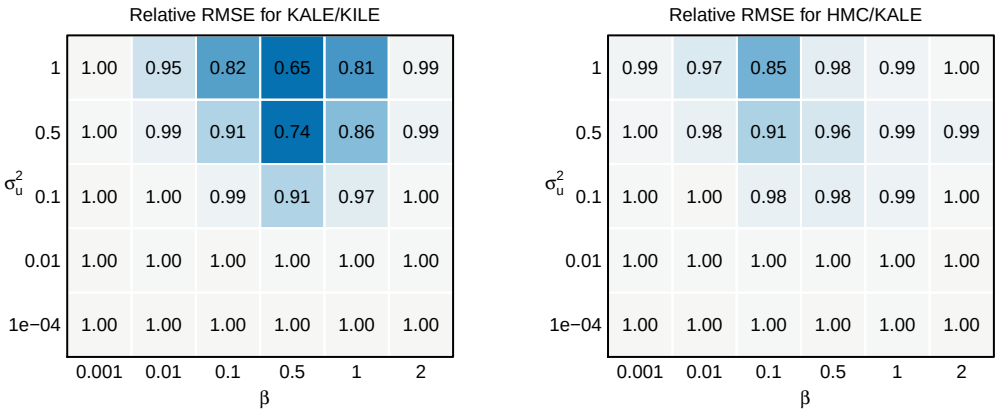
Figure 3.6: Relative RMSE of KALE and KILE (A) and HMC and KALE (B) for each combination of parameters (β, σ_u^2) indicated, and $\sigma_x^2 = 0.0001$. Blue shading represents a relative decrease in RMSE while red shading represents a relative increase in RMSE.

We can see that there is little difference among the three methods when σ_u^2 is sufficiently small (0.0001), or when β is sufficiently large (2). This makes sense, as in the former case, with small

location errors the potential improvement over KILE (which is exact for $\sigma_u^2 = 0$) is negligible, and in the latter case, observations are too weakly correlated for nearby points to be informative. Larger values of σ_u^2 give KALE a significant reduction in RMSE versus KILE, with the reduction as large as 79% for the case of large magnitude location errors ($\sigma_u^2 = 1$) and a moderately smooth signal ($\beta = 0.1$).

The idea of a moderately smooth signal requires further elaboration: for a given σ_u^2, when x is very smooth (β very small), the process is roughly constant within small neighborhoods, meaning $y(s) \approx x(s)$ and location errors are less of a concern for accurate inference and prediction. On the other hand, when β is very large and the process is highly variable in small regions of the input space, location errors are less of a concern because there is very little information in the data to begin with. Location errors are most influential when the process x has more moderate variation across neighborhoods corresponding to the plausible range of the location errors.

HMC offers further reductions in RMSE over KALE in roughly the same regions of the parameter space in which KALE improves over KILE, although the additional improvement is less dramatic. The maximum RMSE reduction we observe is about 28%, once again for a moderately smooth signal with larger magnitude location errors.



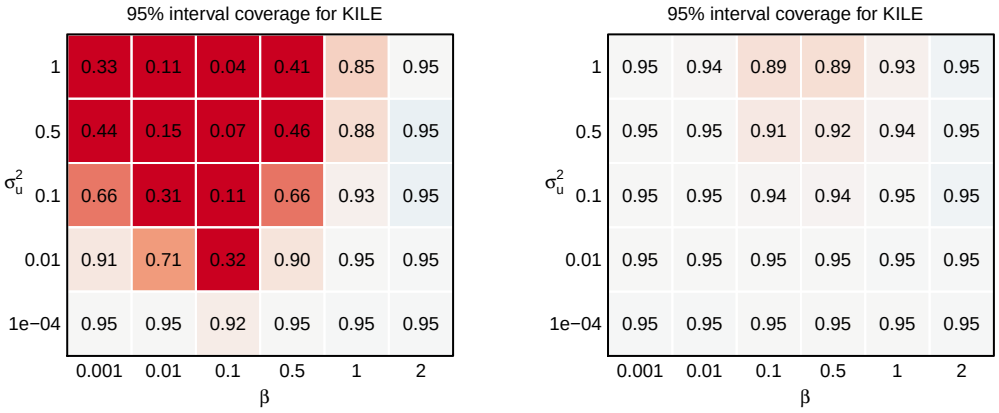
(a) RMSE ratio of KALE to KILE. *(b) RMSE ratio of HMC to KALE.*

Figure 3.7: Relative RMSE of KALE to KILE (A) and HMC and KALE (B) for each combination of parameters (β, σ_u^2) indicated, and $\sigma_x^2 = 0.1$. Blue shading represents a relative decrease in RMSE while red shading represents a relative increase in RMSE.

When the nugget variance σ_x^2 is increased (Figure 3.7 shows results for $\sigma_x^2 = 0.1$), differences

in RMSE among the three methods become smaller (the differences are wiped out entirely at $\sigma_x^2 = 1$, which is not pictured). This is not due to a shared σ_x^2 term in the RMSE value for all methods, as this is subtracted out. Rather, the similarity of all three methods reflects the fact that a larger nugget leaves less information in the data that can be effectively used for prediction. However, the differences that we do observe (both comparing KALE to KILE and HMC to KALE) occur primarily when the magnitude of location errors σ_u^2 is large.

In the case where all parameters are fixed and known, both KALE and HMC produce intervals with exact coverage (subject to Monte Carlo or numerical approximation errors) in all simulations. KILE, however, can severely undercover in the presence of location errors. Figure 3.8 shows coverage as low as 4% when the magnitude of the location errors is high ($\sigma_u^2 = 1$), $\beta = 0.1$, and the nugget variation is minimal ($\sigma_x^2 = 0.0001$). Undercoverage still persists in this region of the parameter space for $\sigma_x^2 = 1$, the largest nugget variance used in our simulations.



(a) $\sigma_x^2 = 0.0001$. *(b)* $\sigma_x^2 = 1$

Figure 3.8: 95% interval coverage for KILE for $\sigma_x^2 = 0.0001$ *(A) and* $\sigma_x^2 = 1$ *(B). With moderately smooth signals and large location errors, we see severe undercoverage that does not disappear even for* $\sigma_x^2 = 1$.

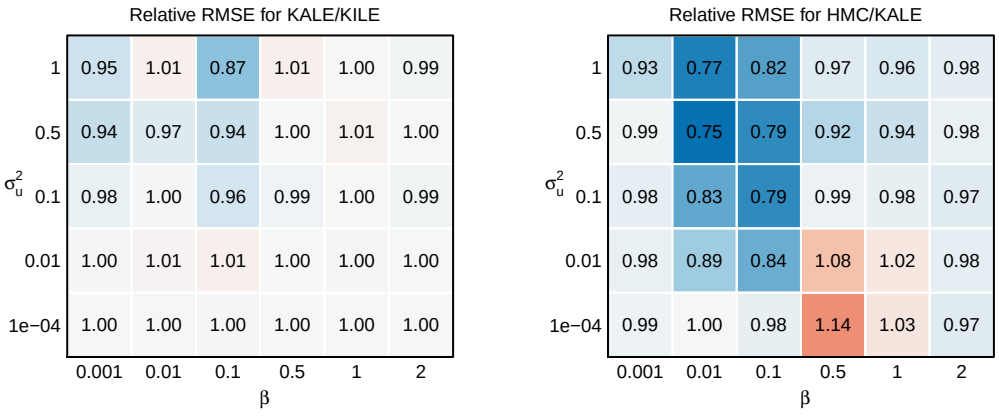
3.4.2 Unknown covariance parameters

In typical applied settings, the analyst will not know model parameters such as those of the covariance function (τ^2, β), the nugget variance σ_x^2, or even the variance of the location errors σ_u^2. Due to identifiability issues with our choice of covariance function in this simulation (3.5), we

assume σ_u^2 is known but estimate all other parameters before making predictions at unobserved locations.

For KILE and KALE, parameter estimation is accomplished through maximum (pseudo-) likelihood, as in (3.8). Parameter estimates are then plugged into Kriging equations (3.4)–(3.7) to obtain corresponding point and interval estimates. Because c and k are both squared exponential (3.5), the pseudolikelihood estimation procedure estimates the same covariance function for y, however the estimated parameters (and therefore Kriging equations, based on k^*) will differ. The plug-in approach ignores uncertainty in parameter estimates, so plug-in MSE estimates will be too optimistic. Various techniques exist for adjusting MSE from estimated parameters [Smith (2004); Zhu & Stein (2006)], though there is no need to incorporate such techniques into our analysis since exact (up to Monte Carlo error) MSEs are provided by simulation.

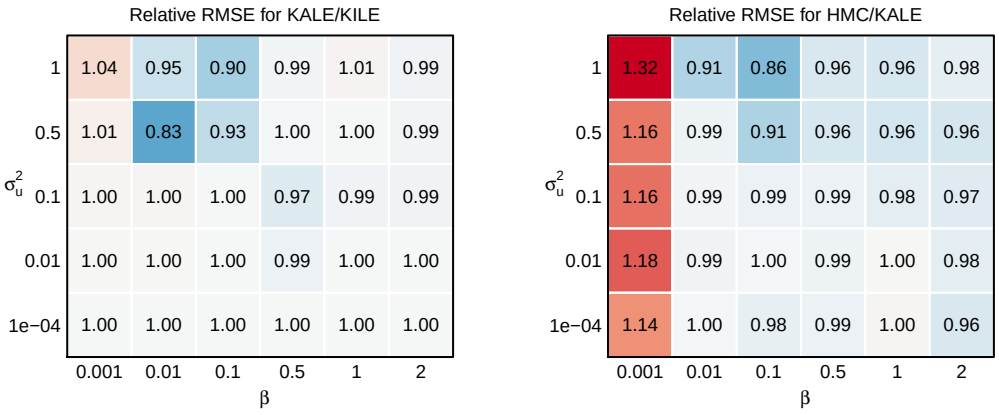
For HMC, we supply unknown parameters with prior distributions and sample parameters and predictions jointly from the posterior distribution $\pi(\theta, \mathbf{x}_k^*|\mathbf{y}_n)$. The priors we use are flat over a reasonable range (see Table 3.1), which guarantees both a proper posterior and a posterior mode that agrees with the maximum likelihood estimate of θ. This second condition supports fair comparisons between predictions derived from HMC parameter estimates versus those based on the maximum (psueolikelihood) parameter values.



(a) RMSE ratio of KALE to KILE. *(b) RMSE ratio of HMC to KALE.*

Figure 3.9: Relative RMSE of KALE and KILE (A) and HMC and KALE (B) for each combination of parameters (β, σ_u^2) indicated, and $\sigma_x^2 = 0.0001$. Parameters are assumed unknown and first estimated to obtain point predictions.

Figure 3.9 provides the relative RMSE of KALE vs KILE, and HMC vs KALE, for predictions when parameters must first be estimated (using $\sigma_x^2 = 0.0001$). We notice that there does not appear to be a great advantage in KALE over KILE when parameters are first estimated. This is because, as mentioned earlier, the marginal process y still has a squared exponential covariance function 3.5, so Kriging equations for KALE and KILE will be very similar. On the other hand, we notice a modest improvement when using HMC over Kriging, except in a small region of the parameter space ($\sigma_u^2 \leq .01$ and $\beta \in [0.5, 1]$).

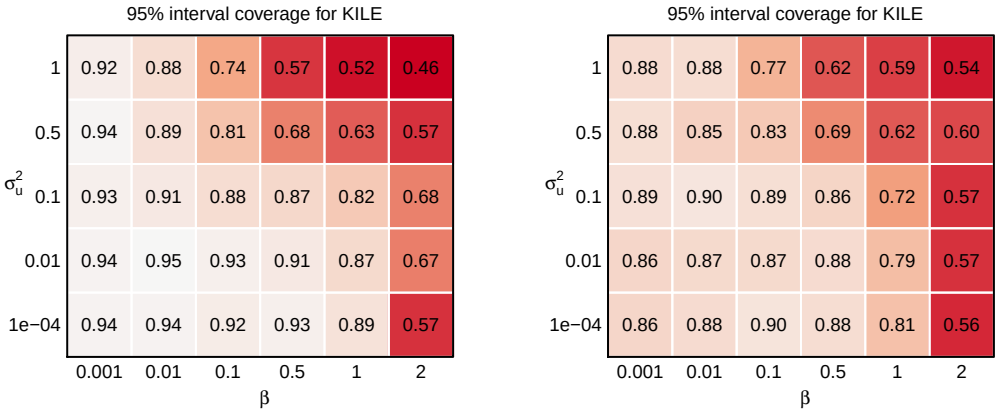


(a) RMSE ratio of KALE to KILE. *(b) RMSE ratio of HMC to KALE.*

Figure 3.10: Relative RMSE of KALE and KILE (A) and HMC and KALE (B) for each combination of parameters (β, σ_u^2) indicated, and $\sigma_x^2 = 0.1$. Parameters are assumed unknown and first estimated to obtain point predictions.

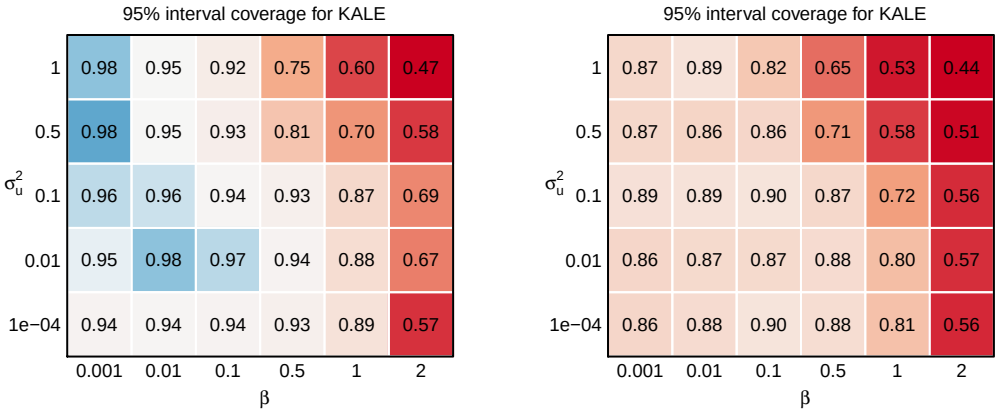
When the nugget variance is increased to $\sigma_x^2 = 0.1$, we see the results in Figure 3.10. We still see relatively similar performances from KALE and KILE. HMC offers a small improvement over KALE when $\beta \geq 0.01$, though for $\beta = 0.001$ we actually see significantly higher MSEs with HMC. At $\beta = 0.001$ the process is extremely smooth, as the most distant pairs of observations still have a correlation of 0.88. We are thus more concerned with overestimating β than underestimating it; as the former shrinks predictions towards 0 while the latter shrinks towards (approximately) the mean of all observations. As we use a flat prior for β, where almost all mass is located $\beta > .001$, the posterior tends to overestimate β, leading to draws with relatively high MSE.

Neither Kriging or HMC guarantees prediction intervals with the correct coverage in the regime



(a) $\sigma_x^2 = 0.0001$ *(b)* $\sigma_x^2 = 1$

Figure 3.11: 95% *interval coverage for KILE for* $\sigma_x^2 = 0.0001$ *(A) and* $\sigma_x^2 = 1$ *(B).*

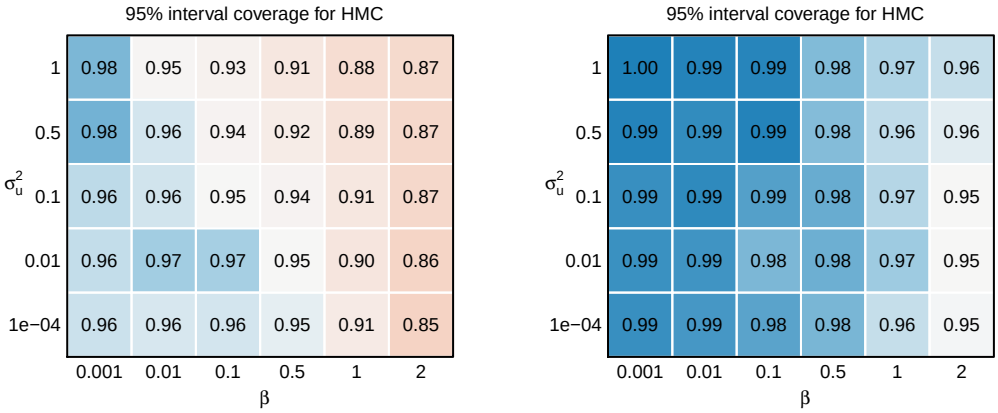


(a) $\sigma_x^2 = 0.0001$ *(b)* $\sigma_x^2 = 1$

Figure 3.12: 95% *interval coverage for KALE for* $\sigma_x^2 = 0.0001$ *(A) and* $\sigma_x^2 = 1$ *(B).*

where parameters must first be estimated*. We nevertheless present coverage results in Figures 3.11–3.13. While we don't expect any method used to provide exact coverage, Kriging (both KALE and KILE) suffer from significant undercoverage for some regions of the parameter space, while HMC is consistent in offering at least 85% coverage throughout our simulations. In a regime without location errors, Zimmerman & Cressie (1992) advocate Bayesian procedures under non-informative priors over frequentist procedures in order to obtain interval estimates with good coverage; our simulation results, albeit in the context of location errors, agree with this finding.

*Though HMC would give proper "Bayes coverage" when simulating θ according to the prior used.



(a) $\sigma_x^2 = 0.0001$ *(b)* $\sigma_x^2 = 1$

Figure 3.13: 95% *interval coverage for HMC for* $\sigma_x^2 = 0.0001$ *(A) and* $\sigma_x^2 = 1$ *(B).*

3.4.3 Summary

Our simulation results confirm the theoretical guarantee of KALE dominating KILE in prediction MSE when the covariance function is known, and furthermore HMC dominating KALE. The magnitude of differences in MSE between these methods is greatest when the process is moderately smooth relative to the spatial sampling (e.g. $0.01 \leq \beta \leq 0.5$), when the magnitude of location errors σ_u^2 is largest, and when nugget variation σ_x^2 is smallest. For such regions of the parameter space, KILE fails to deliver prediction intervals with proper coverage, whereas KALE and HMC can give valid prediction intervals for any parameter values.

An important consequence in adjusting for location errors with a known covariance function is the corresponding adjustment to the nugget. The discussion in Stein (1999) (Sections 3.6 and 3.7) emphasizes the importance of correctly specifying the high-frequency behavior of the process when interpolating (correctly specifying the low-frequency behavior is less crucial), including the nugget term. Estimating parameters, including the nugget term σ_x^2, implicitly corrects for model misspecification when ignoring location errors. Thus we see little difference in predictive performance between KALE and KILE when parameters are first estimated. Depending on the choice of prior, KALE/KILE may give lower MSE predictions than HMC, which averages over posterior parameter uncertainty; however, interval coverage is better for HMC (using weak prior information) than for KALE/KILE.

3.5 Interpolating Northern Hemisphere Temperature Anomolies

To illustrate the methods discussed in this paper, we consider interpolating northern hemisphere temperature anomolies during the summer of 2011 using the publicly available CRUTEM3v data set[†] [Brohan et al. (2006)]. Figure 3.14 shows our data. These data are used in geostatistical reconstructions of the Earth's temperature field, which interpolate temperatures at unobserved points in space-time in order to better understand the historical behavior of climate change (see, for example, Tingley & Huybers (2010) and Richard et al. (2012)). Each observation is a spatiotemporal average: temperature readings are averaged over the April–September period and each $5° \times 5°$ longitude-latitude grid cell. These values are then expressed as anomolies relative to the global average during the period 1850–2009, which is calculated using an ANOVA model [Tingley (2012)]. Apart from this spatiotemporal averaging, numerous other preprocessing steps adjust this data for differences in altitude, timing, equipment, and measurement practices between sites, along with other potential sources of error; please see Morice et al. (2012) and Jones et al. (2012) for more details.

Our analysis, restricted to interpolating a single year of data, and without using external data such as temperature proxies [Mann et al. (2008)], is intended as a proof of concept rather than as a refinement or improvement to existing analyses of these data. We wish to illustrate the potential impact of location errors on conclusions drawn from these data.

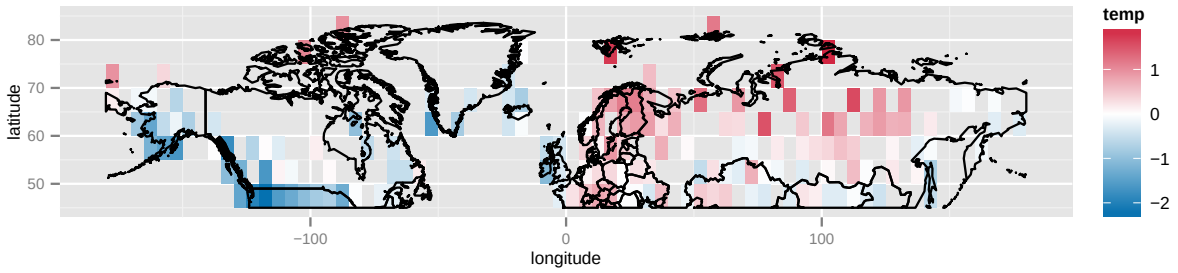


Figure 3.14: CRUTEM3v data for summer 2011, with 2011 mean subtracted so that measurements represent spatial anomolies. Generally speaking, we see lower (cooler) anomolies in North America and positive (warmer) anomolies in Europe. Higher latitudes also tend to have positive anomolies.

[†]http://www.cru.uea.ac.uk/cru/data/temperature/

The "gridding", or spatial averaging across $5^\circ \times 5^\circ$ cells, complicates analyses using Gaussian process models [Director & Bornn (2015)]. However, assuming a smooth temperature field, we know that the recorded spatial average must be realized exactly at some location in each grid box (closer to the center if a lot of points have been averaged together). This frames the spatial averaging problem as a location measurement error problem: instead of observing the temperature $x(s)$ at each grid center s, we observe the temperature at an unknown location displaced from the grid center: $y(s) = x(s + u)$.

Following Tingley & Huybers (2010), we assume an exponential covariance function for $x(s)$, where distance is calculated along the Earth's surface. As s is given in terms of longitude/latitude ($s = (\psi, \phi)$), this has the form

$$\begin{aligned} c(s_1, s_2) &= \tau^2 \exp(-\beta\Delta) + \sigma_x^2 \mathbf{1}_{s_1 = s_2} \\ \Delta &= 2r \arcsin \sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\psi_2 - \psi_1}{2}\right)}, \end{aligned} \tag{3.14}$$

where $r = 6371$ is the radius of the earth (in km). At higher latitudes (ϕ), the centers of each grid cell are closer together, so nearby observations are more strongly correlated. The nugget term σ_x^2 represents some combination of measurement error in temperature readings and high-frequency spatial variation that is inestimable using the gridded observation samples.

We assume the following model for location errors u_i, which are additive displacements of long/lat coordinates $s_i = (\psi_i, \phi_i)$:

$$u_i \sim \mathcal{N}\left(\mathbf{0}, \sigma_u^2 \left(\frac{180}{\pi r}\right)^2 \begin{pmatrix} \frac{1}{\cos^2(\phi_i)} & 0 \\ 0 & 1 \end{pmatrix}\right). \tag{3.15}$$

This prior is equivalent to assuming that distance along the Earth's surface (great-circle distance) between each grid center and the corresponding observation location is chi distributed, $d(s_i, s_i + u_i) \sim \sigma_u \chi_2$. Combining (3.15) and (3.14), we use Monte Carlo to compute k.

We treat parameters $\tau^2, \beta, \sigma_x^2$ as unknown, but fix $\sigma_u^2 = 7500$. At this value, the median magnitude of the location errors in great-circle distance is 102km, which consistent with analyzing

the coordinates of the temperature recording sites used to compile the CRUTEM3v data[‡].

3.5.1 Kriging

We first apply Kriging approaches to interpolate the CRUTEM3v data, both adjusting for and ignoring location errors (3.15). Because parameters $\tau^2, \beta, \sigma_x^2$ are unknown, we first need to estimate them using maximum likelihood (when ignoring location errors) or maximum pseudo-likelihood (3.8) (when adjusting for location errors). These can then be plugged in to covariance functions c and k to obtain "empirical" Kriging equations we can use for interpolation [Zimmerman & Cressie (1992)].

We small differences in parameter estimates when ignoring location errors (assuming $\sigma_u^2 = 0$) and adjusting for them (assuming $\sigma_u^2 = 7500$), summarized in Table 3.2. Consequently, when

σ_u^2	$\hat{\tau}^2$	$\hat{\beta}$	$\hat{\sigma}_x^2$
0	1.1671	1.4275×10^{-4}	0.0747
7500	1.1649	1.4677×10^{-4}	0.0699

Table 3.2: Covariance function parameter estimates when ignoring location errors (assuming $\sigma_u^2 = 0$) and adjusting for location errors (assuming $\sigma_u^2 = 7500$).

we interpolate data at the centers of grid cells for which no data was observed, we see differences between the KALE and KILE approaches. Figure 3.15 shows the differences between KALE and KILE interpolations (both point and interval estimates). Relative to the range of the data (most anomolies are in the interval $(-1, 1)$), the discrepency between KALE and KILE does not seem very significant.

3.5.2 HMC

Using HMC, parameter inference and interpolations are made simultaneously. The resulting point and interval predictions differ substantially from the Kriging results. However, because HMC incorporates parameter uncertainty in predictions, this comparison is not sufficient to illustrate the impact of location errors on conclusions from this data. A more appropriate comparison

[‡]Station locations are vieweable at https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/

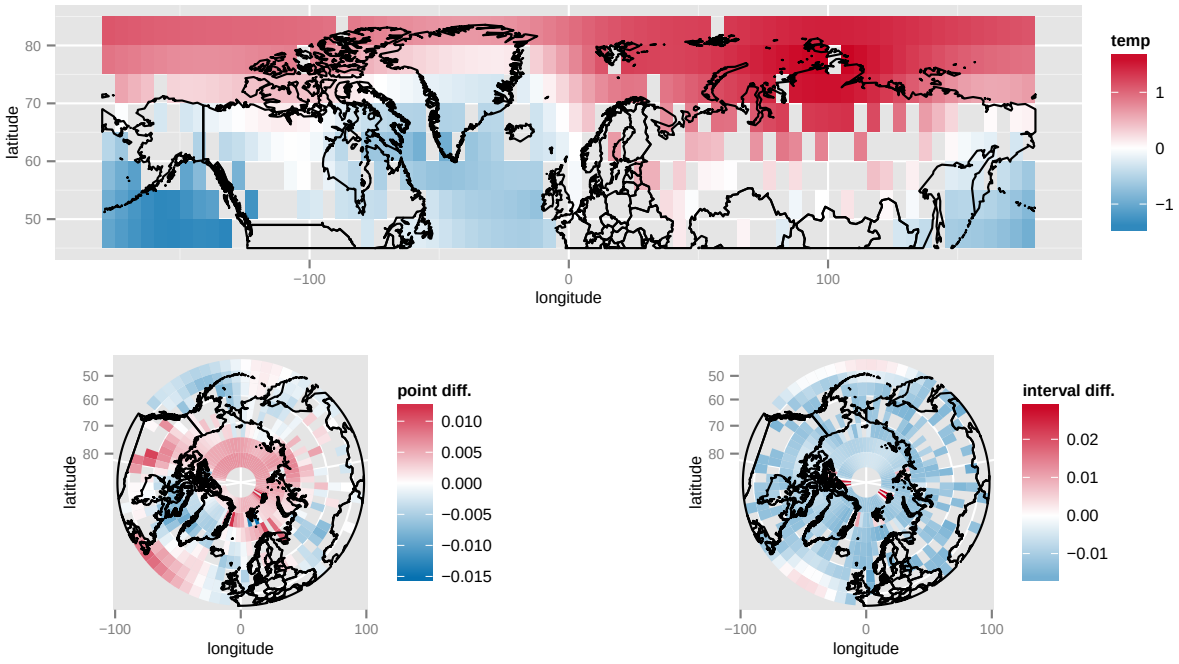


Figure 3.15: Kriging results for interpolating temperature anomolies from summer 2011. The top plot shows interpolations at unobserved grid centers given by KALE. The bottom left plot shows the difference in estimates between KALE and KILE (KALE – KILE), and the bottom right plot shows difference in 95% interval widths between KALE and KILE.

is between HMC with a location error model ($\sigma_u^2 = 7500$), and HMC assuming with no location errors ($\sigma_u^2 = 0$). These results are plotted in Figure 3.16.

Using HMC, accounting for location errors produces more significant differences in inference/prediction than was observed for Kriging. This is particularly true for interval predictions, where adjusting for location errors yields intervals as much as 0.1 wider, which is a significant discrepency when most observations lie in $(-1, 1)$.

Figure 3.17 shows posterior densities for unknown parameters of the covariance function based on HMC draws from the $\sigma_u^2 = 7500$ and $\sigma_u^2 = 0$ models (the Kriging estimates of these parameters are vertical lines). HMC under location error model ($\sigma_u^2 = 7500$) gives slightly larger β estimates than when using $\sigma_u^2 = 0$, meaning observations are inferred to be less strongly correlated. This yields prediction intervals that tend to be wider (see Figure 3.16). The most extreme descrepencies occur in the arctic, where distances between grid points are closest. The fact that modeling location errors adds additional uncertainty to arctic predictions is of particular inter-

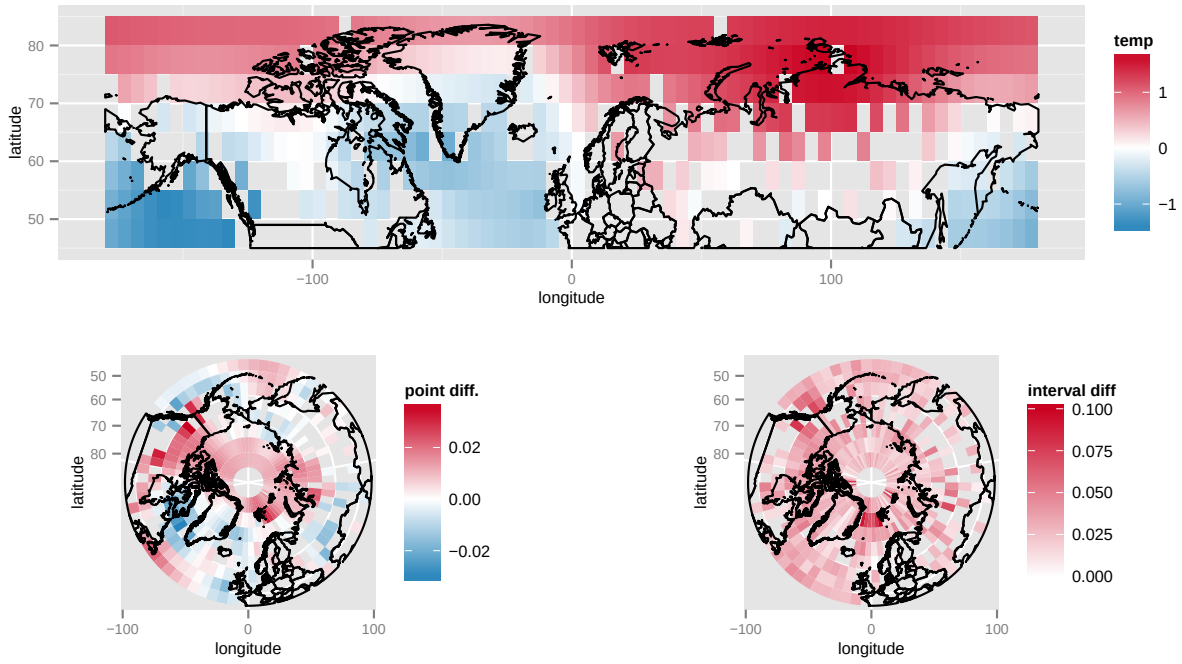


Figure 3.16: Results for interpolating temperature anomolies from summer 2011 using HMC. The top plot shows interpolations at unobserved grid centers, assuming location errors $\sigma_u^2 = 7500$. The bottom left plot shows the difference in estimates between the location error model and the model with $\sigma_u^2 = 0$. The bottom right plot shows difference in 95% interval widths.

est to climate scientists, as accurate climate reconstruction for the arctic region is essential for understanding recent climate change patterns [Cowtan & Way (2014)].

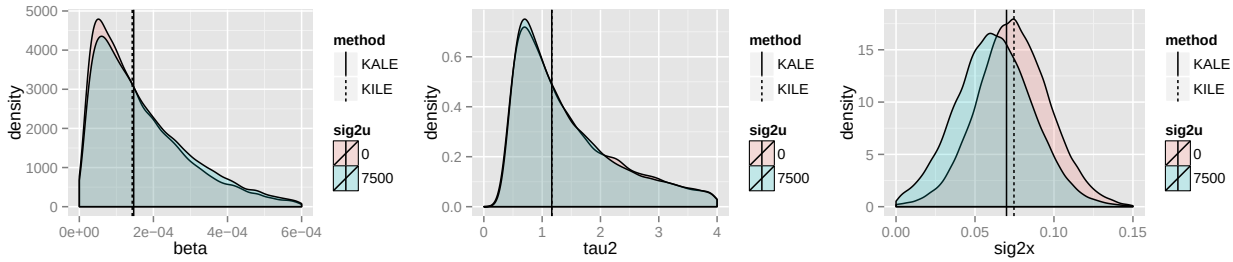


Figure 3.17: Density of posterior draws from HMC using $\sigma_u^2 = 7500$ (blue) and $\sigma_u^2 = 0$ (red). Point estimates of these parameters from Kriging (Table 3.2) are shown as vertical lines.

The difference between predictions obtained under the $\sigma^2 = 0$ and $\sigma_u^2 = 7500$ models using HMC suggests that modeling location errors, even when they are small in magnitude, meaningfully impacts parameter estimates and predictions at unobserved locations. The fact that results for HMC (assuming $\sigma_u^2 = 7500$) also differ from the results using KALE, while the KILE results do so less, demonstrates that moment procedures such as Kriging may be ineffective in adjusting

for these errors.

3.6 Conclusion

In this paper, we have explored the issue of Gaussian process regression when locations in the input space $\mathbb{S}$ are subject to error. Even when location errors are quite small in magnitude, it is essential to adjust Kriging equations in order to obtain good point and interval estimates; further improvements can be made by using MCMC to sample directly from the distribution of the measurement of interest given the sampled data.

Both MCMC and Kriging will be infeasible for large data sets, due to the cost of the covariance matrix inversion. A useful future study would be to adapt the procedures discussed in this paper to methods for inference and prediction for large spatial data sets, such as the predictive process approach [Banerjee et al. (2008)], low rank representations [Cressie & Johannesson (2008)], likelihood approximations [Stein et al. (2004)], and Markov random field approximations [Lindgren et al. (2011)]. It will also be useful to extend the analysis of this paper to regimes where location errors may be correlated with the process of interest x. For example, in climate data, regions with extreme climates will be harder to sample, thus there may be greater error in the spatial refencing of such sampling than for regions that are easier to sample.

A

Full Specification of Multiresolution Transition Models

In this appendix we provide full details on parametrizing and fitting the hierarchical models for the multiresolution models introduced in Sections 2.4 and 2.5. Our intent is to present our methodology with enough specificity that it could be implented and reproduced by readers with access to the data (the data is not publicly available) and appropriate computational resources. EPV represents a unique inferential challenge, as our data is high dimensional and the parameter space of our multiresolution transtion models—including spatial random effect surfaces for all ballcarrier and macrotransition type combinations—is extremely rich. Thus, even for readers not wishing to reproduce our results, this appendix may provide a valuable example of hierarchical

spatiotemporal modeling.

A.1 Macrotransition partial likelihood

As discussed in Section 2.6, parameters for macro and microtransition models are estimated separately using partial likelihoods (2.14). We now focus on inference for the macrotransition model, beginning with the partial likelihood. Following (2.14), the competing risks model (2.6)–(2.7) specifies a partial likelihood function for all unknown model components—$\boldsymbol{\beta}_j^\ell$ and ξ_j^ℓ for all players ℓ and macrotransitions j, as well as $\tilde{\xi}_j^\ell$ for $j \leq 4$. Let $\mathcal{T}^\ell$ comprise the time intervals for which player ℓ possesses the ball, with $\mathcal{T}_j^\ell \subset \mathcal{T}^\ell$ the time points at which a macrotransition of type j occurs. Then the (partial) likehood can be written

$$\begin{aligned} L(\boldsymbol{\beta}, \xi, \tilde{\xi}) &= \prod_{\text{possessions}} \left(\prod_{t=0}^{T-\epsilon} \mathbb{P}(M(t)^c | \mathcal{F}_t^{(Z)})^{\mathbf{1}[M(t)^c]} \prod_{j=1}^{6} \mathbb{P}(M_j(t) | \mathcal{F}_t^{(Z)})^{\mathbf{1}[M_j(t)]} \right) \\ &= \prod_{\ell} \prod_{j=1}^{6} \left(\prod_{t \in \mathcal{T}_j^\ell} \lambda_j^\ell(t) \right) \exp\left(- \int_{\mathcal{T}^\ell} \lambda_j^\ell(s) ds \right), \end{aligned} \tag{A.1}$$

with $(\boldsymbol{\beta}_j^\ell, \xi_j^\ell, \tilde{\xi}_j^\ell)$ parameterizing $\lambda_j^\ell(t)$ as in (2.7). This likelihood is identical to that of a model that assumes macrotransition events occur according to an inhomogeneous Poisson Process with intensity $\lambda_j^\ell(t)$ [Laird & Olivier (1981)]. Notice that (A.1) factors across player-macrotransition pairs, however exact likelihood inference is impossible for this model due to the infinite-dimensional spatial effect parameters ξ and $\tilde{\xi}$ contained in λ, over which we need to integrate in the second term of (A.1).

Since the observed data is discretized to every 1/25th of a second, we do not observe players' locations (and consequently, do not observe some of the time-referenced covariates) continuously, making it appropriate to replace the integral in the rightmost term in (A.1) with a sum over the finite collection of times at which data is observed. Let $\mathcal{T}_0^\ell$ index times at which data is observed but a macrotransition does not occur (in reality, macrotransitions occur almost surely between points at which data is observed, yet our data references them to the nearest time at which loca-

tions are recorded). Then the likelihood (A.1) is approximated by

$$L(\boldsymbol{\beta}, \xi, \tilde{\xi}) = \prod_{\ell} \prod_{j=1}^{6} \left(\prod_{t \in \mathcal{T}_j^\ell} \lambda_j^\ell(t) \exp(-\lambda_j^\ell(t)) \right) \exp\left(- \sum_{t \in \mathcal{T}_0^\ell} \lambda_j^\ell(t) \right), \tag{A.2}$$

which yields (A.1) in the limit as the temporal resolution of the data increases. While the intractable integral in (A.1) has been replaced by a sum in (A.2), evaluation of the likelihood may be extremely computationally expensive if $|\mathcal{T}_0^\ell|$ is large (depending on the functional form assumed for the spatial random effect surface). In our data set, for some players ℓ, $|\mathcal{T}_0^\ell|$ is as large as 300000, which makes evaluating (A.2) impossible when assuming a Gaussian process prior for ξ, as discussed in Appendix A.3.

A.2 Covariates

As revealed in (2.7), the hazards $\lambda_j^\ell(t)$ are parameterized by spatial effects (ξ_j^ℓ and $\tilde{\xi}_j^\ell$ for pass events), as well as coefficients for situation covariates, $\boldsymbol{\beta}_j^\ell$. The covariates used may be different for each macrotransition j, but we assume for each macrotransition type the same covariates are used across players ℓ.

Among the covariates we consider, `dribble` is an indicator of whether the ballcarrier has started dribbling after receiving possession. `ndef` is the distance between the ballcarrier and his nearest defender (transformed to $\log(1 + d)$). `ball_lastsec` records the distance traveled by the ball in the previous one second. `closeness` is a categorical variable giving the rank of the ballcarrier's teammates' distance to the ballcarrier. Lastly, `open` is a measure of how open a potential pass receiver is using a simple formula relating the positions of the defensive players to the vector connecting the ballcarrier with the potential pass recipient.

For $j \leq 4$, the pass event macrotransitions, we use `dribble`, `ndef`, `closeness`, and `open`. For shot-taking and turnover events, `dribble`, `ndef`, and `ball_lastsec` are included. Lastly, the shot probability model (which, from (2.9) has the same parameterization as the macrotransition model) uses `dribble` and `ndef` only. All models also include an intercept term.

A prior distribution is assumed for each coefficient for each macrotransition model jointly

across players; we provide this in A.4.

A.3 Spatial effects

The parameters ξ_j^ℓ for all players ℓ and macrotransitions $j \in \{1, \ldots, 6\}$, as well as $\tilde{\xi}_j^\ell$ for $j \leq 4$, are infinite-dimensional as they are functions from the court space $\mathbb{S}$ to $\mathbb{R}$. We assume each is a realization of a Gaussian process (sometimes called a Gaussian random field in the 2-dimensional case). Generally speaking, if ξ is a 0 mean Gaussian process with covariance function $C(\mathbf{z}, \mathbf{z}^*)$, then for locations $\mathbf{z}, \mathbf{z}^* \in \mathbb{S}$,

$$\begin{pmatrix} \xi(\mathbf{z}) \\ \xi(\mathbf{z}^*) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} C(\mathbf{z}, \mathbf{z}) & C(\mathbf{z}, \mathbf{z}^*) \\ C(\mathbf{z}^*, \mathbf{z}) & C(\mathbf{z}^*, \mathbf{z}^*) \end{pmatrix}\right). \tag{A.3}$$

The form of the joint distribution (A.3) extends to arbitrarily many $\mathbf{z} \in \mathbb{S}$, and naturally provides interpolation and uncertainty quantification for values of the spatial field at unobserved locations (see, e.g., Rasmussen (2006)). The covariance function, $C(\mathbf{z}, \mathbf{z}^*)$, is called isotropic if it is a function only of $\Delta = ||\mathbf{z} - \mathbf{z}^*||$. A common choice of isotropic covariance function is the Matérn, where

$$C(\Delta) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(\kappa\Delta)^\nu K_\nu(\kappa\Delta), \tag{A.4}$$

with K_ν being the modified Bessel function of the second kind and order $\nu > 0$, $\kappa > 0$ being a scaling parameter for the distance Δ, and σ^2 giving the marginal variance of any point $\xi(\mathbf{z})$. Typically, ν is fixed by the analyst, and κ and σ^2 are estimated from the data, perhaps in a Bayesian fashion [Neal (1997)]. Likelihood evaluations for parameters of the covariance function are $\mathcal{O}(n^3)$ where n is the number of spatially referenced observations associated with a particular field (assuming different parameter values for each field), which is prohibitively expensive for large data sets such as that for our macrotransition model (A.2).

An alternative to specifying a form of covariance function is to represent Gaussian processes using functional bases; that is, for $\phi_1, \ldots, \phi_d : \mathbb{S} \to \mathbb{R}$ and any $\mathbf{z} \in \mathbb{S}$,

$$\xi(\mathbf{z}) = \sum_{i=1}^{d} w_i \phi_i(\mathbf{z}), \tag{A.5}$$

with $\mathbf{w} = (w_1 \ w_2 \ \dots \ w_d)'$ and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Because the $\{\phi_i\}$ are fixed, this yields a finite representation of the process because the map ξ is completely determined by $\mathbf{w}$, which offers many computational advantages [Higdon (2002); Quiñonero-Candela & Rasmussen (2005)], notably that the $\mathcal{O}(n^3)$ cost of evaluating the likelihood of the covariance parameters reduces to at worst $\mathcal{O}(n^2 d + d^3)$, depending on the structure of $\boldsymbol{\Sigma}$. Denoting $\boldsymbol{\phi}(\mathbf{z}) = (\phi_1(\mathbf{z}) \ \phi_2(\mathbf{z}) \ \dots \ \phi_d(\mathbf{z}))'$, the representation (A.5) yields the covariance function $C(\mathbf{z}, \mathbf{z}^*) = \boldsymbol{\phi}(\mathbf{z})' \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{z}^*)$. In general, for fixed d, it is not possible to find bases $\phi_1, \dots, \phi_d$ and covariance matrix $\boldsymbol{\Sigma}$ such that the representation (A.5) is equivalent to a process specified using a Matérn covariance function, however it is possible to obtain very accurate approximations.

Following Lindgren et al. (2011), we assume a functional basis $\{\phi_i, i = 1, \dots, d\}$ induced by a triangular mesh of d vertices on the court space $\mathbb{S}$ (in practice, the triangulation is defined on a larger region that includes $\mathbb{S}$, due to boundary effects). The mesh is formed by partitioning $\mathbb{S}$ into triangles, where any two triangles share at most one edge or corner (see figure A.1 for an illustration). With some arbitrary ordering of the vertices of this mesh, $\phi_i : \mathbb{S} \to \mathbb{R}$ is the unique function taking value 0 at all vertices $j \neq i$, 1 at vertex i, and linearly interpolating between any two points within the same triangle used in the mesh construction. Thus, with this basis, fields ξ are piecewise linear on the triangles of the mesh.

As Lindgren et al. (2011) show, there are a couple advantages to using this particular functional basis for the spatial field representation (A.5) in addition to the generic computational advantages of having a discrete representation of an infinite-dimensional parameter. The first is that it is possible to find $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\nu, \kappa, \sigma^2)$ (for closed-form expression, see Lindgren et al. (2011)) with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ such that $\xi(\mathbf{z}) = \sum_{i=1}^{d} w_i \phi_i(\mathbf{z})$ closely approximates a Gaussian random field with Matérn covariance (A.4). The second advantage is that the precision $\boldsymbol{\Sigma}^{-1}$ is sparse, equivalent to a conditional independence structure for $\mathbf{w}$.

For the spatial fields in the macrotransition model, ξ_j^ℓ for all players ℓ and $j \in \{1, \dots, 6\}$, as well as $\tilde{\xi}_j^\ell$ for $j \leq 4$, we assume the representation (A.5) and the functional basis illustrated in Figure A.1. Reducing the spatial effects for each log-hazard model (2.7) to unknown d-vectors offers many computational benefits, and eases the implementation of hierarchical models that exploit the structural variation of our model parameters across players. To introduce the appro-

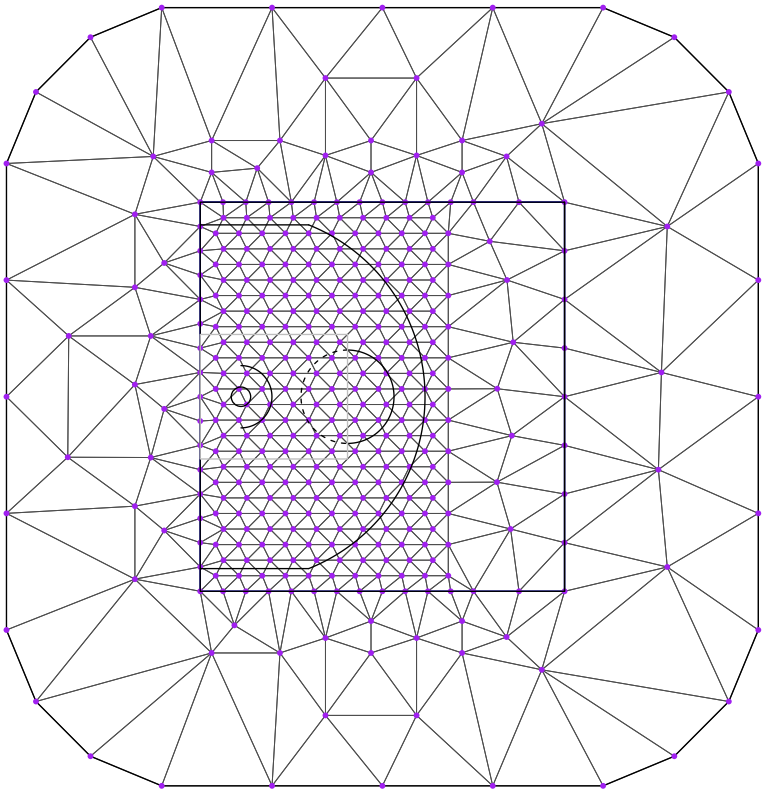


Figure A.1: Triangulation of $\mathbb{S}$ *used to build the functional basis* $\{\phi_i, i = 1, \ldots, d\}$*. Here,* $d = 383$*.*

priate notation, let $\xi_j^\ell(\mathbf{z}) = \boldsymbol{\phi}(\mathbf{z})'\mathbf{w}_j^\ell$, where $\mathbf{w}_j^\ell \in \mathbb{R}^d$ and $\boldsymbol{\phi}(\mathbf{z}) = (\phi_1(\mathbf{z})\ \phi_2(\mathbf{z})\ \ldots\ \phi_d(\mathbf{z}))'$ with $\phi_i : \mathbb{S} \to \mathbb{R}$ for $i = 1, \ldots, d$. Similiarly define $\tilde{\mathbf{w}}_j^\ell$ (loading the same basis functions $\{\phi_i\}$). Note that the basis functions ϕ_i are the same for each macrotransition j across all players, so that between-player variation in the spatial effects for each macrotransition components is provided by the weight vectors, $\mathbf{w}_j^\ell$.

This discrete representation of the spatial fields reduces the likelihood (A.2) to that of a Poisson regression, as

$$\log(\lambda_j^\ell(t)) = [\mathbf{W}_j^\ell(t)]'\boldsymbol{\beta}_j^\ell + \boldsymbol{\phi}(\mathbf{z}_\ell(t))'\mathbf{w}_j^\ell + \left(\tilde{\boldsymbol{\phi}}(\mathbf{z}_j(t))'\tilde{\mathbf{w}}_j^\ell \mathbf{1}[j \leq 4]\right) \tag{A.6}$$

is now linear in all unknown parameters.

A.4 Between-player structure

Beyond the information in the data for estimating the unknown components of (A.6) (which are, for $j \in \{1, \ldots, 6\}$, $\boldsymbol{\beta}_j^\ell$, $\mathbf{w}_j^\ell$, and for $j \leq 4$, $\tilde{\mathbf{w}}_j^\ell$), we have prior knowledge on their structure due to spatial smoothness and the natural clustering of players by team and by position. Our data includes position labels (e.g. center, point guard, power forward) for each player in the NBA. A player's position on a basketball team usually depends on his size and other physical attributes, as well as on his skill set and intended role in the team's strategic scheme. We therefore assume that parameters of players' macrotransition model cluster by position.

Rather than use the labeled positions in our data, we define position as a distribution of a player's location during his time on the court. Specifically, we divide the offensive half of the court into 4-square-foot bins (575 total) and count, for each player, the number of data points for which he appears in each bin. Then we stack these counts together into a $L \times 575$ matrix (there are $L = 461$ players in our data), denoted $\mathbf{G}$, and take the square root of all entries in $\mathbf{G}$ for normalization. We then perform non-negative matrix factorization on $\mathbf{G}$ in order to obtain a low-dimensional representation of players' court occupancy that still reflects variation across players [Miller et al. (2013)]. Specifically, this involves solving:

$$\hat{\mathbf{G}} = \underset{\mathbf{G}^*}{\operatorname{argmin}}\{D(\mathbf{G}, \mathbf{G}^*)\}, \text{ subject to } \mathbf{G}^* = \underset{L \times r}{\mathbf{U}} \underset{r \times 575}{\mathbf{V}} \text{ and } U_{ij}, V_{ij} \geq 0 \text{ for all } i, j, \quad \text{(A.7)}$$

where r is the rank of the approximation $\hat{\mathbf{G}}$ to $\mathbf{G}$ (we use $r = 5$), and D is some distance function, such as a Kullback-Liebler type

$$D(\mathbf{G}, \mathbf{G}^*) = \sum_{i,j} G_{ij} \log\left(G_{ij}/G_{ij}^*\right) - G_{ij} + G_{ij}^*.$$

The rows of $\mathbf{V}$ are non-negative basis vectors for players' court occupancy distributions and the rows of $\mathbf{U}$ give the loadings for each player. With this factorization, $\mathbf{U}_i$ (the ith row of $\mathbf{U}$) provides player i's "position"—a r-dimensional summary of where he spends his time on the court. Moreover, the smaller the difference between two players' positions, $||\mathbf{U}_i - \mathbf{U}_j||$, the more alike are their roles on their respective teams, and the more similar we expect the parameters of their

macrotransition models to be a priori.

Formalizing this, let $\mathbf{H}$ be a $L \times L$ matrix consisting of 0s, then set $H_{ij} = 1$ if player j is one of the eight closest players in our data to player i using the distance $||\mathbf{U}_i - \mathbf{U}_j||$ (the cutoff of choosing the closest eight players is arbitrary). This construction of $\mathbf{H}$ does not guarantee symmetry, which is required for the expressions that follow, thus we set $H_{ji} = 1$ if $H_{ij} = 1$. Let $n_i = \sum_{j=1}^{L} H_{ij}$ count the number of neighbors for player i. For any parameter $\boldsymbol{\theta}$, with θ_i being the value for player i, we assume a conditional autogressive model (CAR) [Besag (1974)]:

$$\theta_i | \theta_{(-i)}, \tau \sim \mathcal{N}\left(\frac{1}{n_i} \sum_{j:H_{ij}=1} \theta_j, \frac{\tau^2}{n_i} \right). \tag{A.8}$$

This can be expressed as a joint distribution for θ,

$$P(\boldsymbol{\theta}|\tau) \propto \tau^{-L} \exp\left(-\frac{1}{2\tau^2} \sum_{i,j:H_{ij}=1} (\theta_i - \theta_j)^2 \right), \tag{A.9}$$

which is improper ((A.9) is constant for constant location shifts of $\boldsymbol{\theta}$), though as a prior distribution will yield a proper posterior in typical applied settings [Besag et al. (1991)]. While $\mathbf{H}$ derives from the data and may therefore seem problematic as a component of a prior distribution, $\mathbf{H}$ is ancillary for the macrotransition parameters. As seen in (A.1), the macrotransition likelihood conditions on players' locations, modeling only their decisions at the locations they occupy. The CAR prior thus plays a key role in our estimation of the parameters of the macrotransition model, allowing for information sharing between players and easing inference for situations in which player-specific data is sparse.

A.5 Parameter estimation for the macrotransitions

With the likelihood (A.2), finite representation for the Gaussian random fields (A.5), and prior form (A.9) introduced, we now connect these components together and discuss inference.

Each unknown coefficient β of the macrotransition model is assumed the CAR prior structure. Specifically, we assume the vector $\boldsymbol{\beta}_{j,i} = (\beta_{j,i}^1 \ \beta_{j,i}^2 \ \dots \ \beta_{j,i}^L)'$ has the distribution given in (A.9) for the generic parameter $\boldsymbol{\theta}$, and that $\{\boldsymbol{\beta}_{j,i}, i = 1, \dots, p_j \text{ and } j = 1, \dots, 6\}$ are independent a

priori. We may analogously define $\mathbf{w}_{j,i} = (w^1_{j,i} \;\; w^2_{j,i} \;\; \ldots \;\; w^L_{j,i})'$, the vector of the loadings across players for the ith basis function for the jth macrotransition model's spatial field. While we may expect, a priori, components of this vector to covary according to a CAR structure, we also assume spatial covariance among the loadings for any particular player-macrotransition model, i.e., $\mathbf{w}^\ell_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_j = \mathbf{\Sigma}(1, \kappa_j, \sigma^2_j))$ as discussed in Appendix A.3. This suggests a Kronecker structure for the prior covariance of the basis loadings across players and space.

The computation demands of fitting such a model are prohibitive. Even assuming prior independence of all parameters across macrotransition types, $\mathbf{w}_j = \begin{pmatrix} \mathbf{w}^1_j & \ldots & \mathbf{w}^L_j \end{pmatrix}'$ would enter the model as a Ld-dimensional random effect ($Ld = 176563$ in our current specification), with an unknown covariance matrix itself parameterized by $\tau^2_j, \kappa_j, \sigma^2_j$. The conditional independence structure implied by the CAR model (A.8) does not hold when spatial structure is included, as all components of $\mathbf{w}^\ell_j$ for all ℓ depend on κ_j and σ^2_j. Coarsening the mesh that induces the functional basis $\boldsymbol{\phi}$ reduces d, relieving some computational burden, but also impacting our ability to detect small-scale spatial variation.

Our approach is to consider h-vectors $\mathbf{v}^\ell_j \in \mathbb{R}^h$ where the ith component is a linear combination of the d components of $\mathbf{w}^\ell_j$: $v^\ell_{j,i} = \sum_{m=1}^k a^j_{i,m} w^\ell_{j,m}$. The idea here is that $\mathbf{v}^\ell_j$ is an h-dimensional representation of the d-dimensional vector of loadings $\mathbf{w}^\ell_j$. With a rough estimate of the matrix $\mathbf{w}_j$, the weights $a^j_{i,m}$ can be estimated by matrix factorization, such as SVD or NMF. This is exactly the route we take. For each macrotransition j, we estimate $\mathbf{w}^\ell_j$ independently for each player ℓ by fitting the jth macrotransition model (A.2) and (A.6) using the spatial prior $\mathbf{w}^\ell_j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\nu = 1, \kappa, \sigma^2))$ and vague priors on $\boldsymbol{\beta}^\ell_j, \kappa, \sigma^2$. This was done using the software R-INLA (`www.r-inla.org`), which uses integrated nested Laplace approximations [Rue et al. (2009)] for approximate Bayesian inference for generalized linear models with latent Gaussian Markov parameters, along with possibly non-Gaussian hyperparameters. Estimates $\mathbf{w}^1_j, \mathbf{w}^2_j, \ldots, \mathbf{w}^L_j$ were stacked on top of each other as row vectors to form $\mathbf{w}_j$, which was then exponentiated and factored using NMF with KL loss for $h = 10$ ($r = 10$ in the notation of (A.7)). Using NMF instead of SVD gives slightly better out-of-sample predictive results for the final estimated macrotransition model; note also that components of $\mathbf{w}_j$ enter the likelihood (and log-likelihood) exponentiated.

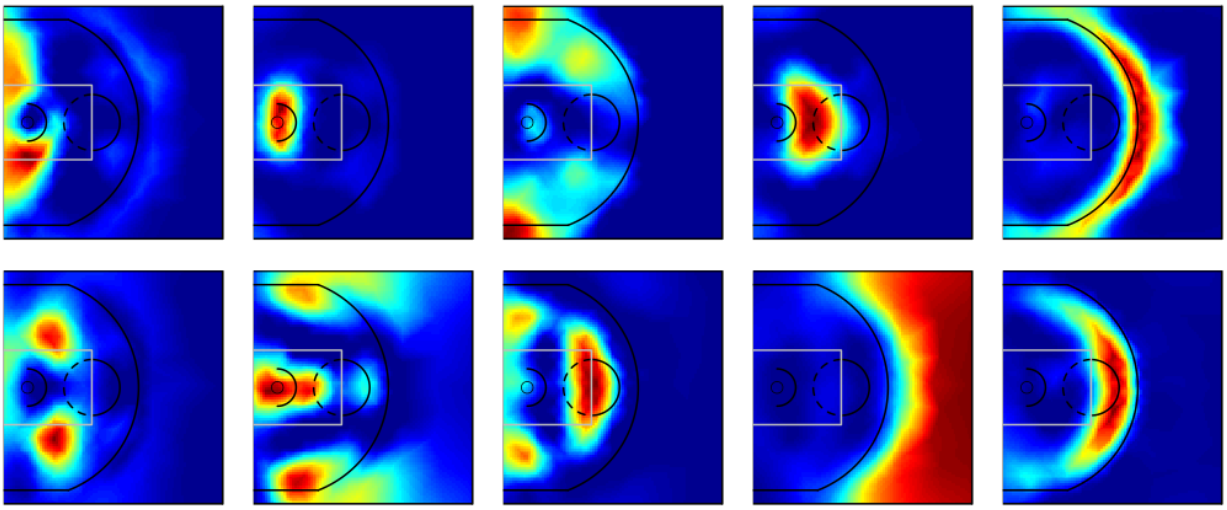


Figure A.2: The functional bases $\psi_{j,i}$ for $i = 1, \ldots, 10$ and j corresponding to the shot-taking macrotransition. Unlike SVD, there is no interpretation of the ordering of the bases. These bases correspond to well-known basketball motifs, such as layups (2), three point shots (3 and 5), and perimeter shots (6, 7, and 10).

Our construction of $\mathbf{v}_j^\ell$ implies a new functional basis representation for ξ_j^ℓ. Since $v_{j,i}^\ell = \sum_{m=1}^d a_{i,m}^j w_{j,m}^\ell$, for any $\mathbf{z} \in \mathbb{S}$ and $i \in \{1, \ldots, h\}$ we may write $\psi_{j,i}(\mathbf{z}) = \sum_{m=1}^d a_{i,m}^j \phi_m(\mathbf{z})$. $\boldsymbol{\psi}_j = (\psi_{j,1}\ \psi_{j,2}\ \ldots\ \psi_{j,h})'$ now provides a new functional basis corresponding to the loadings $\mathbf{v}_j^\ell$ for each player ℓ—note that after incorporating the weights $a_{i,m}^j$ into the basis functions, we no longer tether the definition of $\mathbf{v}_j^\ell$ to these weights. Correspondingly, the basis functions $\boldsymbol{\psi}_j$ depend on j—meaning they differ by macrotransition type—unlike the basis functions $\boldsymbol{\phi}$. Similar to Miller et al. (2013), these functional bases allow for spatial fields that are not simply locally smooth, but smooth across regions where players employ similar strategies—for an illustration of this, see Figure A.2. Thus, they represent information sharing across players and across space. Note that for pass events, we analogously construct $\tilde{\boldsymbol{\psi}}_j$ and $\tilde{\mathbf{v}}_j^\ell$ for all ℓ and $j \leq 4$.

Besides dimension reduction, another advantage to the new functional basis $\boldsymbol{\psi}_j$ is that, due to obtaining weights $a_{i,m}^j$ by NMF, the components of $\mathbf{v}_j^\ell$ may be assumed to be uncorrelated a priori*, both within and across players ℓ. The linear combinations that comprise $\boldsymbol{\psi}_j$ already provide spatial covariation. This allows us to model $\mathbf{v}_{j,i} = (v_{j,i}^1\ v_{j,i}^2\ \ldots\ v_{j,i}^L)'$ identically to other model parameters $\boldsymbol{\beta}_{j,i}$; namely, we assume $\mathbf{v}_{j,i}$ have prior structure given by (A.9), and that $\mathbf{v}_{j,i}$

*This is not a theoretical property of NMF, but approximately holds due to the similarity in results from NMF and SVD.

are a priori independent across j and i (as are $\tilde{\mathbf{v}}_{j,i}$).

To summarize, we rewrite the model components introduced in this section and concisely reveal the form of the posterior:

$$
\begin{aligned}
&\text{(likelihood)} && L(\boldsymbol{\beta}, \mathbf{v}, \tilde{\mathbf{v}}) = \prod_{\ell} \prod_{j=1}^{6} \left(\prod_{t \in \mathcal{T}_j^\ell} \lambda_j^\ell(t) \right) \exp\left(- \int_{\mathcal{T}^\ell} \lambda_j^\ell(s) ds \right) \\
&\text{where} && \log(\lambda_j^\ell(t)) = [\mathbf{W}_j^\ell(t)]' \boldsymbol{\beta}_j^\ell + \boldsymbol{\psi}_j(\mathbf{z}_\ell(t))' \mathbf{v}_j^\ell + \left(\tilde{\boldsymbol{\psi}}_j(\mathbf{z}_j(t))' \tilde{\mathbf{v}}_j^\ell \mathbf{1}[j \leq 4] \right), \\
&(\boldsymbol{\beta} \text{ prior}) && P(\boldsymbol{\beta}_{j,i} | \tau_{j,i}^\beta) \propto \left(\tau_{j,i}^\beta \right)^{-L} \exp\left(- \frac{1}{2(\tau_{j,i}^\beta)^2} \sum_{q,r: H_{qr}=1} (\beta_{j,i}^q - \beta_{j,i}^r)^2 \right) \\
& && \qquad \text{for } i = 1, \ldots, p_j \text{ and } j = 1, \ldots, 6, \\
&(\mathbf{v} \text{ prior}) && P(\mathbf{v}_{j,i} | \tau_{j,i}^v) \propto \left(\tau_{j,i}^v \right)^{-L} \exp\left(- \frac{1}{2(\tau_{j,i}^v)^2} \sum_{q,r: H_{qr}=1} (v_{j,i}^q - v_{j,i}^r)^2 \right) \\
& && \qquad \text{for } i = 1, \ldots, h \text{ and } j = 1, \ldots, 6, \\
&(\tilde{\mathbf{v}} \text{ prior}) && P(\tilde{\mathbf{v}}_{j,i} | \tau_{j,i}^{\tilde{v}}) \propto \left(\tau_{j,i}^{\tilde{v}} \right)^{-L} \exp\left(- \frac{1}{2(\tau_{j,i}^{\tilde{v}})^2} \sum_{q,r: H_{qr}=1} (\tilde{v}_{j,i}^q - \tilde{v}_{j,i}^r)^2 \right) \\
& && \qquad \text{for } i = 1, \ldots, h \text{ and } j = 1, \ldots, 4, \\
&\text{and } (\tau \text{ prior}) && P(\tau_{j,i}^*) \propto \left(\tau_{j,i}^* \right)^{-2} \exp\left(\frac{-1}{\tau_{j,i}^*} \right) \text{ for all } j, \, i, \text{ and } * = \beta, v, \tilde{v}.
\end{aligned}
\tag{A.10}
$$

This looks daunting, but it is just a mixed-effects Poisson regression, with a very particular structure for the random effects. However, due to the scale of the data, considerable computational resources are needed for inference. The posterior factors across macrotransition models j, thus we can estimate components corresponding to each macrotransition model separately. Each macrotransition model has over 10 million data points entering the likelihood (this is the size of $\sum_\ell (|\mathcal{T}_j^\ell| + |\mathcal{T}_0^\ell|)$), and at least 6006 parameters, corresponding to $(L+1)(p_j + h)$ for $L = 461$, $p_j = 3$ (all macrotransitions have $p_j \geq 3$), and $h = 10$ (the $L+1$ term accounts for τ as well). Approximate Bayesian inference was performed using R-INLA.

A.6 Parameter estimation for microtransitions

Both μ_x and μ_y are assumed to be realizations of Gaussian processes with Matérn covariance (A.4) with $\nu = 1$, though approximated with the functional basis (A.5) used in the macrotransition model and illustrated in Figure A.1. Like the spatial fields in the macrotransition model, we use R-INLA to fit the microtransition model (2.12) independently for $x(t)$ and $y(t)$. Note that for each player, separate models are fit to predict his motion during times he is the ballcarrier and times he is not the ballcarrier but still on offense. There are thus $L \times 2 \times 2$ microtransition models of the form (2.12)—one for each of L players, two situations (carrying ball and not carrying ball) and two dimensions (x and y). There is no information sharing between models; in principle we can imagine different components are connected a priori, yet the data is so informative that any appropriate prior would not be influential. Unlike the macrotransition model, where we may only observe a handfull of macrotransitions depending on the player and macrotransition type (e.g. turnovers are fairly rare for all players), all players are constantly moving, so the data alone are sufficient for precise inference.

B

EPV-Derived Quantities

While detailed studies of EPV curves and multiresolution transitions, as in Section2.7, are the most immediate and impactful application of EPV, we may also consider metrics that aggregate a season's worth of EPV curves.

B.1 EPV-Added

EPV-Added (EPVA) quantifies a player's overall offensive value of all of his movements and decisions while handling the ball, relative to the estimated value contributed by a league-average player receiving ball possession in the same situations. The notion of *relative* value is important because the martingale structure of EPV (ν_t) prevents any meaningful aggregation of EPV across a specific player's possessions; for instance, $\mathbb{E}[\nu_t - \nu_{t+\epsilon}] = 0$ for all t, meaning that *on average*

EPV does not change during any specific player's ball handling. For instance, while we see the EPV skyrocket after LeBron James receives the ball and eventually attack the basket in Figure 2.2, the definition of EPV prevents such increases being observed on average. James does not always attack the basket given the spatial situation he encountered when receiving the ball, and even when he does, he does not always beat the defense and gain a clear lane to the basket.

To calculate the baseline EPV at any time point for a league average player, we start by considering an alternate version of the transtion probability matrix between coarsened states $\mathbf{P}$. For each player $\ell_1, \ldots, \ell_5$ on offense, there is a disjoint subset of rows of $\mathbf{P}$, denoted $\mathbf{P}_{\ell_i}$, that correspond to possession states for player ℓ_i. Each row of $\mathbf{P}_{\ell_i}$ is a probability distribution over transitions in $\mathcal{C}$ given possession in a particular state. Technically, since states in $\mathcal{C}_{\text{poss}}$ encode player identities, players on different teams do not share all states which they have a nonzero probability of transitioning to individually. To get around this, we remove the columns from each $\mathbf{P}_{\ell_i}$ corresponding to passes to players not on player ℓ_i's team, and reorder the remaining columns according to the position (guard, center, etc.) of the associated pass recipient. Thus, the interpretation of transition distributions $\mathbf{P}_{\ell_i}$ across players ℓ_i is as consistent as possible. We create a baseline transition profile of a hypothetical league-average player by averaging these across all players: (with slight abuse of notation) let $\mathbf{P}_r = \sum_{\ell=1}^{L} \mathbf{P}_\ell / L$. Using this, we create a new transition probability matrix $\mathbf{P}_r(\ell)$ by replacing player ℓ's transition probabilities ($\mathbf{P}_\ell$) with the league-average player's ($\mathbf{P}_r$). The baseline (league-average) EPV at time t is then found by evaluating $\mathbb{E}_{\mathbf{P}_r(\ell)}[h(C_T)|C_t]$, which contrasts with the coarsened approximation, $\mathbb{E}_{\mathbf{P}}[h(C_T)|C_t]$, to ν_t. Denote this baseline (coarsened) EPV $\nu_t^{r(\ell)} = \mathbb{E}_{\mathbf{P}_r(\ell)}[h(C_T)|C_t]$.

If player ℓ has possession of the ball at time t_1 until time t_2, the quantity $\nu_{t_2} - \nu_{t_1}^{r(\ell)}$ estimates the value contributed player by ℓ relative to a league-average player during his ball possession. We calculate EPVA for player ℓ (EPVA(ℓ)) by summing such differences over all a player's touches (and dividing by the number of games played by player ℓ to provide standardization):

$$\text{EPVA}(\ell) = \frac{1}{\#\text{ games for } \ell} \sum_{\{t_s, t_e\} \in \mathcal{T}^\ell} \nu_{t_e} - \nu_{t_s}^{r(\ell)} \tag{B.1}$$

where $\mathcal{T}^\ell$ contains all intervals of form $[t_s, t_e]$ that span player ℓ's ball possession.

It must first be noted that for any $[t_s, t_e] \in \mathcal{T}^\ell$, $\mathbb{E}[\nu_{t_e} - \nu_{t_s}^{r(\ell)}] = \mathbb{E}[\nu_{t_s} - \nu_{t_s}^{r(\ell)}]$ due to ν_t being a martingale. However, as defined, EPVA(ℓ) sums over terms with more variation, since ν_t is more variable later in the possession. This is helpful for identifying particular plays in which players accrue very high (low) EPVA. Secondly, the choice to average over games implicitly rewards players who have high usage, even if their value added per touch might be low. Often, one-dimensional offensive players accrue the most EPVA per touch since they only handle the ball when they are uniquely suited to scoring; for instance, some centers (such as Miami's Chris Andersen) only receive the ball right next to the basket, where their height offers a considerable advantage for scoring over other players in the league. Thus, averaging by game—not touch—balances players' efficiency per touch with their usage and importance in the offense. Lastly, using coarsened EPV as a baseline $\nu_t^{r(\ell)}$ exploits the fact that, when averaging possessions over the entire season, the results are (in expectation) identical to using full-resolution EPV, assuming corresponding multiresolution transition probability models for this hypothetical league-average player—a consequence of (2.10).

Table B.1 provides a list of the top and bottom 10 ranked players by EPVA using our 2013-14 data, which is complete until February 7, 2014. Generally, players with high EPVA effectively adapt their decision-making process to the spatiotemporal circumstances they inherit when gaining possession. They receive the ball in situations that are uniquely suited to their abilities, so that on average the rest of the league is less successful in these circumstances. Players with lower EPVA are not necessarily "bad" players in any conventional sense; their actions simply tend to lead to fewer points than other players given the same options. Of course, EPVA provides a limited view of a player's overall contributions since it does not quantify players' actions on defense, or other ways that a player may impact EPV while not possessing the ball (though EPVA could be extended to include these aspects).

As such, we stress the idea that EPVA is not a best/worst players in the NBA ranking. Analysts should also be aware that the "league-average player" being used as a baseline is completely hypothetical, and we heavily extrapolate our model output by considering value calculations assuming this nonexistant player possessing the ball in all the situations encountered by an actual NBA player. The extent to which such an extrapolation is valid is a judgment a basketball ex-

Player	EPVA	Player	EPVA
Dirk Nowitzki	6.08	Ricky Rubio	-0.07
Kevin Durant	6.08	Luke Ridnour	0.18
Jose Calderon	5.33	Tayshaun Prince	0.26
Damian Lillard	5.28	Shaun Livingston	0.38
Kevin Love	5.13	Beno Udrih	0.47
Stephen Curry	4.63	P.J. Tucker	0.55
Channing Frye	4.58	Al-Farouq Aminu	0.59
Kyle Lowry	4.50	Andre Miller	0.68
Paul George	4.40	Gerald Henderson	0.71
LeBron James	4.38	Cody Zeller	0.71

Table B.1: Top 10 and bottom 10 players by EPV-added (EPVA) in 2013-14 (per game, minimum 500 touches during season).

pert can make. Alternatively, one can consider EPV-added over *specific* players (assuming player ℓ_2 receives the ball in the same situations as player ℓ_1), using the same framework developed for EPVA. Such a quantity may actually be more useful, particularly if the players being compared play similar roles on their teams and face similar situations (and the degree of extrapolation is minimized).

B.2 Shot Satisfaction

Another EPV-derived player metric we consider is called *shot satisfaction*. For each shot attempt a payer takes, we wonder how satisfied the player is with his decision to shoot—what was the expected point value of his most reasonable passing option at the time of the shot? If for a particular player, the EPV measured at his shot attempts is higher than the EPV conditioned on his possible passes at the same time points, then by shooting the player is consistently making the best decision for his team. On the other hand, players with pass options at least as valuable as shots should regret their shot attempts (we term "satisfaction" as the opposite of regret) as passes in these situations have higher expected value.

Specifically, we calculate

$$\text{SATIS}(\ell) = \frac{1}{|\mathcal{T}^{\ell}_{\text{shot}}|} \sum_{t \in \mathcal{T}^{\ell}_{\text{shot}}} \nu_t - \mathbb{E}\left[h(C_T) \mid \bigcup_{j=1}^{4} M_j(t) \right] \tag{B.2}$$

where $\mathcal{T}^{\ell}_{\text{shot}}$ indexes times a shot attempt occurs, $\{t : M_5(t)\}$, for player ℓ. Recalling that macro-

Player	Shot Satisfaction
Jose Calderon	0.34
Martell Webster	0.34
Spencer Hawes	0.33
Andre Iguodala	0.33
Channing Frye	0.32
Kyle Lowry	0.32
Mike Miller	0.31
Marvin Williams	0.31
Kyle Korver	0.30
Jodie Meeks	0.29

Player	EPVA
Ricky Rubio	-0.01
Tayshaun Prince	0.00
DeMar DeRozan	0.03
LaMarcus Aldridge	0.04
Tyreke Evans	0.05
Shaun Livingston	0.05
Gerald Henderson	0.05
Kevin Garnett	0.06
Jarrett Jack	0.06
Anthony Davis	0.07

Table B.2: Top 10 and bottom 10 players by shot satisfaction in 2013-14 (per game, minimum 500 touches during season).

transitions $j = 1, \ldots, 4$ correspond to pass events (and $j = 5$ a shot attempt), $\bigcup_{j=1}^{4} M_j(t)$ is equivalent to a pass happening in $(\epsilon, t + \epsilon]$. Unlike EPVA, pass satisfaction SATIS(ℓ) is expressed as an average per shot (not per game), which favors player such as three point specialists, who often take fewer shots than their teammates, but do so in situations where their shot attempt is extremely valuable. Table B.2 provides the top/bottom 10 players in shot satisfaction for our 2013-14 data. While players who attempt many three-pointers (e.g. Calderon, Miller, Korver) and/or players shots near the basket (e.g. Iguodala) have the most shot satisfaction, players who primarily take mid-range or long-range two pointers (e.g. Aldridge, Garnett) or poor shooters (e.g. Rubio, Prince) have the least. However, because almost all shot satisfaction numbers are positive, players still shoot relatively efficiently—almost every player generally helps his team by shooting rather than passing in the same situations, though some players do so more than others.

C

Some Proofs

This Appendix provides additional proof details for results derived in this dissertation.

C.0.1 Proof of Proposition 3.2.1

k is a valid covariance function if and only if for all n, $\mathbf{s}_n$, and $\{a_i \in \mathbb{R}, i = 1, \ldots, n\}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(s_i, s_j) \geq 0.$$

From (3.3), this condition can be rewritten:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(s_i, s_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \int_{\mathbb{S}} c(s_i + u_i, s_j + u_j) dg_{\mathbf{s}_n}(\mathbf{u}_n) \\ &= \int_{\mathbb{S}} \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j c(s_i + u_i, s_j + u_j) dg_{\mathbf{s}_n}(\mathbf{u}_n)\end{aligned}$$

As c is a valid covariance function, the integrand in this expression is always non-negative, so the integral is also non-negative. Thus k is a valid covariance function.

Note that for the common scenario where location errors are independent, so that $g_{\mathbf{s}_n}$ is a product measure $g_{s_1} \times \ldots \times g_{s_n}$, then Proposition 3.2.1 is a special case of kernel convolution [Rasmussen (2006)].

C.0.2 Proof of Proposition 3.2.2

Without loss of generality, we can assume $\tau^2 = 1$ and fix β, Δ arbitrarily. Using the fact that $k^*(s, s^*) = \mathbb{E}[\exp(-\beta||s + u - s^*||^2)]$ evaluates the moment generating function of a noncentral χ^2_p random variable $||s + u - s^*||^2$, we get that

$$\mathbb{E}[(\hat{x}_{\text{KALE}}(s^*) - x(s^*))^2] = 1 - \left(\frac{1}{1 + 2\beta\sigma_u^2}\right)^p \exp\left(\frac{-2\beta\Delta^2}{1 + 2\beta\sigma_u^2}\right).$$

Call this quantity $c(\sigma_u^2)$. Differentiating, we get

$$c'(\sigma_u^2) = \frac{2\beta[2\beta(p\sigma_u^2 - \Delta^2) + p]}{(1 + 2\beta\sigma_u^2)^{p+2}} \exp\left(\frac{-2\beta\Delta^2}{1 + 2\beta\sigma_u^2}\right).$$

If $\beta\Delta^2 \leq p/2$, then $c'(\sigma_u^2) > 0$ for all $\sigma_u^2 > 0$. Since $c(\sigma_u^2)$ is left continuous at 0, continuous on $\mathbb{R}_+$, and $c(0) = c_0$, this means $\beta\Delta^2 \leq p/2$ implies $c(\sigma_u^2) \geq c_0$ for all σ_u^2.

Otherwise, if $\beta\Delta^2 \leq p/2$, then for all $0 < \sigma_u^2 < \frac{\Delta^2}{k} - \frac{1}{2\beta}$, $c'(\sigma_u^2) < 0$. Once again, because $c(\sigma_u^2)$ is left continuous at 0, continuous on $\mathbb{R}_+$, and $c(0) = c_0$, this means $c(\sigma_u^2) < c_0$ for σ_u^2 in this interval.

C.0.3 Proof of Proposition 3.2.3

Let $W = x(s^*) - \hat{x}_{\text{KALE}}(s^*)$. We can explicitly write the dependence of W on $\mathbf{u}_n$:

$$W|\mathbf{u}_n \sim \mathcal{N}(0, V(\mathbf{u}_n))$$

where

$$\begin{aligned} V(\mathbf{u}_n) &= \sigma^2 + \gamma' \mathbf{C}(\mathbf{s}_n + \mathbf{u}_n, \mathbf{s}_n + \mathbf{u}_n)\gamma - 2\gamma' \mathbf{C}(\mathbf{s}_n + \mathbf{u}_n, s^*), \\ \gamma &= \mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)^{-1} \mathbf{K}^*(\mathbf{s}_n, s^*), \end{aligned}$$

and $\sigma^2 = \mathbb{V}[x(s^*)]$. Thus

$$\begin{aligned} \mathbb{P}(W < z) &= \mathbb{E}[\mathbb{P}(W < z|\mathbf{u}_n)] \\ &= \mathbb{E}\left[\Phi\left(\frac{z}{\sqrt{V(\mathbf{u}_n)}}\right)\right]. \end{aligned}$$

C.0.4 Proof of Theorem 3.2.4

For any n, the KILE MSE in predicting $x(s^*)$ given $\mathbf{y}_n$ is

$$\begin{aligned} \mathbb{E}[(x(s^*) - \hat{x}_{\text{KILE}}(x^*))^2] &= v(s^*) - 2\mathbf{C}(s^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{K}^*(\mathbf{s}_n, s^*) \\ &\quad + \mathbf{C}(s^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{C}(\mathbf{s}_n, s^*). \end{aligned} \tag{C.1}$$

$\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)$ is symmetric and positive definite, so assume the eigendecomposition $\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n) = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ and similarly $\mathbf{K}(\mathbf{s}_n, \mathbf{s}_n) = \mathbf{R}\mathbf{\Omega}\mathbf{R}'$. Further letting $\mathbf{a} = \mathbf{C}(s^*, \mathbf{s}_n)\mathbf{Q}$, $\mathbf{b} = \mathbf{Q}'\mathbf{K}(\mathbf{s}_n, s^*)$, and $\mathbf{D} = \mathbf{Q}'\mathbf{R}\mathbf{\Omega}^{\frac{1}{2}}$, we can write (C.1) as

$$\begin{aligned} \mathbb{E}[(x(s^*) - \hat{x}_{\text{KILE}}(x^*))^2] &= v(s^*) - 2\mathbf{a}'\mathbf{\Lambda}^{-1}\mathbf{b} + \mathbf{a}'\mathbf{\Lambda}^{-1}\mathbf{D}\mathbf{D}'\mathbf{\Lambda}^{-1}\mathbf{a} \\ &= v(s^*) - 2\sum_{i=1}^{n} \frac{a_i b_i}{\lambda_i} + \sum_{j=1}^{n}\left(\sum_{i=1}^{n} \frac{a_i D_{ij}}{\lambda_i}\right)^2. \end{aligned} \tag{C.2}$$

Now let $\xi = \lambda_{(1)}^{-1}$ be the reciprocal of the smallest eigenvalue in $\mathbf{\Lambda}$. We can then write (C.2) as

$$\mathbb{E}[(x(s^*) - \hat{x}_{\text{KILE}}(x^*))^2] = \xi^2 \left(\sum_{j=1}^{n} a_{(1)}^2 D_{(1),j}^2 \right) + h(\xi), \tag{C.3}$$

where h is linear in ξ.

Without loss of generality, assume $s_1 \to s_2$. Thus $\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)$ becomes rank $n-1$, with $\lambda_{(1)} \to 0$ and for all $i > 1$, $\lambda_{(i)} \to \lambda_{(i)}^* > 0$. Thus $\xi \to \infty$.

For the vast majority of distributions $\mathbf{u}_n \sim g_{\mathbf{s}_n}(\mathbf{u}_n)$, we can use Dominated Convergence to show that c continuous on $\mathbb{S}^2$ implies k continuous. However, $\mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)$ does not become singular, since $\mathbb{P}(u_1 \neq u_2) < 1$ implies $\lim_{s_2 \to s_1} k(s_1, s_2) \neq \mathbb{V}[y(s_1)]$. Assuming continuity and the fact that $\mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)$ is nonsingular in the limit, all terms besides ξ in (C.2) converge; that is $\mathbf{a} \to \mathbf{a}^*$, $\mathbf{b} \to \mathbf{b}^*$, and $\mathbf{D} \to \mathbf{D}^*$ as $s_1 \to s_2$. Moreover, we cannot have $D_{(1),j}^* = 0$ for all j, as this contradicts $\mathbf{K}(\mathbf{s}_n, \mathbf{s}_n)$ remaining full-rank. Lastly, since $\mathbf{C}(s^*, \mathbf{s}_n) \neq \mathbf{0}$ and $\mathbf{Q}$ is orthogonal, $a_i \neq 0$ and $a_i^* \neq 0$ for all $i = 1, \ldots, n$.

Thus the quadratic coefficient in (C.3), $\sum_{j=1}^{n} a_{(1)}^2 D_{(1),j}^2$ is strictly positive, and $h(\xi) = \mathcal{O}(\xi)$. Because $\xi \to \infty$, we get

$$\lim_{s_1 \to s_2} \mathbb{E}[(x(s^*) - \hat{x}_{\text{KILE}}(x^*))^2] = \infty.$$

For pathological choices of $g_{\mathbf{s}_n}$ where k is not continuous everywhere and limits for $\mathbf{b}$ and $\mathbf{D}$ may not exist, all components of these terms can be still be bounded, which is sufficient for Theorem 3.2.4 to hold.

C.0.5 Proof of Proposition 3.3.1

Bayes rule predictors by definition satisfy $R_\pi(\pi) \leq R_\pi(\tilde{\pi})$, which confirms the two inequalities in the statement of Proposition 3.3.1. The equality $R_\pi(\pi_0) = R_{\pi_0}(\pi_0)$ holds since the risk of the

Bayes estimator under π_0 is a quadratic form, and therefore constant for all $\pi \in \Pi_{\mathbf{0},\mathbf{C}}$:

$$\begin{aligned} R_{\pi_0}(\pi_0) &= \mathbb{E}_{\pi}[(\mathbb{E}_{\pi_0}[x(s^*)|\mathbf{x}_n] - x(s^*))^2] \\ &= \mathbb{E}_{\pi}[(\mathbf{C}(s^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{x}_n - x(s^*))^2] \\ &= c(s^*, s^*) - \mathbf{C}(s^*, \mathbf{s}_n)\mathbf{C}(\mathbf{s}_n, \mathbf{s}_n)^{-1}\mathbf{C}(\mathbf{s}_n, s^*) \\ &= R_{\pi}(\pi_0). \end{aligned}$$

References

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4), 825–848.

Barber, J. J., Gelfand, A. E., & Silander, J. A. (2006). Modelling map positional error to infer true feature location. *Canadian Journal of Statistics*, 34(4), 659–676.

Berbeco, R. I., Nishioka, S., Shirato, H., Chen, G. T., & Jiang, S. B. (2005). Residual motion of lung tumours in gated radiotherapy with external respiratory surrogates. *Physics in Medicine and Biology*, 50(16), 3655–3667.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192–236.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Freudenheim, J. L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*, 14(4), 408–412.

Boots, B. & 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* (pp. 293–300). San Francisco, CA.

Boshnakov, G. N. (2009). Analytic expressions for predictive distributions in mixture autoregressive models. *Statistics and Probability Letters*, 79(15), 1704–1709.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F., & Jones, P. D. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 111(D12).

Bukiet, B., Harold, E. R., & Palacios, J. L. (1997). A markov chain approach to baseball. *Operations Research*, 45(1), 14–23.

Burke, B. (2010). Win probability added (wpa) explained. *www.advancedfootballanalytics.com*, (website).

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Carvalho, A. X. & Tanner, M. A. (2005). Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification. *IEEE Transactions on Neural Networks*, 16(1), 39–56.

Cervone, D., Pillai, N., Pati, D., Berbeco, R., & Lewis, J. H. (2014). Code supplement to "a location-mixture autogressive model for online forecasting of lung tumor motion". DOI:XYZ.

Cowtan, K. & Way, R. G. (2014). Coverage bias in the hadcrut4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, 140(683), 1935–1944.

Cox, D. R. (1975a). A note on partially bayes inference and the linear model. *Biometrika*, 62(3), 651–654.

Cox, D. R. (1975b). Partial likelihood. *Biometrika*, 62(2), 269–276.

Cressie, N. & Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 209–226.

Cressie, N. & Kornak, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical science*, 18(4), 436–456.

Cressie, N. A. & Cassie, N. A. (1993). *Statistics for spatial data*, volume 900. Wiley New York.

De Gooijer, J. G. & Kumar, K. (1992). Some recent developments in non-linear time series modelling, testing, and forecasting. *International Journal of Forecasting*, 8(2), 135–156.

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

Director, H. & Bornn, L. (2015). Connecting point-level and gridded moments in the analysis of climate data. *Journal of Climate*.

D'Souza, W., Naqvi, S. A., & Cedric, X. Y. (2005). Real-time intra-fraction-motion tracking using the treatment couch: a feasibility study. *Physics in Medicine and Biology*, 50(17), 4021–4033.

Ernst, F., Dürichen, R., Schlaefer, A., & Schweikard, A. (2013). Evaluating and comparing algorithms for respiratory motion prediction. *Physics in Medicine and Biology*, 58(11), 3911–3929.

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

Ernst, F. & Schweikard, A. (2009). Forecasting respiratory motion with accurate online support vector regression (SVRpred). *International Journal of Computer Assisted Radiology and Surgery*, 4(5), 439–447.

Fanshawe, T. & Diggle, P. (2011). Spatial prediction in the presence of positional error. *Environmetrics*, 22(2), 109–122.

Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2014). Characterizing the spatial structure of defensive skill in professional basketball. *arXiv preprint*, arXiv:1405.0231.

Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164–181.

Gabrosek, J. & Cressie, N. (2002). The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, 34(3), 262–285.

Gelman, A. & Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of Markov chain Monte Carlo*, (pp. 163–174).

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 243–268.

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

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. & 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. & 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* (pp. 289–297). Scottsdale, AZ.

Goldner, K. (2012). A Markov model of football: Using stochastic processes to model a football drive. *Journal of Quantitative Analysis in Sports [online]*, 8(1).

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69(1), 95–105.

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.

Heyde, C. C. (1997). *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. Springer Science & Business Media.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (pp. 37–56). New York, NY: Springer.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Hollinger, J. (2005). *Pro Basketball Forecast, 2005-06*. Washington, D.C: Potomac Books.

Homan, M. D. & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, 15(1), 1593–1623.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366.

Ihler, A., Hutchins, J., & Smyth, P. (2006). Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 207–216). New York, NY: ACM.

Jones, P., Lister, D., Osborn, T., Harpham, C., Salmon, M., & Morice, C. (2012). Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117(D5).

Kalbfleisch, J. D. & 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., & Schmitz, R. (2010). A state-based probabilistic model for tumor respiratory motion prediction. *Physics in Medicine and Biology*, 55(24), 7615–7631.

Krauss, A., Nill, S., & Oelfke, U. (2011). The comparative performance of four respiratory motion predictors for real-time tumour tracking. *Physics in Medicine and Biology*, 56(16), 5303–5317.

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

Laird, N. & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374), 231–240.

Lan, S., Streets, J., & Shahbaba, B. (2013). Wormhole hamiltonian monte carlo. *arXiv preprint arXiv:1306.0063*.

Lau, J. W. & So, M. K. (2008). Bayesian mixture of autoregressive models. *Computational Statistics and Data Analysis*, 53(1), 38–60.

Lawson, A. B. (1994). Using spatial gaussian priors to model heterogeneity in environmental epidemiology. *The Statistician*, (pp. 69–76).

Le, N. D., Martin, R. D., & Raftery, A. E. (1996). Modeling flat stretches, bursts, outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436), 1504–1515.

Lee, T. C., Judge, G., & Zellner, A. (1968). Maximum likelihood and bayesian estimation of transition probabilities. *Journal of the American Statistical Association*, 63(324), 1162–1179.

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

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Methodological)*, 73(4), 423–498.

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

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., & Ni, F. (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences*, 105(36), 13252–13257.

Mardia, K. V. & Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate environmental statistics*, 6(347-385), 76.

Matheron, G. (1962). *Traité de géostatistique appliquée*. Editions Technip.

McCulloch, R. E. & Tsay, R. S. (1994). Statistical analysis of economic time series via markov switching models. *Journal of Time Series Analysis*, 15(5), 523–539.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, (pp. 538–558).

Meshkani, M. R. & Billard, L. (1992). Empirical bayes estimators for a finite markov chain. *Biometrika*, 79(1), 185–193.

Miller, A., Bornn, L., Adams, R., & Goldsberry, K. (2013). Factorized point process intensities: A spatial analysis of professional basketball. In *Proceedings of the 31st International Conference on Machine Learning* (pp. 235–243).

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadcrut4 data set. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117(D8).

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, (pp. 515–529).

Morris, C. N. (1995). Hierarchical models for educational data: An overview. *Journal of Educational and Behavioral Statistics*, (pp. 190–200).

Murphy, M. J. & Dieterich, S. (2006). Comparative performance of linear and nonlinear neural networks to predict irregular breathing. *Physics in Medicine and Biology*, 51(22), 5903–5914.

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

Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint*, arXiv:physics/9701026.

Neal, R. M. (2005). Hamiltonian importance sampling. Talk presented at the Banff International Research Station (BIRS) workshop on Mathematical Issues in Molecular Dynamics, June 2005.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2.

Omidiran, D. (2011). A new look at adjusted plus/minus for basketball analysis. *MIT Sloan Sports Analytics Conference [online]*, 2011.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V., & Breslow, N. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, (pp. 541–554).

Qin, Z. S. & Liu, J. S. (2001). Multipoint metropolis method with application to hybrid monte carlo. *Journal of Computational Physics*, 172(2), 827–840.

Quiñonero-Candela, J. & Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6, 1939–1959.

Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Riaz, N., Shanker, P., Wiersma, R., Gudmundsson, O., Mao, W., Widrow, B., & Xing, L. (2009). Predicting respiratory tumor motion with multi-dimensional adaptive filters and support vector regression. *Physics in Medicine and Biology*, 54(19), 5735–5748.

Richard, A. et al. (2012). A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*.

Roberts, G. O., Rosenthal, J. S., et al. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4), 351–367.

Rottmann, J., Keall, P., & Berbeco, R. (2013). Markerless epid image guided dynamic multi-leaf collimator tracking for lung tumors. *Physics in Medicine and Biology*, 58(12), 4195–4204.

Ruan, D. & Keall, P. (2010). Online prediction of respiratory motion: multidimensional processing with low-dimensional feature learning. *Physics in Medicine and Biology*, 55(11), 3011–3025.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 71(2), 319–392.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, (pp. 409–423).

Salazar, R. & Toral, R. (1997). Simulated annealing using hybrid monte carlo. *Journal of Statistical physics*, 89(5-6), 1047–1060.

Schweikard, A., Glosser, G., Bodduluri, M., Murphy, M. J., & Adler, J. R. (2000). Robotic motion compensation for respiratory movement during radiosurgery. *Computer Aided Surgery*, 5(4), 263–277.

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

Shao, X. & Li, L. (2011). Data-driven multi-touch attribution models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 258–264). New York, NY: ACM.

Sharp, G. C., Jiang, S. B., Shimizu, S., & Shirato, H. (2004). Prediction of respiratory tumour motion for real-time image-guided radiotherapy. *Physics in Medicine and Biology*, 49(3), 425–440.

Smith, R. L. (2004). Asymptotic theory for kriging with estimated parameters and its application to network design.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Stan Development Team (2014). Rstan: the r interface to stan, version 2.5.0.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stein, M. L., Chi, Z., & Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2), 275–296.

Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time-series motif from multi-dimensional data based on mdl principle. *Machine Learning*, 58(2-3), 269–300.

Thomas, A., Ventura, S. L., Jensen, S. T., & Ma, S. (2013). Competing process hazard function models for player ratings in ice hockey. *The Annals of Applied Statistics*, 7(3), 1497–1524.

Tibshirani, R. (1996). A comparison of some error estimates for neural network models. *Neural Computation*, 8(1), 152–163.

Tingley, M. P. (2012). A bayesian anova scheme for calculating climate anomalies, with applications to the instrumental temperature record. *Journal of Climate*, 25(2), 777–791.

Tingley, M. P. & Huybers, P. (2010). A bayesian algorithm for reconstructing climate anomalies in space and time. part i: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 23(10), 2759–2781.

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

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

Tong, H. & 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. & Moeanaddin, R. (1988). On multi-step non-linear least squares prediction. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37(2), 101–110.

Varin, C., Reid, N. M., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1), 5–42.

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

Veregin, H. (1999). Data quality parameters. *Geographical information systems*, 1, 177–189.

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., Mix, W., Colt, J. S., & Hartge, P. (2005). Positional accuracy of two methods of geocoding. *Epidemiology*, 16(4), 542–547.

Warnes, J. & Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial gaussian processes. *Biometrika*, 74(3), 640–642.

Wong, C. S. & Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1), 95–115.

Wong, W. H. (1986). Theory of partial likelihood. *The Annals of Statistics*, (pp. 88–123).

Yang, T. Y. & Swartz, T. (2004). A two-stage bayesian model for predicting winners in major league baseball. *Journal of Data Science*, 2(1), 61–73.

Ye, L. & 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* (pp. 947–956). Paris, France.

Zhu, Z. & Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1), 24–44.

Zimmerman, D. L. & Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the institute of statistical mathematics*, 44(1), 27–43.

Zimmerman, D. L., Li, J., & Fang, X. (2010). Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in medicine*, 29(9), 1025–1036.

Zimmerman, D. L. & Sun, P. (2006). Estimating spatial intensity and variation in risk from locations subject to geocoding errors. *Iowa City: University of Iowa*.