November 2015

Physical Activity Classification with Conditional Random Fields

Evan L. Ray
University of Massachusetts Amherst

PHYSICAL ACTIVITY CLASSIFICATION WITH CONDITIONAL RANDOM FIELDS

A Dissertation Presented

by

EVAN L. RAY

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2015

Department of Mathematics and Statistics

PHYSICAL ACTIVITY CLASSIFICATION WITH CONDITIONAL RANDOM FIELDS

A Dissertation Presented

by

EVAN L. RAY

Approved as to style and content by:

John Staudenmayer, Chair

Krista Gile, Member

Michael Lavine, Member

Patty Freedson, Member

Farshid Hajir, Department Head
Mathematics and Statistics

ACKNOWLEDGEMENTS

I would like to thank my committee members for their valuable insights, questions, comments, and guidance. I am also grateful to Jeffer Sasaki and Stephen Intille for their willingness to make their data available for use in this work and their helpfulness in answering my questions. Finally, I have been lucky to have the support of my friends and family.

ABSTRACT

PHYSICAL ACTIVITY CLASSIFICATION WITH CONDITIONAL RANDOM FIELDS

SEPTEMBER 2015

EVAN RAY, B.S., UNIVERSITY OF MASSACHUSETTS, BOSTON

M.S., UNIVERSITY OF MASSACHUSETTS, AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS, AMHERST

Directed by: Professor John Staudenmayer

In this thesis we develop methods for classifying physical activity using accelerometer recordings. We cast this as a problem of classification in time series with moderate to high dimensional observations at each time point. Specifically, we observe a vector of summary statistics of the accelerometer signal at each point in time, and we wish to use these observations to estimate the type and intensity of physical activity the individual engaged in as it changes over time.

Our methods are based on Conditional Random Fields, which allow us to capture temporal dependence in an individual's physical activity type without requiring us to model the distribution of the observed features at each point in time. We develop three novel estimation strategies for Conditional Random Fields, evaluate their performance on classification tasks through simulation studies and demonstrate their use in applications with real physical activity data sets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Scientists would like to be able to classify the type and intensity of physical activity that a person is performing in order to help answer questions about the relationship between physical activity and health and assess the effectiveness of interventions designed to increase physical activity levels. One approach to the objective measurement of physical activity that the research community favors is through the use of an accelerometer worn by the individual that records acceleration over time. Advantages of accelerometers include their relatively low costs, low maintenance requirements, and minimal burden on the subject wearing the device. However, current methods for inferring physical activity type and intensity from accelerometer recordings are imperfect. As a result, measurement error is introduced into statistical analyses relating physical activity type or intensity to health outcomes, potentially masking the relationship between these variables and causing bias in the estimated model parameters. Similarly, our ability to assess the effectiveness of interventions is limited.

In this thesis we explore methods for classifying physical activity according to its type or intensity level using accelerometer data. We classify according to activity type using categories like Sitting, Standing, Locomotion, and Cycling. The particular set of activity type categories used varies with the data set; we will describe these data in more detail later. We classify activity intensity as either Sedentary, Light, Moderate, or Vigorous activity. These categories are a commonly used binning of an underlying continuous measure of energy expenditure.

Our models do not use the accelerometer measurements directly. Instead, we divide the time during which acceleration was recorded into short windows a few seconds in length and calculate a vector of features for each window that summarizes the accelerometer signal within

that shorter time span. Our models relate this vector of features to the observed activity type or intensity in each window. Our overall strategy is to first estimate the model parameters using data from subjects whose activity type or intensity is known at each point in time, and then use the fitted models to classify activity in new data where only the accelerometer recordings are available.

Our proposed methods for activity type and intensity classification are variations on Conditional Random Fields (CRFs). The CRF specifies a conditional distribution for the sequence of activity types that an individual engages in given the observed accelerometer features in each window. CRFs were originally formulated in the computer science literature and have been applied to physical activity classification with accelerometer data in one previous study. They are useful because they provide a mechanism for capturing temporal dependence in the activity type without requiring us to specify a model for the accelerometer features, which have a complex distribution that is difficult to model well.

Our contributions are three new estimation strategies for CRFs. Our first estimation strategy is based on a parametric specification of the CRF model, and combines bagging of observation sequences with boosting to estimate the model parameters. Our second strategy uses a non-parametric specification for the component of the model that relates the accelerometer features to the activity type labels, and again employs bagging and boosting in the estimation procedure. Our third strategy is also based on this non-parametric model specification, and uses an estimation algorithm similar to that employed in random forests. In simulation studies and applications to real physical activity data, we demonstrate that the CRF model can lead to improved classification performance relative to models that do not account for sequential dependence in activity types or that attempt to model the distribution of the accelerometer features associated with each activity type. We also show that two of our three proposed methods consistently offer good classification performance, while the performance of other approaches to estimating CRFs and similar models is less consistent.

Throughout this work, we will use capital letters to denote random variables and lower case letters to denote realizations of those random variables. Bold letters indicate vectors. The training data for our models consist of N pairs $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T_i})$ is a sequence of length T_i containing the true class at each time window and $\mathbf{x}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T_i})$ is a sequence of length T_i containing the observed accelerometer features at each time window. N corresponds to

the number of subjects in a study, and T_i is the number of windows in the observation sequence for the ith subject. There are S classes, so that each $y_{i,t} \in \{1, \ldots, S\}$. The feature vectors have dimension D, so that $\mathbf{x}_{i,t} = (x_{i,t,1}, \ldots, x_{i,t,D}) \in \mathbb{R}^D$.

The rest of this paper is organized as follows. In Section 1.2 of this Chapter, we present a taxonomy of classification models and estimation strategies for those models that will help to organize our work and describe how our methods fit into the existing literature. In Chapter 2 we review the literature related to classification methods in general and CRFs in particular, as well as previous approaches to physical activity classification and energy expenditure estimation. We then describe the physical activity data that are available to use in training our models in Chapter 3. We conduct two small preliminary simulation studies to motivate the use of CRFs in Chapter 4. We then discuss our proposed estimation strategies for CRF models in Chapter 5. We conduct a simulation study comparing our methods to several existing approaches in Chapter 6, and then apply our methods to classification of physical activity type in Chapter 7 and to classification of physical activity intensity in Chapter 8. Finally, we summarize our contributions, discuss the limitations of our methods, and offer suggestions for future work in Chapter 9.

1.2 A Taxonomy of Classification Models and Estimation Strategies

In this Section, we present a taxonomy for classification models and estimation strategies that have been applied with physical activity and similar data. This taxonomy is a set of 7 characteristics of a model and estimation strategy that capture the differences between many of the methods that have been developed to perform classification in structured data sets. Each of these 7 characteristics can take one of several values. Thus, any particular model and estimation algorithm can be viewed as falling into a cell of a factorial design.

We divide this taxonomy into two sections: one describing the model specification, and one describing the estimation algorithm. The taxonomy for the model specification is as follows:

1. **Static vs. Dynamic**:
    - **Static**: The model does not account for sequential dependence in the data.
    - **Dynamic**: The model does account for sequential dependence in the data.
2. **Joint vs. Conditional (aka Generative vs. Discriminative)**:

- **Joint**: The model specifies the joint distribution for $(\mathbf{Y}, \mathbf{X})$. In the classification setting, these are often referred to as generative models since they specify how the data are generated.
- **Conditional**: The model specifies the conditional distribution for $\mathbf{Y}|\mathbf{X}$. In the classification setting, these are often referred to as discriminative models.

3. **Parameterization of the Observation Distributions or Feature Functions**:

   - **Restrictive**: In generative models, the distributions of the features in each class are often modeled by a given parametric family, or a mixture of distributions in such a family with a small number of mixture components (such as a mixture of Gaussians with only a few mixture components). In the case of discriminative models, feature functions take the place of the observation distributions. Intuitively, these feature functions represent how likely an observation is to fall in a particular class given the observed features $\mathbf{X}$, although their exact interpretation depends on the details of the model formulation. If the conditional model is derived by conditioning on $\mathbf{X}$ in a joint model for $(\mathbf{Y}, \mathbf{X})$ where the observation distributions are in an exponential family, the feature functions often correspond to the sufficient statistics for the exponential family and the estimation task reduces to estimating the natural parameters of the family. However, it is also possible to formulate more general parametric specifications.
   - **Flexible**: A non-parametric form can also be used for either the observation distributions or the feature functions. For instance, kernel density estimation could be used to estimate the observation distributions in generative models. In discriminative models, it is common to use decision or regression trees for the feature functions.

4. **Number of Components,** M:

   - $M > 1$: The model combines inferences from multiple "component models". These are often referred to as ensemble methods. The number of component models, M, may be specified manually or selected using a method such as cross-validation. Classification is performed by combining the inferences from these distinct models; the following are three methods for combining the component models among many other possibilities:

  - Logarithmic opinion pools (LOPs)
  - Linear opinion pools (LIPs)
  - Majority vote
- $M = 1$: The model consists of a single component model.

We now present our taxonomy for estimation algorithms. In this taxonomy, we focus on approaches that have been developed for ensemble methods, when $M > 1$.

1. **Independent vs. Dependent Estimation of Component Models**:
   - **Sequential**: The component models are estimated sequentially. In each stage, the parameter values for earlier models are held fixed and the estimates for later models depend on these earlier estimates. These methods are often referred to as boosting algorithms, although some authors reserve boosting to refer to a subset of these methods.
   - **Backfitting**: Backfitting is similar to the sequential estimation strategy described above, but we iteratively update the parameters for each component model until some criteria are met.
   - **Independent Estimation**: The parameters of the component models are estimated separately.
2. **Feature Subset Selection**:
   - **Manually Selected Subsets Per Component Model**: The full set of features is manually divided into subsets. A separate component model is fit using each subset of features.
   - **Randomly Selected Subsets Per Component Model**: The full set of features is randomly divided into subsets. A separate component model is fit using each subset of features.
   - **Randomly Selected Subsets Within Each Component Model**: Some estimation strategies use randomly selected feature subsets within the estimation of each component model. An example of this is random forests, where a possibly different subset of the features is randomly selected to use in finding a split point in each tree node.

- **All Features Used**: All features are used at every stage in the estimation procedure.

3. **Bagging**:

    - **Bagging Observation Sequences**: Each component model is trained using a randomly selected subset of the observation sequences. Thus, for each observation sequence all of the observations are either "in the bag", meaning that they are used in estimating the parameters for a given component model, or "out of the bag", meaning that they are not used in training that component.
    - **Bagging Time Point Observations**: Each component model is trained using a randomly selected subset of the observations. In each observation sequence, some of the observations may be "in the bag" and others may be "out of the bag".
    - **No Bagging**: All observations are used in training each component model.

We note that this taxonomy is not exhaustive and does not seek to uniquely describe every classification method that has been developed; rather, it provides a useful structure for thinking about existing classification methods and identifying gaps in the literature that may be fruitfully filled. In Chapter 2, we review existing methods in the literature and describe how they fit into this taxonomy. The methods we develop in Chapter 5 represent three promising cells in the factorial design described by this taxonomy that have not yet been explored.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this Chapter, we review the literature on classification in the presence of sequential dependence and methods that have been proposed for physical activity classification in particular. We begin with a general discussion of the differences between the generative and discriminative approaches to classification in Section 2.2. We then discuss Hidden Markov Models (HMMs) in Section 2.3. The HMM is a generative model that is commonly used to perform classification in the presence of sequential dependence. We then review two discriminative methods for classification with sequential dependence in Section 2.4. We discuss some ideas for how estimation of these models can be performed effectively when we observe a moderate to large number of covariates at each time point in Section 2.5. In Section 2.6 we discuss ensemble methods, which combine inferences from many slightly different model fits. Finally, in Section 2.7 we review previous models that have been used for activity classification and energy expenditure estimation.

2.2 Generative (Joint) vs. Discriminative (Conditional) Modeling

In this Section, we discuss the differences between the generative and discriminative approaches to modeling, and the circumstances when one or the other of these approaches may be preferred. We also note that methods combining the generative and discriminative approaches have been suggested [e.g., Rubinstein, 1998, Raina et al., 2004]. We will not discuss these ideas further here.

The generative approach proceeds by specifying a joint model for the class $\mathbf{Y}$ and covariates $\mathbf{X}$ at each time point. These are referred to generative models because they model the

data generating process. The model parameters $\boldsymbol{\theta}$ can be estimated in either the frequentist or Bayesian paradigm. Once the parameters have been estimated, we can perform classification in a new data set where only $\mathbf{X}$ is observed by using Bayes' rule to compute the probabilities $P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}; \widehat{\boldsymbol{\theta}})$. We will discuss one common dynamic joint model for $(\mathbf{Y}, \mathbf{X})$, the Hidden Markov Model (HMM) in Section 2.3.

In contrast, the discriminative approach seeks to directly optimize the performance of the final classifier, which discriminates between different classes based on the values of the observed covariates $\mathbf{x}$. Often, but not always, this is done by maximizing the likelihood of a conditional model for $\mathbf{Y}|\mathbf{X}$. We will cover some approaches that have been developed for discriminative modeling in the presence of sequential dependence in Section 2.4.

The classic discussion of the relative merits of generative and discriminative methods centers on a comparison of normal discriminant analysis (NDA) and logistic regression (LR). NDA is a generative model for i.i.d. data specifying that the observed covariates $\mathbf{X}$ follow a mixture of normals, with a different normal distribution associated with each class Y. Conditioning on $\mathbf{X}$ in this generative model yields LR, a discriminative model for the distribution of $\mathbf{Y}|\mathbf{X}$, the probabilities of each class. Although both NDA and LR can be specified in terms of the same parameters, the resulting maximum likelihood parameter estimates are different [Efron, 1975].

Efron [1975] examines the asymptotic behavior of these models in the case where the generative NDA model is true. He shows that with both models the error rate of the classifier obtained using the MLEs converges to the Bayes error rate as the training sample size goes to infinity. However, the asymptotic variance of the classification error rate from NDA is smaller than the variance of the error rate from LR. Thus, the generative approach is preferred when the joint model for $(\mathbf{Y}, \mathbf{X})$ is accurate.

When the data generating mechanism specified by the joint model for $(\mathbf{Y}, \mathbf{X})$ is not accurate, it is less clear which method is preferred. One point in favor of LR is that it can be obtained by conditioning on $\mathbf{X}$ in a joint model similar to NDA as long as the distribution of $\mathbf{X}$ associated with each class is in an exponential family. For example, if the joint model specifies that the data follow a mixture of gamma distributions with one mixture component for each class, we again obtain the LR model by conditioning on $\mathbf{X}$ in the joint model. Thus, LR is an accurate model in a wider range of circumstances than NDA. Also, it can be shown using the theory of Vapnik-Chervonenkis dimensions that for any generative/discriminative model pair where the

discriminative model is obtained by conditioning on $\mathbf{X}$ in the generative model, the asymptotic error rate of the discriminative model is no worse than the asymptotic error rate of the generative model [Ng and Jordan, 2002]. However, Ng and Jordan [2002] show that under certain conditions NDA approaches its asymptotic error more quickly than LR. They conclude that in many problems, NDA is preferred when the sample size is relatively small, and LR is a better approach when the sample size is large.

Similar asymptotic results have been obtained for estimation of parameters in the Hidden Markov Model (HMM), which specifies a joint model for $(\mathbf{Y}, \mathbf{X})$ and captures sequential dependence. Nádas [1983] compares the MLEs of the HMM parameters with the MLEs in the model for $\mathbf{Y}|\mathbf{X}$ obtained by conditioning on $\mathbf{X}$ in the HMM. He shows that if the HMM is an accurate model of the data generating process, the MLE from the joint model is asymptotically relatively efficient in comparison to the MLE from the conditional model. In a later expansion on this work, Nádas et al. [1988] give an example where the Markov structure in the HMM is misspecified and demonstrate that the MLE from the conditional model yields a classifier with an asymptotically lower misclassification rate than the MLE from the joint model.

We can summarize this discussion as follows. In general, if the data generating model is correctly specified, generative approaches are superior. However, if the joint model is misspecified, discriminative methods are often preferred since they make fewer assumptions. Xue and Titterington [2008] conclude that

> For real world datasets... there is no theoretically correct, general criterion for choosing between the discriminative and the generative classifiers; the choice depends on the relative confidence we have in the correctness of the specification of either $p(y|\mathbf{x})$ or $p(y, \mathbf{x})$.

In many applications with sequential dependence the HMM is an inadequate representation of the data generating process, suggesting that discriminative methods may be preferred. We will discuss this issue in the context of physical activity classification in Chapters 4 and 5.

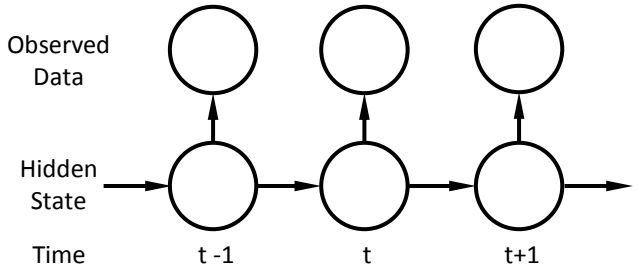


Figure 1. A diagram illustrating the HMM. Dependence is shown with arrows. The distribution of the state at a given time depends on the state at the previous time. The distribution of the observed data at a given time depends on the state at that time.

2.3 Hidden Markov Models

In this Section we review the literature about Hidden Markov Models (HMMs). In Subsection 2.3.1 we introduce HMMs and a related model, the Finite Mixture Model (FMM). We review frequentist inference for HMMs in Subsection 2.3.2 and Bayesian inference in Subsection 2.3.3.

2.3.1 HMM Basics

HMMs have been successfully applied in a wide variety of fields including speech recognition [Rabiner, 1989], financial time series [Rydén et al., 1998], detection of fraudulent telephone activity [Scott, 1999], and information extraction from text [Taghva et al., 2005]. The model can be used when observations are made at regular intervals in space or time. We will refer to observations made in time since that is our context, but the ideas apply to both settings.

The model is based on the assumption that each observation $\mathbf{X}_{i,t}$ follows a specified parametric distribution with parameters depending on an unobserved (or hidden) state $Y_{i,t}$ corresponding to that point in time. In the most basic formulation, the state changes over time according to a Markov process; that is, given the states that occurred up through time t, the probability distribution for the state at time $t+1$ depends only on the state at time t (and not on the earlier states). The states come from a state space which can be discrete or continuous, and finite or infinite. We focus on the case of a discrete state space of size $S < \infty$. In that case, the transition probabilities for the hidden Markov chain are often organized into an $S \times S$ transition matrix Q, where entry (r, s) is the probability that the state at time $t+1$ is s given the state at time t is r. A schematic depiction of this model is given in Figure 1.

Finite mixture models are a related class of model that are also based on the assumption that the parameters of the observed data distribution are associated with an unobserved state. However, in FMMs the hidden states are assumed to be i.i.d. with a fixed multinomial distribution for all observations. It can be seen FMMs are a special case of HMMs where the rows of the transition matrix are all equal to each other. To see this, note that if we restrict the rows of the transition matrix of a HMM to be the same, the probability distribution for the state at time $t+1$ is the same no matter what the state is at time t. This restriction therefore removes the dependence of future states on the current state, so that the model is equivalent to a FMM. Many of the same issues arise when fitting HMMs and FMMs, so we will draw from the literature about both of these models in this review.

The unknown parameters of the HMM are the distribution of the initial state, the transition probabilities for moving from one state to another, and the parameters for the observed data distribution associated with each state in the state space. Depending on the formulation, the number of states in the state space may also be treated as a model parameter. In different applications, we may be interested in learning about these model parameters, the unobserved state at each time point, or both.

Any parametric distribution can be used to model the distribution of the observed data given the state. One common choice is a multivariate Gaussian distribution, which has been used in a Bayesian setting by Spezia [2009, 2010] for HMMs, and by Lavine and West [1992], Stephens [2000a], Zhang et al. [2004], Dellaportas and Papageorgiou [2006], and Komárek [2009], among many others, for FMMs. In an applied univariate problem, Stephens [2000a] noted that when a Gaussian distribution was used, extra states may have been included to account for fat tails in the observed data. To correct for this, he used t distributions. Jasra et al. [2007] have also used a t distribution in a multivariate setting. Another common option in HMMs, especially in the speech recognition literature, is to use a multivariate normal FMM for the observation distribution in each state [e.g., Hain et al., 2003, Dimitrakakis and Bengio, 2004].

An important aspect of our data, which we present in more detail in Chapter 3, is that we observe a classification at each time point in addition to the feature vector. In other words, we have labeled data and training the HMM is a supervised learning problem.

There are several common approaches to using HMMs with labeled data. One simple option is to assume that there is a one-to-one correspondence between the observed classes and states

in the HMM – that is, the data used to estimate the model parameters contain observations of the hidden states. An advantage to this approach is that parameter estimation is comparatively simple. Another alternative is to allow for a group of states within the HMM for each observed class. This idea has been used in models of physical activity [Pober et al., 2006] as well as speech recognition [Dimitrakakis and Bengio, 2004] and video segmentation [Xie et al., 2004].

2.3.2 Frequentist Inference for HMMs

If we identify the state with the observed classes $\mathbf{y}_{i,t}$ in the training data, we can easily obtain maximum likelihood estimates of the initial state distribution and the transition matrix using the observed initial state and number of transitions between each pair of states. Also, the observed features are conditionally independent given the states, so we can treat the observed feature vectors from a given state as an i.i.d. sample from the distribution associated with that state [Frühwirth-Schnatter, 2006]. If these state-specific distributions are modeled by a simple parametric form, maximum likelihood estimates for the state-specific distributions are often obtained easily. On the other hand, if the observation distribution associated with each state is modeled by a Gaussian FMM, or if we associate multiple states in the HMM with each observed class, estimation is more involved.

The most common approach to estimating the model parameters when the state memberships are not observed is to obtain local maximum likelihood parameter estimates for a fixed state space size with an EM algorithm. For the special case of a discrete observation vector at each time point, this algorithm is known as the Baum-Welch algorithm in the HMM literature; it can be easily extended to handle finite mixtures of continuous distributions (with some restrictions on the form of the distribution for each finite mixture component) [Baum et al., 1970, Rabiner, 1989]. The basic idea of this algorithm is to iteratively re-estimate the model parameters according to the following procedure:

(1) Estimation step: Based on the current model parameters and the observed data, calculate

   (a) the probability of transitions from each state to each other state for every time t, and

   (b) the probability that the observation at time t came from the kth mixture component of the distribution associated with state s, for each time t, state s, and mixture component k.

(2) Maximization step: Using the quantities calculated in step (1), update the model parameter estimates:

   (a) the initial state distribution is updated using the probabilities of each state at time 1,

   (b) the transition probabilities are updated using the total expected number of transitions from each state to each other state at all times, and

   (c) the parameters of the distribution associated with each mixture component are updated using the observed data at each time t and the probabilities that the observation at time t came from each mixture component of the distribution associated with each state.

The calculations in step (1) can be implemented efficiently using a recursion algorithm. The Baum-Welch algorithm always converges to a local maximum of the likelihood function, but not necessarily to a global maximum [Rabiner, 1989].

In addition to the fully parameterized covariance matrix, reduced parameterizations of the covariance matrices for the Gaussian components have been considered. These are common in high-dimensional settings, where there may not be enough data to estimate all $D(D-1)/2$ parameters in each covariance matrix. There are three main ideas here: (1) impose restrictions on the spectral decomposition of the covariance matrix [Banfield and Raftery, 1993, Celeux and Govaert, 1995, Gales, 1999, 2002, Bouveyron et al., 2007], (2) represent the covariance as a low-rank component capturing most of the variation in the covariates plus a component with small variance [Tipping and Bishop, 1999, McLachlan et al., 2003, Attias, 1999, Saul and Rahim, 2000], or (3) model the inverse of the covariance, which is often sparse [Dharanipragada and Visweswariah, 2006, Olsen and Gopinath, 2004, Vanhoucke and Sankar, 2004, Axelrod et al., 2002, Bilmes, 2000]. EM or forward search algorithms are typically used for inference with these parameterizations in the frequentist setting. We will discuss other approaches for handling high dimensions in Section 2.5.

In order to determine the size of the state space, we can perform maximum likelihood estimation separately for several candidate sizes and then compare these models with likelihood ratio tests or information criteria such as AIC and BIC [Rydén et al., 1998]. Another option is to maximize a penalized likelihood function [Leroux and Puterman, 1992].

Once the model parameters have been estimated, there are several possible methods to perform classification in a sequence where the states are unobserved. One option is to calculate the

probabilities of each state at each time point given the observed data. We obtain a separate vector of class probabilities for each point in time, averaging over the possible values of the states at all other times. We may either report these estimated class probabilities, or choose the single class at each time point with the highest estimated probability. This approach maximizes the expected number of correctly estimated states, and is therefore optimal under misclassification loss [Rabiner, 1989]. However, the sequence of states obtained with this method may have a very low probability of occurring. Another option is to find the most likely sequence of states given the observed data and the parameter values for the model. This can be done using the Viterbi algorithm [Forney Jr, 1973]. The choice of which classification procedure to use depends on the goals of the researcher.

2.3.3 Bayesian Inference for HMMs

In this Subsection we discuss inference for HMMs in a Bayesian setting. We begin by discussing common choices for prior distributions, and then we review algorithms that have been developed to sample from the posterior distribution of the model parameters and the distribution of the sequence of hidden states. Throughout this Subsection, we focus on models in which a Gaussian or a finite mixture of Gaussians is used for the observation distribution for each state.

Prior Distributions

An important consideration in specifying prior distributions for HMMs is that the direct use of improper priors for the model parameters can be problematic. This is because it is possible that one or more of the states could be unrepresented by any of the observations (unless our model formulation assumes one HMM state for each observed class and the observation distribution assumes only one Gaussian mixture component for each state). There is an additive term in the likelihood corresponding to that possibility. When we form the posterior, the observed data do not contribute any information to our knowledge of the parameters for that state, so the integral of that term is infinity if the prior is improper. However, improper priors can be used if done carefully, for instance in a multilevel prior specification or if some dependency is induced between the parameters for different states [Marin et al., 2005].

There are two common choices for the prior distribution on the size of the state space: a discrete uniform distribution on the integers from 1 to a chosen maximum, and a truncated

Poisson distribution. In either case, the cutoff for the maximum size must be large enough that the true value lies within the support of the prior distribution. Scott [2002] points out that this is not an onerous condition since we could allow for the possibility that every observation is drawn from a different distribution by choosing the cutoff to be equal to the number of observations. In practice, it usually suffices to choose a much smaller number than this.

It is common to take independent Dirichlet priors for each row of the transition matrix, since the row entries must sum to 1. Most authors take the parameter for the prior to be a vector of 1's, which results in a uniform distribution on the space of valid transition matrices (i.e., where all components are non-negative and each row sums to 1).

Cappé et al. [2003] introduce an alternative parameterization in terms of an $S \times S$ matrix Ω, where each entry of Ω is non-negative but may be greater than 1. The transition matrix is obtained from Ω by scaling so that each row sums to 1. They argue that removing the constraint that each row sums to 1 improves mixing of the MCMC sampler. It also facilitates the development of some of the transition moves in some of the MCMC samplers discussed in the next Subsection. Cappé et al. [2003] use i.i.d. gamma priors for each entry in the Ω matrix, noting that this results in a Dirichlet prior on each row of the original transition matrix.

Spezia [2010] uses the same parameterization of the transition matrix as Cappé et al. [2003], but with a more informative prior. This prior places more weight on large values for the diagonal terms and small values for the off-diagonal terms. This captures the phenomenon that the state is more likely to remain the same from one time point to the next than to change.

Most authors specify data-driven priors for the mean vectors of the observed data given the hidden states, following ideas used by Richardson and Green [1997] in the univariate case. The basic idea is to take the prior mean equal to a measure of the center of the observed data, and the prior variance and covariance to be functions of the range of observed values for each covariate. Dellaportas and Papageorgiou [2006] work with centered and scaled data and note that as a result the prior mean can reasonably be set to 0 in each component.

Spezia [2009] claims that it is possible to achieve a non-informative prior distribution with this data-driven approach by taking the prior variance to be large relative to the spread of the observed data. However, Stephens [2000b] observed that estimation of the number of mixture components in a FMM was sensitive to the prior specification for all parameters. Hierarchical priors are often used to give less informative prior distributions for the mean and covariance

parameters [e.g., Robert et al., 2000, Cappé et al., 2003, Richardson and Green, 1997].

The simplest prior for the covariance matrices of the observed data given the state is Inverse-Wishart (or a Wishart prior on the precision matrix). This specification was used by Spezia [2010], Stephens [2000a], Dellaportas and Papageorgiou [2006], Jasra et al. [2007], and Komárek [2009]. However, many other ideas have been suggested, especially when the number of variables observed at each time point is large. As in the frequentist case, many authors have worked with reduced parameterizations based on the spectral decomposition of the covariance matrices [Bensmail et al., 1997, Bensmail and Meulman, 2003, Erar, 2011, Zhang et al., 2004].

Algorithms to Sample from the Posterior Distribution

We now discuss algorithms that can be used to sample from the posterior distribution of the parameters. We focus on Markov Chain Monte Carlo (MCMC) algorithms. Several other classes of algorithms have been proposed that can be used to approximate the posterior distribution of the parameters in HMMs and related models, including particle filtering or Sequential Monte Carlo [Liu and West, 2001], particle Markov Chain Monte Carlo [Andrieu et al., 2010], and Variational Bayes [Ji et al., 2006]. We focus on MCMC methods because they have been shown to be effective in fitting HMMs.

Some considerations in designing MCMC algorithms to use in fitting HMMs and FMMs include how to handle the unobserved states, ways to improve mixing and escape from local modes in the posterior distribution, and the method used to explore different state space sizes. We will review ideas that have been proposed for each of these problems.

One idea for dealing with the hidden states, referred to in the literature as *completion*, is to treat them as missing data and sample their values as part of the MCMC sampler. This is useful because the conditional posterior distributions of the model parameters given the states generally have simple forms. We can therefore use a Gibbs sampler with some Metropolis-with-Gibbs steps if non-conjugate priors are used. This method was originally developed by Diebolt and Robert [1994]; Scott [2002] is a review paper focusing on this technique.

An alternative to completion is to use a more general Metropolis-Hastings algorithm. In this approach, we evaluate the full likelihood by integrating over the distribution of the state sequence, rather than sampling one sequence from that distribution at each iteration and then drawing from the conditional distribution of the parameters given the states. In order to achieve

reasonable acceptance rates, the high dimensional parameter vector is typically broken into subcomponents with separate proposals made for each. This approach is dependent on the availability of an algorithm for calculating the likelihood quickly, since the likelihood must be calculated for each proposal. A recursion algorithm to do this has been developed in the literature [Scott, 2002, Zucchini and MacDonald, 2009].

Scott [2002] gives several arguments for why using completion may be a better approach. One consideration is that even using the likelihood recursion, computing the acceptance ratio for a Metropolis-Hastings move is more computationally expensive than sampling from the full conditional posterior in the Gibbs sampler. Another advantage to using completion is that some of the parameters may be highly correlated in the posterior distribution, but have low correlation conditional on the state sequence. If that is the case, mixing may be improved if we sample the state sequence as part of the algorithm. Finally, we are often interested in the unobserved states, so we will have to sample their values at some point anyways.

However, several authors have argued that the more general Metropolis-Hastings algorithm has faster convergence or better mixing than algorithms using completion and Gibbs steps. One line of reasoning given by Cappé et al. [2003] is that the sampler will converge faster because the dimension of the space we must sample from is reduced. Integrating over the distribution of the state sequence eliminates the stochasticity that would be introduced by sampling from it. Also, Celeux et al. [2000] and Jasra et al. [2005] have observed that the Gibbs sampler using completion is more likely to get stuck in a local mode of the posterior distribution because the sampled states reinforce the current parameter values.

The problem of escaping local modes is a serious consideration for MCMC samplers used for HMMs and FMMs since the posterior distribution is typically highly multimodal. One way of escaping local modes in a sampler that doesn't use completion is tempering. The idea behind tempering is to generate proposals by making use of a sequence of distributions which are successively more "flattened" versions of the true posterior we would like to sample from. One way of obtaining these flattened distributions is to let them be proportional to a power of the true posterior, where the powers are at a range of values between 0 and 1. Given the current parameter value for the chain, a proposal for the next is generated by moving up through these distributions from the true posterior through the flattened versions, and then back down. The proposed parameter can make larger moves through this technique and so escape local modes,

but will still be in relatively high-probability regions of the posterior. This method was first suggested by Neal [1996] and was used in the context of FMMs by Celeux et al. [2000]. A similar idea known as population Monte Carlo maintains a separate MCMC chain for each of the flattened distributions, and allows transition moves to exchange states between the different chains. This was used for FMMs by Jasra et al. [2007]. Finally, Scott [2002] has observed that methods for changing the size of the state space (discussed next) can help to resolve this problem by providing a mechanism for adding new states with parameters that are not close to the parameter values for existing states.

Another major consideration for designing a MCMC sampler is how different state space sizes are explored. The simplest idea is to run a separate sampler for each possible state space size, keeping the size of the state space fixed within a single sampler. The posterior probability of a given size S is then approximately proportional to the prior probability of S times the average likelihood of the parameter values sampled in the corresponding MCMC chain. The proportionality is resolved through the constraint that the posterior probabilities must sum to 1 [Scott, 2002]. It is also possible to use criteria such as Bayes Factors or information complexity measures to compare these separately-estimated models and select the number of states [Bensmail and Meulman, 2003, Erar, 2011].

The requirement to run many MCMC samplers is a major drawback since it is computationally burdensome. Several other methods have been proposed that allow the state space size to vary within a single MCMC sampler. Of these, the most popular is reversible jump MCMC (RJMCMC); another option is birth-death MCMC (BDMCMC).

Both of these options provide mechanisms for adding and removing states from the state space. RJMCMC was introduced by Green [1995]; it is a general method applicable to problems where the dimension of the parameter space varies. The algorithm has been applied to FMMs by Richardson and Green [1997], Cappé et al. [2003], Jasra et al. [2007], and Dellaportas and Papageorgiou [2006], and to HMMs by Spezia [2010]. The idea is to augment a standard MCMC sampler for fixed state space size with additional move types that add or delete a state, split one state into two, or merge two states into one. In order to ensure that the detailed balance equations are satisfied and the sampler converges to the desired posterior distribution, these dimension-changing moves must be reversible in the sense that a move that splits one state into two can be "undone" by a move that merges those two into one state, and vice versa.

BDMCMC is another approach that was developed for FMMs by Stephens [2000b] and extended to HMMs by Cappé et al. [2003]. The basic idea is similar to RJMCMC, but new states are added and removed in continuous time. In the simplest formulation, "births" of new states occur according to a Poisson process with a fixed rate parameter. "Deaths" of states occur according to independent Poisson processes with rate parameters depending on the likelihood of the corresponding state parameters; states that do not contribute to explaining the data die at higher rates than those that do. Cappé et al. [2003] included other types of moves, such as split and merge moves, within this framework.

2.4 Discriminative Approaches with Sequential Dependence

In this Section, we discuss two discriminative approaches to classification in data with sequential dependence. We discuss Conditional Random Fields (CRFs) in Subsection 2.4.1 and a method for combining discriminative models for i.i.d. data with the time dependence structure of HMMs in Subsection 2.4.2. There is also an extensive line of research in the speech recognition literature that optimizes functions other than the negative log-likelihood to estimate the parameters of the HMM in a discriminative fashion. One of these approaches is equivalent to maximizing the likelihood of the CRF. We will not discuss the others of these methods in detail; see Jiang [2010] and He et al. [2008] for more thorough reviews of these methods.

2.4.1 Conditional Random Fields

Conditional Random Fields (CRFs) were originally proposed in the computer science literature by Lafferty et al. [2001] and have been applied to a variety of problems, including part-of-speech tagging [Lafferty et al., 2001], gene prediction [Bernal et al., 2007] and physical activity classification [Vinh et al., 2011] among many others. The idea of the CRF is to use a graphical model for the conditional distribution of $\mathbf{Y}_i|\mathbf{X}_i$. Estimation of CRF models can be performed in either a frequentist or Bayesian setting, but Bayesian methods are less well developed; we focus on frequentist methods in this review.

The model uses a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the dependence relationships among the variables $Y_{i,t}$. The vertices of the graph are the components of the vector $\mathbf{Y}$ and the edges specify

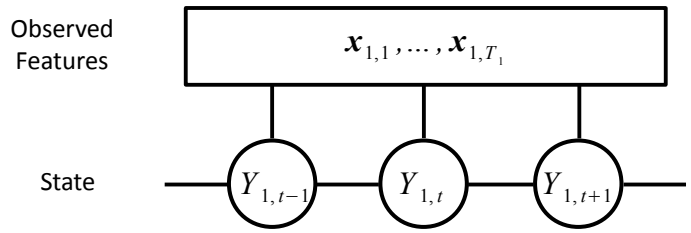


Figure 2. A diagram illustrating the linear chain CRF with a single observation sequence, $i = 1$. Random variables are shown with nodes. Boxes are random variables that we condition on, and circles are random variables that we model. The distribution of the state $Y_{i,t}$ depends on the states at the adjacent two time points in the sequence and the feature values at all times. Formulations with more complex dependence relationships are also possible, and entail adding more edges to the graph.

the conditional dependence structure of $\mathbf{Y}|\mathbf{X}$ as follows:

$$p(Y_{i,t}|\mathbf{X}, Y_{i^*,t^*}, (i,t) \neq (i^*,t^*)) = p(Y_{i,t}|\mathbf{X}, Y_{i^*,t^*}, Y_{i,t} \sim Y_{i^*,t^*}),$$

where $Y_{i,t} \sim Y_{i^*,t^*}$ indicates that the graph contains an edge connecting $Y_{i,t}$ and Y_{i^*,t^*} [Lafferty et al., 2001]. In words, $\mathbf{Y}_i|\mathbf{X}_i$ has the Markov property with respect to the graph.

If the graph has the structure of a chain where $Y_{i,t}$ is connected to $Y_{i,t-1}$ and $Y_{i,t+1}$ as illustrated in Figure 2, it is common to assume that the conditional distribution has a p.m.f. with the following form [Sutton and McCallum, 2011]:

$$f(\mathbf{y}|\mathbf{x}) = \frac{exp\left\{\sum_{k=1}^{K}\sum_{t=1}^{T_i-1}\lambda_k f_k(y_{i,t}, y_{i,t+1}, \mathbf{x})\right\}}{\sum_{\mathbf{y}_i^*} exp\left\{\sum_{k=1}^{K}\sum_{t=1}^{T_i-1}\lambda_k f_k(y_{i,t}^*, y_{i,t+1}^*, \mathbf{x})\right\}}. \tag{2.4.1}$$

Here, the index $\mathbf{y}_i^*$ of the summation in the denominator runs over all possible sequences of classes. If we have N observation sequences, the joint conditional distribution is a product of N terms of this form. It can be shown that the conditional model for $\mathbf{Y}|\mathbf{X}$ that results from conditioning on $\mathbf{X}$ in the first-order HMM is a special case of the CRF that can be achieved by particular choices of the functions f_k in Equation (2.4.1) [Lafferty et al., 2001, Sutton and McCallum, 2011]. However, the CRF is more general in two ways: first, the graph structure can be arbitrary, and second, more general functions f_k can be used.

The formulation of the CRFs given above specifies a conditional model for the full observation sequences $\mathbf{Y}_i$. Some authors have modified this formulation to give a model for the individual states $Y_{i,t}$ [Torralba et al., 2004, Taskar et al., 2004].

Usually, the only parameters that are estimated are the weights λ_k; the graph structure and the functions f_k are treated as fixed at the time the λ_k are estimated. However, several authors use a two-stage estimation strategy in which parameters for these functions are estimated first, and then the functions are combined through the CRF. We will discuss a special case of this where each f_k is derived from a CRF or HMM in Subsection 2.6.1, and another variation where the f_k are treated as general functions which are estimated non-parametrically in Subsection 2.6.4.

In CRFs with relatively simple graphical structures such as the linear chain model pictured in Figure 2, recursive algorithms can be used to compute the value of the likelihood function and its gradient efficiently. We will present these algorithms in detail in Chapter 5. When the graph has a more complicated structure, exact evaluation of the likelihood is computationally infeasible and approximate methods must be used. In either case, estimates of the CRF parameters are most often obtained using general purpose gradient ascent type algorithms; see Sutton and McCallum [2011] for a general discussion of the application of these methods to CRFs. Several authors have introduced boosting methods for estimating CRFs; we will discuss these methods in Section 2.6.4.

2.4.2 Discriminative HMMs, McShane et al. [2013]

McShane et al. [2013] take a different approach to discriminative estimation of HMM parameters by rewriting the model in a form that replaces estimation of the multivariate density for $\mathbf{X}_{i,t}|Y_{i,t}$ with the simpler problem of estimating the distribution of $Y_{i,t}|\mathbf{X}_{i,t}$. With this reparameterization, they are able to employ well-developed classification methods that are applicable for i.i.d. data in order to perform classification with a HMM. This preserves the advantages of the generative approach in terms of the relative ease of parameter estimation, while gaining some of the superior classification performance of discriminative approaches when the distribution $\mathbf{X}_{i,t}|Y_{i,t}$ is difficult to model. This approach can be viewed as a hybrid approach in which the parameters describing transitions between activity types are estimated by maximum likelihood in the joint HMM, but the parameters describing the connections between the features $\mathbf{X}$ and the classes $\mathbf{Y}$ are estimated discriminatively.

2.5 Moderate to High Dimensional Observations

When the dimension D of the random vector $\mathbf{X}_{i,t}$ observed at each time point is large, the basic formulations of the HMM and CRF models discussed above can overfit the training data, resulting in reduced classification performance on new data. Overfitting can result if the dimension is large relative to the number of observations because noise in the training data can falsely indicate differences in the distributions of observed data between states.

In this Section we review ideas that have proposed for addressing this problem with HMMs and CRFs in both the frequentist and Bayesian frameworks. These strategies generally combine one or more of the following ideas:

1. an initial dimension-reducing data transformation before fitting the model
2. a reduced number of the observed covariates, either overall or for each state
3. penalized estimation
4. boosting, a method to combine inferences from multiple simplified models

The first of these strategies is more commonly applied with HMMs than with CRFs. However, variable selection, penalized estimation and boosting have been applied with both HMMs and CRFs. We review methods for performing an initial dimension reduction in Subsection 2.5.1 and methods for variable selection and penalized estimation in Subsection 2.5.2. We will discuss boosting in Section 2.6.

2.5.1 Initial Dimension Reduction

One option is to reduce the dimensionality of the data before fitting the classification model. One way of doing this is through principle components analysis (PCA), which has been recommended as a first step when fitting FMMs by Dellaportas and Papageorgiou [2006] and Jasra et al. [2007]. Some other approaches that can be used to perform an initial data reduction include linear discriminant analysis (LDA) and its extensions. We discuss these ideas in this Subsection.

The idea behind PCA is to create a small set of new variables that are linear combinations of the originally observed variables and which capture most of the variability in the original data. PCA does this by projecting the observed data vectors, which lie in $\mathbb{R}^D$, onto $D' < D$ linearly independent subspaces of $\mathbb{R}^D$. These subspaces are generally taken to be those spanned by the

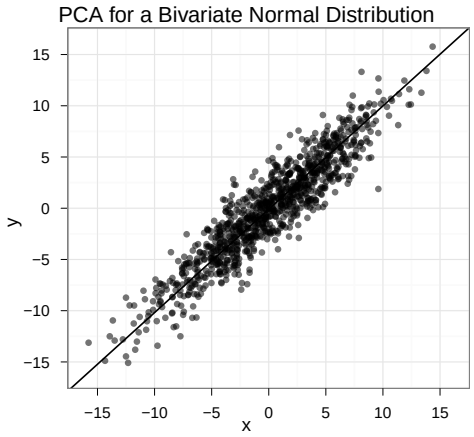
eigenvectors corresponding to the D' largest eigenvalues of the sample covariance matrix [Ravishanker and Dey, 2002]. This is illustrated in Figure 3(a) for the case of reducing 2-dimensional data to 1 dimension.

The projection of a high dimensional Gaussian random variable onto a lower dimensional space also follows a Gaussian distribution. Liu et al. [2003] argue that this provides some justification for the use of PCA with Gaussian HMMs and FMMs: if the observed data follow a multivariate Gaussian distribution conditional on the state, the variables created using PCA will also follow a Gaussian distribution conditional on the state. Thus, the HMM or FMM applies equally well to the variables obtained through PCA as it did to the original variables.

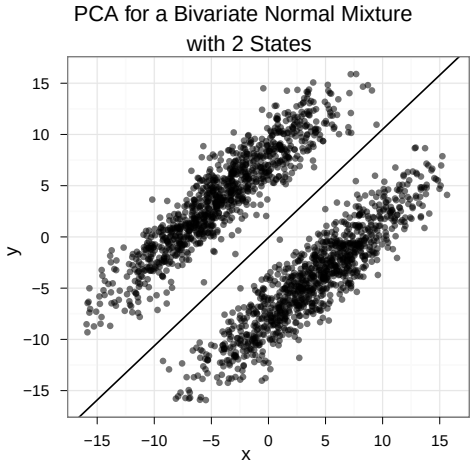
However, PCA can make it more difficult to distinguish between different classes. An example of how this might occur is shown in Figure 3(b). The difficulty here is that observations from two different Gaussian components are projected to the same place on the axis defined by the first principle component. As a result, there is no way to tell the observations from different states apart based on the subspace selected by PCA. If a different projection had been used, we would have been able to capture the differences between the states in a smaller number of variables. Liu et al. [2003] have proposed a Bayesian approach that helps to remedy this problem for FMMs by simultaneously fitting the model and selecting which of the principle components are most informative for distinguishing between the states.

Linear discriminant analysis (LDA) is another technique that projects the data onto a lower-dimensional linear subspace, but this method seeks to find the optimal subspace which maintains separation among the states. In order to use this method, the data set used to train the model must include the true state corresponding to each observation. Given this data, the observed features are projected onto the subspace of dimension D' which maximizes the variance among the means for the different states relative to the variances for observations within each state [Hastie et al., 2009].

There are two practical problems with using the standard implementation of LDA for HMMs. First, it requires that the state be known for the data used to fit the model; often this is not available. Even when some classification information is available, there may be multiple states in the HMM corresponding to each observed class. Second, the standard approach is based on the assumption that the observed features follow multivariate Gaussian distributions with the same covariance matrix in each state. These problems can both be resolved to some extent through the



(a) The diagonal line represents the first principal component; it indicates the direction in which the observed data varies the most. A new variable is created by computing the distance between the data mean and the projection of each observed (x, y) onto this line.



(b) The diagonal line again represents the first principal component. The projection of the data points onto this line fall into the same region; the first principal component does not give us enough information to distinguish between the two states.

Figure 3. Principle Components Analysis

use of other dimension-reduction methods such as mixture discriminant analysis [Hastie and Tibshirani, 1996] or local Fisher discriminant analysis [Sugiyama, 2007]. These methods both extend LDA to the case that the data follow a Gaussian FMM within each class.

2.5.2 Variable Selection and Shrinkage

Another set of ideas focuses on selecting the best subset of the observed variables to use in fitting the model. Approaches to this task can be divided between variable selection methods where each variable is either in or out of the model, and shrinkage methods where the influence of each observed covariate in determining cluster membership is scaled down according to the amount of information contributed by that covariate. Some shrinkage methods may also perform variable selection, for instance by shrinking the parameters associated with a particular variable to 0. Shrinkage methods can often be interpreted either as frequentist estimates obtained by maximizing a penalized likelihood function or as Bayesian estimates under an informative prior [Hastie et al., 2009]. In this Section we review several approaches for estimation

with variable selection and shrinkage.

Dy and Brodley [2004] propose a forward search algorithm for feature selection in a Gaussian FMM. In each step, features are evaluated based on how much they increase the likelihood or a geometric measure of separability between clusters similar to that used in LDA. Covariates that lead to better clustering according to one of these measures are added to the model one at a time. Raftery and Dean [2006] propose a similar algorithm where models are compared using BIC as an approximation to Bayes Factors. A variable is added (or removed) from the model one at a time if there is evidence that the variable does (or does not) contribute information about the clustering given the other variables in the model.

As we mentioned in Subsection 2.5.1, Liu et al. [2003] propose a method to select the principle components that are most informative in clustering while fitting a Gaussian FMM in the Bayesian framework. This can be considered as a variable selection method where the principle components are selected rather than the originally observed covariates. They achieve this by marginalizing over the mean and covariance parameters and using a Gibbs sampler that alternates between sampling the mixture component allocation of each observation conditional on the informative covariates and the value of a vector indicating whether or not each principle component is informative. This algorithm can be used for clustering, but does not estimate the parameters for the observation distribution in each state.

Variable selection ideas have also been developed for CRFs. For example, McCallum [2003] uses an iterative forward search procedure, beginning with a model that does not include any features and then gradually adding new features and interactions between features that are already included in the model. In each step, the features that result in the largest increase in the CRF log-likelihood are added to the model and all coefficients are updated. The model training process halts when a threshold on the minimum increase in the log-likelihood in each iteration is reached.

We now turn to shrinkage techniques. One simple approach for shrinkage is used by Carbonetto et al. [2003] for Gaussian FMMs. They specify a hierarchical prior distribution for the mean vector in each state that allows for shrinkage of components of the mean vector corresponding to features that are similar in different states. That is, if an observed covariate has a similar mean in all states, estimates of those means are shrunk towards each other. One limitation to this method is that it does not apply shrinkage to the variance of the covariates, which

may be different in different states.

Shrinkage methods have also been suggested for the alternative covariance parameterizations we mentioned in Section 2.3. Two articles focus on penalized estimation in a simplified normal FMM where the covariance matrix for each state is assumed to be the same diagonal matrix. Pan and Shen [2007] work with centered data and find penalized maximum likelihood estimates subject to an L_1 penalty on the means of each covariate in each component of the mixture. Wang and Zhu [2008] introduce two new possibilities for penalization terms, both of which enforce a grouping on the means. Under these schemes, the means for the same covariate in different mixture components are shrunk towards 0 together. Galimberti et al. [2008] introduce an L_1 penalty in a model related to a mixture of factor analyzers; the penalty has the effect of shrinking the factor loadings toward 0 for factors that are less important in explaining the clustering in the data.

In applications of CRFs in computer science, there are often hundreds of thousands of model parameters. In order to reduce the potential to overfit the data, it is very common to introduce a penalty on the magnitude of the coefficients λ_k [Sutton and McCallum, 2011, Smith, 2007, Vail et al., 2007a]. An L_2 penalty is used most commonly, since numerical gradient ascent methods are often used in parameter estimation and the L_2 penalty is differentiable. However, methods to perform estimation with L_1 penalties have also been developed [Vail et al., 2007a]. Use of an L_1 penalty can be regarded as a mechanism for performing variable selection, since the penalty tends to shrink some coefficient estimates to exactly 0.

2.6 Ensemble Methods

In this Section, we discuss approaches for combining several classification models into one larger model. We begin by reviewing functional forms that have been proposed for combining the component models in Subsection 2.6.1. In the following three Subsections, we discuss three specific strategies for obtaining a diverse collection of component models: we cover bagging in Subsection 2.6.2, the use of different subsets of the features in Subsection 2.6.3, and boosting in Subsection 2.6.4. Throughout this Section, we will draw from the literature on ensemble methods for static classification in addition to ensemble methods that have been developed for the dynamic HMM and CRF models. We will use M to denote the total number of component mod-

els in the ensemble, and collect the parameters for all of these component models in the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)})$.

2.6.1 Functional Forms for Combining Multiple HMMs or CRFs

Many different methods have been suggested for combining multiple classification methods in order to obtain one overall model. In the context of HMMs and CRFs, one of the first suggestions for achieving this was the Product of HMMs [Brown and Hinton, 2001], which is closely related to the Logarithmic Opinion Pool of CRFs (LOP) [Smith et al., 2005, Smith, 2007]. The LOP model is given by

$$p(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) = \frac{\exp\left[\sum_{m=1}^{M} \alpha_m log\{p(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}^{(m)})\}\right]}{\sum_{\mathbf{y}'_i} \exp\left[\sum_{m=1}^{M} \alpha_m log\{p(\mathbf{y}'_i|\mathbf{x}_i;\boldsymbol{\theta}^{(m)})\}\right]} \tag{2.6.1}$$

Here, the $\alpha_1, \ldots, \alpha_M$ are non-negative model weights with $\sum_{m=1}^{M} \alpha_m = 1$, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)})$ collects the parameters of the M component CRF models, and $p(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}^{(m)})$ is the conditional probability of the observation sequence $\mathbf{y}_i|\mathbf{x}_i$ from the mth CRF. As before, the summation in the denominator is over all state sequences $\mathbf{y}_i$. The Product of HMMs is essentially the same as this model, but it takes $\alpha_m = 1 \, \forall m = 1, \ldots, M$.

The LOP is a particularly convenient method for combining CRFs because by rearranging the order of summations, we can interpret the resulting combined model as either a CRF where each feature function f_m is the log-probability of the given observation sequence obtained from the mth component CRF, or as a CRF with parameters equal to a weighted average of the parameters in the component CRFs [Sutton et al., 2006]. This observation enables fast computation of the class probabilities at each time point, the model likelihood, and the gradient of the likelihood using the forward-backward recursions developed for CRFs.

Much of the work that has been done on LOP parameter estimation focuses on introducing variability into the set of CRFs that are combined in the LOP. Bagging and the use of feature subsets are two examples of methods that have been suggested to achieve this. This goal can be motivated by the so-called ambiguity decomposition for LOPs, derived by Heskes [1998]. Let $p(y|\mathbf{x};\boldsymbol{\theta})$ be the conditional probability estimate formed by combining the individual estimates $p(y|\mathbf{x};\boldsymbol{\theta}^{(m)})$ as in Equation (2.6.1). If we denote the true distribution being estimated by $q(y|\mathbf{x})$,

the ambiguity decomposition states that

$$K\{q(y|\mathbf{x}), p(y|\mathbf{x};\boldsymbol{\theta})\} = \sum_{m=1}^{M} \alpha_m K\{q(y|\mathbf{x}), p(y|\mathbf{x};\boldsymbol{\theta}^{(m)})\} - \sum_{m=1}^{M} \alpha_m K\{p(y|\mathbf{x};\boldsymbol{\theta}), p(y|\mathbf{x};\boldsymbol{\theta}^{(m)})\} \tag{2.6.2}$$

Here, $K(q,p)$ is the Kullback-Leibler divergence of p from q. The first term on the right hand side of Equation (2.6.2) is a measure of the average quality of the individual conditional density estimates. The second term, referred to as the ambiguity, is a measure of how similar the individual estimates are to the combined estimate. It follows from this decomposition that we can improve the the quality of the combined estimate by simultaneously

1. minimizing the first term; i.e., ensuring that each individual estimate gives an accurate representation of the true conditional distribution, and

2. maximizing the second term; i.e., ensuring that the individual estimates are diverse.

The model weights can either be estimated at the same time as the parameters of the component models, or in a second stage. However, previous researchers have found that estimating the model weights does not yield much improvement in classification performance relative to using uniform weights [Smith and Osborne, 2007, Baldridge and Osborne, 2008, Dietterich et al., 2004].

Another method that has been suggested for combining multiple models is through a linear combination of the class probabilities:

$$p(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m p(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}^{(m)})$$

In this case, we require that the model weights are non-negative and sum to 1 to ensure that the model gives a valid probability distribution over the space of all sequences $\mathbf{y}_i$. Both Sutton et al. [2006] and Smith [2007] used this form to combine the predictions from CRFs in applications to text processing, and they found that the resulting models did not perform as well as when a LOP was used.

Smith [2007] proposed two additional methods, which they refer to as a per-transition product of experts and a per-transition mixture. These are analogous to the LOP and the linear combination discussed above, but specify models for $p(y_{i,t}|y_{i,t-1}, \mathbf{x}_i;\boldsymbol{\theta})$ rather than for the full sequence $\mathbf{y}_i$. Again, they found that these models do not perform as well as the LOP when applied to real data sets.

2.6.2 Bagging

Bagging is the idea of combining many models, where the parameters of each model are estimated using a different sample drawn with replacement from the originally observed data. In settings with sequential dependence, bagging can be implemented by resampling complete observation sequences.

Smith and Osborne [2007] employed this approach with a parametric specification of CRFs, and found that it led to a small improvement in classification performance, though not as large as was achieved by using a manually constructed subset of the features in each model (we discuss this approach next). They motivated the use of bagging via the ambiguity decomposition for LOPs. Bagging is a strategy to obtain a more diverse set of component models.

Bagging has been better studied in combination with static classification methods combined in linear opinion pools. One approach to explaining the benefits of bagging is to view it as a mechanism for reducing the sampling variance of the class or regression estimates [Hastie et al., 2009]. Suppose our classification model has low bias but high sampling variance, in that the estimated class for a particular input $\mathbf{x}$ is very sensitive to the data used to estimate the model parameters. By combining the predictions from many different training data sets, we can reduce that variance while maintaining the low bias. Drawing a bootstrap sample from the originally observed data provides a mechanism for obtaining these different training data sets. Biau et al. [2008] also examine bagging and prove that the procedure yields a consistent classifier as long as the original classifier was consistent, and demonstrate that in some cases bagging can create a consistent classifier even if the original classifier was inconsistent.

Bagging is often applied with classification and regression trees, which have high variance and therefore benefit from the variance reduction that bagging provides [Hastie et al., 2009]. In that context, bagging is one of the foundational ideas behind random forests; we will discuss the other main idea behind random forests, random feature selection, in the next Subsection.

2.6.3 Feature Subsets

Another common way of introducing diversity into the set of component models is by restricting the model estimation process for each component model to use a subset of the observed features. Smith and Osborne [2007], Smith [2007], and Sutton et al. [2006] implement this strategy with CRFs by manually creating M subsets of the full feature set that they believe will be

separately informative and training each model using one of these subsets. The predictions from these separately trained models are then combined in the LOP. In applications, they find that this combined classifier performs better than a single model trained using the full feature set and is more robust to differences between the training data set and the test set.

Smith and Osborne [2007] motivate this approach via the ambiguity decomposition for LOPs, which we discussed in Section 2.6.1. Sutton et al. [2006] discuss a second motivation for using a diverse group of component models trained with different feature subsets. They observe that often, a small subset of the covariates used in CRFs is much more individually informative than the others. If penalized estimation is used to estimate the CRF parameters, this means that the coefficient estimates for those covariates will be large and the coefficients for other model terms will shrink towards zero, even if those other covariates are jointly informative. This effect, referred to as weight undertraining, can be problematic for two reasons. First, sometimes there are differences between the training and test data sets that mean that covariates that were informative in the training set are no longer informative in the test set. Second, it may be that the covariates whose coefficients were shrunk were actually jointly more informative than the covariates with large coefficients. They suggest that these problems can be mitigated by combining several diverse models.

Static random forests use a similar idea as part of the estimation process. In this case, during the process of growing the component classification or regression trees, each node split is obtained by searching over a randomly selected subset of the features. The feature in this subset yielding the largest gain in regression or classification performance is selected.

For random forest algorithms, analysis shows that the performance of the final classifier improves if the quality of the individual trees improves and the correlation between the predictions of the individual trees is reduced [Breiman, 2000, Amit and Geman, 1997, Breiman, 2001, Hastie et al., 2009]. This is similar in spirit to the ambiguity decomposition for LOPs. Use of different random feature subsets in different trees is one way of reducing the correlation between their predictions.

2.6.4 Boosting to Combine Multiple HMMs or CRFs

In this Subsection we review boosting methods in general and discuss some additional ideas that have been developed for using boosting to create ensembles of HMMs and CRFs.

Boosting is a general method for obtaining a strong classifier $F(\mathbf{x})$ by combining several weaker classifiers $f_m(\mathbf{x})$, $m = 1, \ldots, M$. Depending on the formulation, each base classifier may map the feature vector $\mathbf{x}$ to an estimated class or to a vector of estimated probabilities for each class. A key idea in boosting is that we fit the base classifiers sequentially, and each new classifier that is estimated focuses more on cases that were misclassified by earlier classifiers and less on cases that were classified correctly by earlier classifiers.

One of the most common boosting algorithms is AdaBoost (or adaptive boosting), which was originally proposed for binary classification problems and has since been adapted to other settings such as multiple classification and regression. In this algorithm, new classifiers are successively fit to a weighted version of the sample in which observations that were correctly classified by the previous classifiers are given less weight than those that were not. The final classifier is a weighted combination of these base classifiers, where the weight for the mth base classifier is a function of the error rates of the base classifiers obtained in iterations 1 through m [Freund and Schapire, 1995].

The particular weighting scheme used in the AdaBoost algorithm arises from performing approximate stagewise additive modeling under an exponential loss function [Friedman et al., 2000]. That is, the base classifiers are obtained one at a time by approximately minimizing an exponential loss function while holding previously estimated classifiers fixed. The exponential loss is a function of a weighted sum of the classification results from each base classifier.

It is often helpful to "slow down" convergence to a minimum of the loss function by shrinking the contribution of each base classifier $f_m(\mathbf{x})$ to the overall classifier $F(\mathbf{x})$. This can be done by multiplying the loss-minimizing weight for the mth base classifier by ν, where $0 < \nu < 1$. Hastie et al. [2009] recommend using a value of ν that is "very small ($\nu < 0.1$)". In general, there is a tradeoff between the amount of shrinkage used and the number of additive components that are fit: the smaller ν is, the larger M must be to obtain the same explanatory power. The best results are often obtained with a small value of ν and a large number of additive components.

Different loss functions can also be used, resulting in variations on the boosting algorithm [e.g. Hastie et al., 2009, Domingo and Watanabe, 2000]. For binary classification problems, one alternative loss function is the binomial deviance. Minimizing the binomial deviance is equivalent to maximizing the binomial log-likelihood where the probabilities can be interpreted as the logistic transform of the value of the weighted sum of base classifiers [Hastie et al., 2009]. Sev-

eral ways of extending boosting to classification problems with more than 2 classes have been proposed: using an extension of the exponential loss to multiple classes [Zhu et al., 2009], using the multinomial likelihood instead of the binomial likelihood [Friedman et al., 2000, Friedman, 2001], and converting the problem into several binary "one vs. the rest" classification problems [Schapire and Singer, 1999], among others.

It has been shown that the AdaBoost algorithm performs poorly when the Bayes error rate is large or the data used to train the model have incorrect labels [Hastie et al., 2009, Dietterich, 2000]. Intuitively, this is because the exponential weighting scheme assigns increasingly large weights to previously-misclassified observations. In the case of mislabeled training data, this means that the base classifiers that are trained in later boosting iterations focus more on the mislabeled observations. An advantage of using the binomial deviance loss function is that it assigns less weight to incorrectly classified observations than the exponential loss. As a result, it is more robust to mislabeled data and high Bayes error rates [Hastie et al., 2009].

Moreover, it has been proven that in the presence of random misclassification error in the labels of the training data, any boosting algorithm that proceeds by stagewise additive modeling under a convex loss function can generate a classifier that misclassifies as often as 50% of the time [Long and Servedio, 2010]. Alternative boosting algorithms that do not work by optimizing a convex loss function have been suggested to address this problem in the binary classification setting [Freund, 2001, Kalai and Servedio, 2003, Long and Servedio, 2008].

In order to select the number of boosting iterations (i.e., the number M of component models), it is common to use evaluation of the combined model's performance on a held-out validation data set. This strategy can mitigate problems in terms of overfitting the data. In fact, it is been shown that some boosting algorithms approximate L_1 penalized model estimation, with the number of boosting iterations corresponding to the magnitude of the penalty parameter [Hastie et al., 2009, Rosset et al., 2004]. For those algorithms, selecting the number of boosting iterations by validation of model performance on a held-out data set corresponds to performing cross-validation to select the penalty parameter.

Boosting methods have been adapted for use with HMMs and CRFs by several authors. Approaches to boosting with HMMs have been based on the setup for training HMMs for classification with labeled data in which a separate HMM is fit to observation sequences from each class. The most direct approach to applying the Adaboost algorithm with this classification model is

to simultaneously fit a new HMM for every class in each boosting iteration, thereby obtaining a new classifer in each iteration. This can be done by drawing a sample with replacement from the set of all observation sequences using sampling weights derived from the classification error rates from previously fit classifiers and then fitting a HMM to the sampled sequences for each class, as in Dimitrakakis and Bengio [2004], Darmanjian et al. [2007], and Darmanjian and Principe [2008]. Zhang et al. [2005] achieve a similar result by training the new set of HMMs with a modified EM algorithm where the observed samples are explicitly weighted. In both cases, a linear combination of the individual classifiers obtained from each boosting iteration is used for the final classifier, as in the original AdaBoost algorithm.

Foo and Dong [2003] and Foo et al. [2004] take a different approach to boosting with a similar model. Rather than fitting a new set of HMMs to obtain an additional classifier in each boosting iteration, they modify the model so that a weighted sum of HMMs is used for each class-specific model. This weighted sum of HMMs is obtained through a modification of the AdaBoost algorithm, where the individual HMMs in the sum for a given class are fit using a weighted sample of the training sequences for that class. They build up the class-specific models by adding one additional HMM at a time, but a single classifier that compares the data likelihood from each class-specific model is used throughout. Chen and Chen [2009] use a similar approach in a binary classification problem. With this approach, it would also be possible to combine the resulting class-specific HMMs into a single larger HMM.

One of the first applications of boosting to training CRFs was proposed by Dietterich et al. [2004], and uses gradient tree boosting. Rather than specify a linear functional form in known functions f_k as in Equation (2.4.1), they model the conditional distribution of $\mathbf{Y}_i|\mathbf{X}_i$ as

$$P(\mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{exp\left[\sum_t F^{y_{i,t}}\{y_{i,t-1}, w_t(\mathbf{x})\}\right]}{\sum_{\mathbf{y}_i^*} exp\left[\sum_t F^{y^*_{i,t}}\{y^*_{i,t-1}, w_t(\mathbf{x})\}\right]}. \tag{2.6.3}$$

In this expression, $w_t(\mathbf{x}_i)$ is a windowing function that extracts the values of $\mathbf{x}_i$ occurring in a window around the time t. They then approximate each of the functions F^k with a linear combination of regression trees that is updated iteratively. In this process, the mth tree is fit so as to approximate the gradient of the contribution to the log-likelihood from each observation $(y_{i,t}, w_t(\mathbf{x}_i))$ with respect to F^k, evaluated at the approximation of F^k obtained after iteration $m-1$. We will describe this approach in more detail in Section 5.3. In applications to four data sets, Dietterich et al. [2004] find that their method outperforms the feature search method of

McCallum [2003], described in Subsection 2.4.1, on two data sets and does worse on the other two. Variations on this algorithm have since been suggested by Gutmann and Kersting [2007] and Liao et al. [2007]. Torralba et al. [2004] develop another boosting algorithm in an application to image segmentation that applies gradient boosting separately to two problems: (1) learning the local relationship between the covariates at one pixel and that pixel's label, and (2) learning the graph structure and relationships between different pixels. This algorithm is unique in that it does not treat the graph structure as fixed and known.

The boosting method of Dietterich et al. [2004] can be cast as a method for fitting a LOP of CRFs. If we let F_m^k be the mth regression tree used in the approximation of F^k, we can rewrite the model as follows:

$$\begin{aligned} P(\mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) &= \frac{exp\left[\sum_t F^{y_{i,t}}\{y_{i,t-1}, w_t(\mathbf{x})\}\right]}{\sum_{\mathbf{y}_i^*} exp\left[\sum_t F^{y^*_{i,t}}\{y^*_{i,t-1}, w_t(\mathbf{x})\}\right]} \\ &= \frac{exp\left[\sum_t \sum_{m=1}^M \alpha_m F_m^{y_{i,t}}\{y_{i,t-1}, w_t(\mathbf{x})\}\right]}{\sum_{\mathbf{y}_i^*} exp\left[\sum_t \sum_{m=1}^M \alpha_m F_m^{y^*_{i,t}}\{y^*_{i,t-1}, w_t(\mathbf{x})\}\right]} \\ &= \frac{exp\left[\sum_{m=1}^M \alpha_m \sum_t F_m^{y_{i,t}}\{y_{i,t-1}, w_t(\mathbf{x})\}\right]}{\sum_{\mathbf{y}_i^*} exp\left[\sum_{m=1}^M \alpha_m \sum_t F_m^{y^*_{i,t}}\{y^*_{i,t-1}, w_t(\mathbf{x})\}\right]} \end{aligned}$$

This is analogous to the LOP model of Equation (2.6.1), where $\sum_t F_m^{y_{i,t}}\{y_{i,t-1}, w_t(\mathbf{x})\}$ plays the role of $log\{p(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta}^{(m)})\}$. Edakunni et al. [2012] develop a general boosting algorithm for Products of Experts that builds on this connection.

2.7 Physical Activity Classification

Activity classification and energy expenditure estimation are active areas of research, and many different methods have been suggested for performing these tasks using accelerometer data. In the early literature, most approaches were limited by the use of accelerometers that recorded only a univariate discretization of the acceleration experienced over the course of intervals of one or more seconds in length, known as "counts". As better technology has become available that records acceleration along multiple axes at higher frequencies, a wider variety of methods have been developed to take advantage of those data.

These newer methods generally begin by dividing the acceleration time series into a series of windows centered at different points in time. A vector of features is then obtained that summa-

rizes the acceleration signal within each separate window. A statistical classification or machine learning algorithm is used to relate the features extracted from each window to the activity type or energy expenditure level for that window. This general approach of dividing a signal into windows is also commonly used in other fields such as speech recognition [Rosti, 2004, Kumar and Andreou, 1998].

We will review each of these steps in turn. We discuss some choices related to how the windows are constructed in Section 2.7.1 and the features that are extracted in Section 2.7.2. We present models that have been used for activity classification in Section 2.7.3, and models for energy expenditure in Section 2.7.4. We will also mention some research from related disciplines that use similar methods.

2.7.1 Windows

One choice to be made when dividing the signal into windows is how long the windows should be. Window lengths used in previous research have generally varied between 1 and 60 seconds. Several researchers have found that classification accuracy increases as window lengths become longer. Mannini et al. [2013] fit separate models using data obtained from windows of sizes 2, 4, and 12.8 seconds and find that classification accuracy is generally better for the models with longer window lengths. Similarly, Trost et al. [2012] examine window lengths varying from 10 seconds to 60 seconds. They find that a 10 second window is adequate for classifying sedentary activity, moderate-to-vigorous household activity, sports, walking, and running, but a 60 second window is necessary to improve classification rates for light-intensity household activity. Similar findings have been reported by Bonomi et al. [2009a] and Sasaki [2013].

Although longer windows have been consistently found to lead to superior classification accuracy, they have some drawbacks. One is that in real-time classification settings (e.g., for the purposes of making health interventions), longer windows lead to greater latency and slower feedback to the user [Mannini et al., 2013]. Another problem is that using a longer window increases the chances that a single window covers more than one activity class. This is particularly a problem in free-living data where activities are not performed for known periods of time that start and end on the minute [Lyden et al., 2014]. Features calculated from a window that covers multiple activity types may not bear much resemblance to the features that would be observed from either activity type alone, causing difficulties for the classification algorithm.

Two studies have examined the possibility of addressing this problem by determining the start and end points of windows empirically. Anderson [2013] uses statistical change-point algorithms to identify time indices that may mark a boundary between different activity classes. These algorithms work by comparing the distribution of time series values in an interval before and after each point in time. If these distributions are different according to a specified measure, the given point in time is used as the boundary of a window for the classification algorithm. Anderson [2013] found that the window boundaries obtained through these methods were noisy and the approach did not perform as well as other options using fixed window size. Lyden et al. [2014] take a different approach, using specific knowledge of human activity to specify decision rules for identifying periods of inactivity and activity based on the accelerometer signal. They then divide periods of activity into windows of length 40 seconds (or greater). This prevents the windows from including periods of both activity and inactivity, though it is still possible that a window may include more than one type of activity.

Another option is to incorporate multiple window lengths in one classification procedure. Zheng et al. [2013] implement this idea by first dividing the time series into windows of length 10 seconds, and then further subdividing each 10 second window into subwindows of varying lengths from 1 to 10 seconds. They train ten separate models, one for each subwindow length. The predictions of these models are combined through a majority vote to obtain a single classification for each 10 second window. The overall classification accuracy is best with this combined model, though it is outperformed by one or another of the fixed-window-size models in each individual activity category. This suggests that patterns in the accelerometer signal that distinguish between activity classes may be available at several different time frames. Lester et al. [2005] also use features from multiple window lengths. In their setup, the window used to calculate each feature continually slides forward, so that it ends at the time the classification is to be performed.

2.7.2 Features

A wide variety of features have been used to summarize the acceleration signal in each window, drawing from the time domain, the frequency domain, and wavelet transforms, which capture both time and frequency domain characteristics. The time domain features used are standard summary statistics of the observed acceleration within the given window, including

the mean, standard deviation, percentiles, lag 1 autocorrelation, and so on. For accelerometers that measure acceleration along 3 axes, it is also possible to calculate features such as the correlation between these axes and the angle of acceleration.

The goal of the frequency domain features is to identify any cyclical trends in the acceleration time series and determine how much of the variation in the data is explained by those cyclical trends. In order to do that, they estimate properties of the spectral density, which is defined as the Fourier transform of the autocovariance function of the underlying stochastic process that generated the data. Intuitively, the spectral density is a non-negative function of ν that tells us how much of the lag l autocovariance in the time series is explained by oscillations of frequency ν, for each lag l [Shumway, 1988]. This also tells us about cyclical variation in the time series itself, because if the stochastic process that generates the data has periodic oscillations then its autocovariance function has oscillations of the same frequency [Engelberg, 2007]. The value of the spectral density function evaluated at ν is referred to as the power of the frequency ν.

The theory of the spectral density is based on the Fourier transform of the autocovariance rather than the original stochastic process because often the Fourier transform of the stochastic process does not exist, but the Fourier transform of the autocovariance does exist as long as the process is weakly stationary and satisfies certain other regularity conditions [Bartlett, 1978]. However, when we estimate the spectral density based on a finite sample from a time series, it is equivalent to work with the discrete Fourier transform of either the observed values or the sample autocovariance [Shumway, 1988]. An important consideration when working with a finite sample of size T is that we can only estimate the spectral density at a finite set of frequencies of the form $\frac{k}{T} \times$ (sampling frequency) for $k = 1, \ldots, T-1$, and only the first $\lfloor \frac{T}{2} \rfloor$ estimated frequencies will be unique (where $\lfloor a \rfloor$ indicates the greatest integer less than a) [Shumway, 1988]. Commonly used frequency domain features include the frequency and power of the first and second dominant frequencies (those frequencies with the highest and second-highest power), the frequency and power of the first and second dominant frequencies restricted to the band of frequencies from 0.6 to 2.5 Hz, ratios among these quantities, and so on.

2.7.3 Activity Classification

Many different algorithms and models have been applied to classify physical activity, including support vector machines [Anderson, 2013, Gyllensten and Bonomi, 2011, Mannini et al.,

2013, Ravi et al., 2005, Sasaki, 2013, Zhang et al., 2012, Zheng et al., 2013], classification trees [Albinali et al., 2010, Anderson, 2013, Bao and Intille, 2004, Bonomi et al., 2009a,b, Gyllensten and Bonomi, 2011, Mathie et al., 2004, Ravi et al., 2005, Zhang et al., 2012], artificial neural networks [Anderson, 2013, de Vries et al., 2011, Ermes et al., 2008, Gyllensten and Bonomi, 2011, Sasaki, 2013, Staudenmayer et al., 2009, Zhang et al., 2012], and nearest neighbors [Bao and Intille, 2004, Foerster et al., 1999, Ravi et al., 2005], among others. Two studies have also looked at combining inferences from multiple different types of classifiers, with some success [Gyllensten and Bonomi, 2011, Ravi et al., 2005]. A thorough review can be found in Preece et al. [2009]. These models routinely achieve classification accuracy in the range of 85 to 95% with laboratory data, and occasionally higher.

It has been shown repeatedly that classification accuracy suffers when models are trained with data from the laboratory and applied to perform classification with data gathered in a free-living environment [Duncan et al., 2011, Foerster et al., 1999, Gyllensten and Bonomi, 2011, Sasaki, 2013, Lyden et al., 2014]. The primary reason for this is that laboratory data are not representative of free-living data. These differences arise because people engage in a smaller variety of activities in the laboratory than they do in real life, and the way that they perform these activities is different. Only a limited number of activities can be performed in the laboratory; in most studies, subjects perform in the range of 10 to 20 specific tasks. However, people engage in a much wider variety of activity in daily life. This means that models trained with laboratory data are only equipped to recognize a small subset of the activities that people do in free-living settings [Bao and Intille, 2004, Duncan et al., 2011, Gyllensten and Bonomi, 2011]. It has also been shown that there are differences in the ways that people perform common activities in the laboratory and in real life. For example, people may sit up straighter in the laboratory than they do outside of it, and they may walk at a different or more uniform pace when there are other people around them [Foerster et al., 1999, Gyllensten and Bonomi, 2011].

Mixed results have been obtained when performing physical activity classification in free-living data using models that are trained with free-living data. In Bao and Intille [2004], subjects perform a set of 20 activities that are designed to be representative of daily life. In an effort to make the subjects more at ease, they are not observed by a researcher; instead, they annotate the start and end of each activity themselves. Their models achieve classification accuracies in line with results from laboratory studies. Ermes et al. [2008] employ a study design with two

components. In the first, subjects complete a supervised routine of activities and their activities are annotated by the researchers; in the second, participants are not restricted in the activities they could perform, and they annotate their own activity. Classification models trained and tested using data from the combined structured and unstructured components attain overall classification accuracy as high as 89%. Sasaki [2013] obtained data through direct observation of 15 seniors in a free-living environment for 2-3 hours each. Models using this free-living data achieve classification accuracy of 70 to 76%.

Overall, it appears that classification of activity gathered in a truly free-living environment is more difficult than classification of activity in a laboratory setting. This is supported by Gyllensten and Bonomi [2011], who observe that there seems to be more "overlap" among the features from different categories in free-living data than in laboratory data, suggesting that the classification problem is inherently more difficult in the free-living setting.

Two major suggestions have been made regarding changes to the classification models discussed above that could lead to higher success rates in classification. One is that models might benefit by taking into account variation between individuals. Mannini et al. [2013] and Bao and Intille [2004] observe that classification accuracy for some activities is consistently good for all subjects, while for other activities classification accuracy is very dependent on the subject. Bao and Intille [2004] show that using subject-specific models has the potential to increase activity recognition rates.

Another limitation to these models is that, while some temporal dependence may be captured within each window (e.g., through the computation of frequency-domain features), they treat each window as independent of the other windows. This discards some useful information, since there is in fact temporal dependence in activity type. This is because an individual is much more likely to perform the same activity in adjacent time periods than to change activities. It has been suggested that a model that accounts for this temporal dependence might have better classification accuracy than one that does not [Gyllensten and Bonomi, 2011, Mannini and Sabatini, 2010, Bao and Intille, 2004].

Several studies have used HMMs in order to capture temporal dependence in activity classification data. The most direct way to use HMMs for this purpose is to represent the true activity class by the hidden state, with the observed acceleration features following a distribution that depends on the state. Mannini and Sabatini [2010] use this approach and achieve classification

accuracy of 98.4% with partially simulated data. It should be noted that they use subject-specific models, and it is not clear how well the methods would translate to non-synthetic data. Pober et al. [2006] use a HMM that assigns 3 states to each activity class, and obtain overall classification accuracy of about 81%.

A second approach to using HMMs is to first use a flexible model that does not incorporate time dependency to obtain an initial classification, and then use a HMM to smooth those initial classifications over time. Lester et al. [2005] take this approach, using the AdaBoost algorithm with decision stumps for the initial classification that does not take time dependence into account. Class probabilities obtained from this first stage classifer are then used as input for a set of HMMs, with one HMM for each activity class. In order to obtain a classification for time t in an observation sequence where the true class is unknown, they apply each class-specific HMM to the features observed in a window of length Δ beginning at time t. The estimated class at time t is that corresponding to the HMM with the highest likelihood value. Anderson [2013] uses a similar idea, but employs decision trees, support vector machines, or neural networks for the first stage classifier. He uses a single HMM to represent the full time series, with one state corresponding to each activity class. The predicted classes from the first stage classifier are the observed values for the HMM, with the hidden state representing the true activity at each point in time. The Viterbi algorithm is used to estimate the true activity class when it is unknown.

Vinh et al. [2011] use a variation on the CRF referred to as the semi-Markov CRF to classify physical activity using accelerometer data. This model formulation relaxes the simple time dependence structures used in many CRF specifications. In addition to features calculated from the accelerometer recordings, they use time of day as a feature. In similar spirit to the methods we discussed in the previous paragraph, they perform an initial unsupervised clustering step, and use the resulting estimated cluster memberships as inputs to the CRF model. Notably, their objective is to classify activity according to activity categories including commuting, office work, lunch, and dinner; these activities are quite different from the categories used by health researchers, which tend to describe postural activities such as sitting or standing. While we are not aware of other applications of CRFs to physical activity classification with accelerometer data, CRFs have been employed in similar settings. For example, Vail et al. [2007b] use CRFs to classify the behavior of robots playing tag; rather than accelerometer data, their inputs are the positions of the robots over time.

2.7.4 Energy Expenditure Estimation

In this Section, we discuss methods that have been developed to estimate the intensity of physical activity from accelerometer data. Energy expenditure is often expressed in units of Metabolic Equivalent of Task (MET). One MET is the energy used by an individual at rest. The energy expenditure of other activities is expressed as multiples of this resting rate. For example, a three MET activity requires three times as much energy as resting. The intensity of physical activity is also often categorized as sedentary (≤ 1.5 METs), light (> 1.5 and < 3 METs), moderate (≥ 3 and < 6 METs), or vigorous (≥ 6 METs) [Ainsworth et al., 2011]. Many hypotheses about the relationship between physical activity and health outcomes, as well as the physical activity guidelines published by the U.S. Department of Health and Human Services, are given in terms of the amount of time spent in these categories [U.S. Department of Health and Human Services, 2008].

The most common approach to estimating energy expenditure in early research with accelerometer data uses a linear regression model that relates accelerometer counts to intensity, as in Freedson et al. [1998]. The difficulty with this approach is that activities with similar intensities may result in quite different counts when recorded with an accelerometer, and similarly, activities with different intensities may yield similar counts [Pober et al., 2006]. Also, these methods tend to work well only for activities similar to those used in estimating the regression coefficients. For example, if the data used to train the model focused on ambulatory activities, the model will typically perform adequately for ambulatory activities but poorly for other types of activity (and vice versa) [Lyden et al., 2011].

One idea for remedying the problems with linear regression models is to use a more flexible model for the relationship between the accelerometer signal and the activity intensity. This approach was taken by Rothney et al. [2007] and Staudenmayer et al. [2009] using artificial neural networks, and by Sasaki [2013] using support vector regression and random forest regression. All of these methods represent improvements over linear regression models.

Another common option is to use separate models for different types of activity. With this approach, in order to estimate energy expenditure we first estimate the physical activity classification at each point in time and then apply the appropriate models to estimate energy expenditure over time.

One of the first implementations of this idea used two regression models to estimate energy

expenditure based on accelerometer counts. The coefficient of variation of counts in each minute is examined to determine which model to use. This coefficient of variation was found to distinguish between walking/running activities and other activities effectively. The regression models are therefore targeted to work for these two groups of activities [Crouter et al., 2006]. This approach improved upon the methods using a single linear regression, but has also been shown to give inaccurate estimates of energy expenditure in a free-living setting [Lyden et al., 2011].

Lyden et al. [2014] use a decision tree to identify whether a time interval is an interval of activity or one of four types of inactivity by considering the number of seconds in the interval with non-zero counts and the duration of the interval. For the inactivity intervals, a MET value is assigned based on the Compendium of Physical Activities (CPA) [Ainsworth et al., 2011] and calibration studies. The CPA brings together scientific research that has established the energy expenditure required to perform a wide variety of physical activities. For activity intervals, METs are estimated using the neural network published earlier by Staudenmayer et al. [2009].

Pober et al. [2006] simply use a point estimate of the energy expenditure associated with each estimated activity class obtained from their classification model (described in Subsection 2.7.3). With this simple approach, they markedly improve the number of minutes correctly placed into the intensity categories in comparison with the approach of Freedson et al. [1998]. Bonomi et al. [2009b] use the same idea to estimate total daily energy expenditure, and find that the method works well. However, Albinali et al. [2010] examine the same method and find that it results in consistently low estimates of energy expenditure. They also consider using either a subject-specific regression model for each activity class to relate accelerometer output to METs, or a model for each activity class relating subject characteristics such as age, height, weight, and resting heart rate to energy expenditure. They find that both of these methods are better than using point estimates from the CPA, but the subject-specific method is impractical for many purposes since it requires measurements of energy expenditure for each subject to fit the regression models.

Another approach to estimating energy expenditure is to estimate the pace of activities such as walking and cycling; the intensity categorization of these activities is related to their speed [Ainsworth et al., 2011]. This idea has been used by Bonomi et al. [2009a]. They develop separate multiple regression models to estimate walking, running, and cycling speeds based on bodily characteristics and accelerometer features. A limitation of this approach is that the relationship

between movement speed and energy expenditure depends on the terrain. For example, walking at the same speed uphill and along flat ground entail different amounts of energy expenditure [Lester et al., 2009].

Lester et al. [2009] estimate energy expenditure using estimated activity type classifications along with subject weight, speed of movement, and the grade of movement. Detecting movement grade requires integration with other data sources or sensors such as a barometric pressure sensor or GPS unit. In their experiments, this system gives estimates of energy expenditure that are in the range of 79 to 89% of true energy expenditure.

CHAPTER 3

DATA

3.1 Introduction

We have recorded data about subjects' physical activity from two different studies: Mannini et al. [2013] and Sasaki [2013]. We describe the design of each of these studies and the data that were collected in Section 3.2. In Section 3.3, we describe the procedures we used in preprocessing the data. Finally, in Section 3.4 we present some plots and discuss aspects of the data that are important to consider in the model building process.

3.2 Study Methodologies

In this Section we describe the design of the studies where the data that we will use were gathered. Table 1 contains descriptive statistics for the participants in these studies.

In the study by Mannini et al. [2013], each participant performed a subset of 26 activities in the laboratory. These activities were designed to be generally representative of activities people engage in in real life, but subjects were not allowed to perform multi-tasking behaviors other

	Mannini et al. [2013]	Sasaki [2013] Laboratory	Sasaki [2013] Free Living
N	33	35	15
Male/Female	11/22	14/21	6/9
Age Range	18 to 75	65 to 80	65 to 78
Height (mean $\pm$ sd)	168.5 $\pm$ 9.3 cm	168.6 $\pm$ 9.8 cm	169.8 $\pm$ 9.8 cm
Weight (mean $\pm$ sd)	70.0 $\pm$ 15.6 kg	76.4 $\pm$ 14.2 kg	74.5 $\pm$ 11.4 kg

Table 1. Descriptive Statistics for Study Participants

than the walking-carrying-a-load activity. The order and duration of activities were determined by the researchers.

While they performed these activities, the subjects wore Wocket accelerometers on their ankle, thigh, wrist, and hip. Only the data from the ankle and the wrist have been made available to us. These accelerometers recorded acceleration (m/s^2) in each of 3 axes at a frequency of 90 Hz. Staff also recorded the start and stop times for each activity that was performed.

The study by Sasaki [2013] focused on the elderly and had two components: one where participants performed a prescribed set of activities in the laboratory, and one where the subjects were observed in their daily lives. For the laboratory component, each subject performed one of two activity routines consisting of a subset of 17 different activities. The subjects performed each activity for 5 minutes in a specified order, with a resting period of between 2 and 5 minutes between activities. As in Mannini et al. [2013], staff recorded the start and stop times of each activity. The subjects' specific behaviors during the resting periods were not labeled; from a visual examination of the accelerometer signals, it appears that they spent some time sedentary and some time moving during these time spans, but we do not have any way to confirm the specific activity types they engaged in. The subjects wore four accelerometers: ActiGraph GT3X+ accelerometers on the dominant ankle, hip, and wrist that recorded acceleration in 3 axes at a frequency of 80 Hz, and an activPAL on the leg that recorded a measure of acceleration along one axis at a frequency of 10 Hz in addition to a measure of the leg's orientation. We have used only the data from the ActiGraph accelerometers in our work. The subjects also wore an Oxycon portable respiratory gas exchange system which records breath-by-breath oxygen consumption and can be used to determine the intensity of physical activity [Rosdahl et al., 2010].

A subset of the subjects in the laboratory component of the study were recruited to participate in the free living component as well. These subjects wore the same four accelerometers that were used in the laboratory component of the study. Staff followed the subjects as they went about their normal activities and classified the activities in 3 ways: the type of activity performed (15 categories), the location of the activity (Indoors or Outdoors), and a categorical assessment of the intensity of the activity (Sedentary, Light, Moderate, or Vigorous).

3.3 Data Preprocessing

Rather than using the raw acceleration data directly in our models, we divide the signal up into windows several seconds in length and extract several summary statistics describing features of the signal within each window for use in the model. This general modeling approach has been used in every method for classification of physical activity using accelerometer signals that we are aware of, and has also been used successfully in other fields such as speech recognition.

We use non-overlapping windows of length 12.8 seconds. As we discussed in the literature review, it has generally been found that windows of length about 10 seconds or longer yield better classification results than shorter windows. We follow Mannini et al. [2013] and Zhang et al. [2012] in using windows of length 12.8 seconds. This particular window length is convenient because at a sampling frequency of 80 Hz each window contains 1024 observations, facilitating fast computations of frequency-domain features.

The data preprocessing procedure involves two primary tasks: (1) assign activity type and intensity labels to each time window, and (2) extract summary features from the accelerometer signal in each time window. We discuss each of these tasks in turn. We note that the data from Mannini et al. [2013] were already cleaned when they came to us. However, they performed many of the same steps that we did; we will describe differences between our preprocessing methods as appropriate.

In order to assign an activity label to each window, there are five subtasks:

1. adjust the timing of activity type and intensity changes recorded in the direct observation logs,
2. drop time with missing data,
3. compute the intensity category,
4. combine activity type categories to obtain a reduced set of type categories, and
5. assign category labels to windows.

It is necessary to adjust the transition times between different activity categories that were recorded in the direct observation logs in order to correct inaccuracies in the logging process. It takes some time for the direct observers to register that a subject has changed their activity type

and record that change in the logging software. In addition, there may be some misalignment and drift in the synchronization of the clocks used in the accelerometers and in the direct observation logging device. We adjusted the direct observation log times by visually aligning the transition times with the acceleration vector magnitude obtained from the ankle accelerometer. We did not notice any appreciable differences in the clocks between the different accelerometers; separate time adjustments for the specific accelerometer locations were not required.

The recorded transition times in the laboratory data were very accurate, and required only minimal adjustments. In the free living data, most of our changes to the logged transition times were between about 0 and 7 seconds. However, we did make a few larger adjustments to the transition times in the free living data, up to about 40 seconds. These larger adjustments were generally for transitions from the Moving Intermittently category to the Sedentary or Standing category; it seems that the direct observer took extra time to register that the subject was no longer moving intermittently in a few cases. These adjustments corrected many problems in the direct observation labels, but we believe that there may still be some inaccuracies in the activity type labels in the free living data. Although we adjusted the locations of existing transition times that were recorded in the direct observation logs, we did not introduce any new transitions.

The data from Sasaki [2013] contain several time periods where the activity type and intensity were not recorded. In the laboratory component of the study, this occurs during the time periods between the prescribed activities. In the free living component of the study subjects were able to request private time in order to use the restroom, change clothes, and perform similar activities. During these times we do not know what specific activities the subjects were engaged in. We have opted to drop these unlabeled time segments. Development of methods for partially labeled data would be interesting, and particularly useful in the context of physical activity data, but that is beyond the scope of our work.

Mannini et al. [2013] also dropped some times from the observation sequences for several reasons. One problem they encountered was data loss that occurred during wireless transmission of the acceleration signal from the accelerometers. If more than 20% of the samples in a given window of 12.8 seconds were missing, they discarded that time window and began a new window after the missing data. They used linear interpolation to fill in any remaining missing acceleration values. Rather than manually adjusting transition times to match the acceleration signal, they addressed the problem of imprecision in the recorded activity labels by discarding

one window of length 12.8 seconds before and after each recorded transition between different activity types. Finally, they used recordings from the accelerometer placed at the ankle to detect times that were mislabeled as ambulation: if the vector magnitude of the acclerometer signal had a standard deviation of less than 0.1g in any 2 second window, they discarded that time.

In Chapter 8, we will classify physical activity according to its intensity level. As we discussed in Section 3.2, only the free living data from Sasaki [2013] contain direct annotations of a categorical intensity level for each subject's activity over time. The laboratory data from Sasaki [2013] contain recordings of the subjects' breath-by-breath oxygen consumption from an Oxycon portable respiratory gas exchange system. We processed these data to obtain the subjects' steady state energy expenditure during minutes 3 - 5 of each activity in units of METs. In calculating METs, we used a denominator of 3.5 ml $O_2 \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ for the reference metabolic rate in the resting state. Use of this population average value can be criticized, particularly with a pool of elderly subjects as we have in this study [Kozey et al., 2010]. However, we were missing resting metabolic rate measurements for two of the subjects. We opted to use a consistent measure that was available for all of our subjects. In general, it takes a few minutes after beginning an activity for oxygen consumption to reach a steady state. For data gathered in the laboratory where the intensity of each activity remained fairly constant throughout the activity, it is appropriate to use the steady state oxygen consumption to represent the intensity of the activity throughout the entire duration of the activity. After computing energy expenditure in METs, we assigned an intensity category to each time point using the standard categories and MET cutoffs [Ainsworth et al., 2011]: Sedentary (≤ 1.5 METs), Light (> 1.5 and < 3 METs), Moderate (≥ 3 and < 6 METs), and Vigorous (≥ 6 METs).

The data from Mannini et al. [2013] do not contain any direct information about intensity. However, they do contain a detailed system of physical activity type categorization, breaking activity type down into 26 different classes. We combined this detailed information about activity type with the population-average MET values associated with each activity category that are available in the Compendium of Physical Activities [Ainsworth et al., 2011] to obtain approximate intensity levels. We then converted these continuous values to categorical intensity levels using the set of MET cutoffs described above.

Both of the studies by Mannini et al. [2013] and Sasaki [2013] made use of reduced set of activity type categories for modeling purposes, formed by merging the finer categories together.

We will use the same category groupings used in those studies. Mannini et al. [2013] used the following category groupings:

1. Sedentary (lying, sitting, internet search, reading, typing, writing, sorting files, riding an elevator)
2. Ambulation (natural walking, treadmill walking, carrying a box, stairs up/down),
3. Cycling (indoor and outdoor), and
4. Other Activities (sweeping with a broom, painting with a roller or brush).

Sasaki [2013] used two different groupings of activities with either 3 or 5 categories, which varied slightly between the laboratory and free living settings. In the laboratory setting with 5 categories, the groups were formed as follows:

1. Sedentary (lying down, sitting, crossword puzzles, playing cards)
2. Standing (standing)
3. Locomotion (slow walk, 400m walk, carrying groceries)
4. Moving Intermittently (dusting, gardening, vacuuming, self-care, laundry, organizing the room)
5. Recreational (tai-chi, simulated bowling)

The system with 3 categories was formed by merging the Standing group with the Sedentary group and the Recreational Group with the Moving Intermittently group. For the free living setting with 5 categories, the groups were formed as follows:

1. Sedentary (lying down, sitting, sitting with upper body movement, driving)
2. Standing (standing, standing with upper body movement)
3. Locomotion (walking, walking with a load, walking on an incline, stairs)
4. Moving Intermittently (moving intermittently, household activities)
5. Recreational (aerobic exercise, resistance exercise, balance exercise)

Again, the classification system with 3 categories was formed by merging the Standing group with the Sedentary group and the Recreational Group with the Moving Intermittently group.

The final step in the labeling task is assigning activity type and intensity labels to each window. The issue here is how to address the fact that some windows may cover more than one activity category. As we mentioned above, Mannini et al. [2013] discarded one window before and after each transition between different activity types, so this problem does not arise in their data. In processing the data from Sasaki [2013], we have opted to keep these transition times. In order to handle transitions between different activity types, we introduced a new Transition category, which we assigned to windows that contained more than one labeled activity type in the reduced classification system.

Figures 4 through 7 show how the data cleaning process we have described here impacts the activity class labels for a typical subject in the laboratory and free living components of the study from Sasaki [2013]. In Figure 4, we see that the original laboratory data contain annotations of 8 different activity categories, along with several segments of unlabeled activity. Figure 5 shows that after preprocessing, the unlabeled time segments have been dropped, the activity type categories have been merged into a reduced set of categories, and windows overlapping two different activity types have been reclassified with the Transition label. With the laboratory data, only minimal adjustments were required in terms of the start and stop times for each activity. Figure 6 shows the originally recorded data for subject 9 in the free living component of the study. The preprocessed data are shown in Figure 7. Here, the transition times were adjusted, time segments labeled as Private activity have been removed, some of the categories from the original direct observation logs have been merged together, and windows covering more than one category in the direct observation logs have been relabeled with the Transition class.

The second stage of preprocessing the data involves extracting a vector of features from the accelerometer signal in each window. In their study, Mannini et al. [2013] considered several different feature sets, and selected one set of 13 features which they felt achieved an optimal balance between the time required to compute the features and the resulting classification performance. In our work with their data set, we use that feature set in order to facilitate comparisons between the effectiveness of our methods and the support vector machine they employed as a classifier. These features are listed in Table 2. They consist of time and frequency domain summaries of the vector magnitude of the accelerometer signal. In our applications to the data from Sasaki [2013],

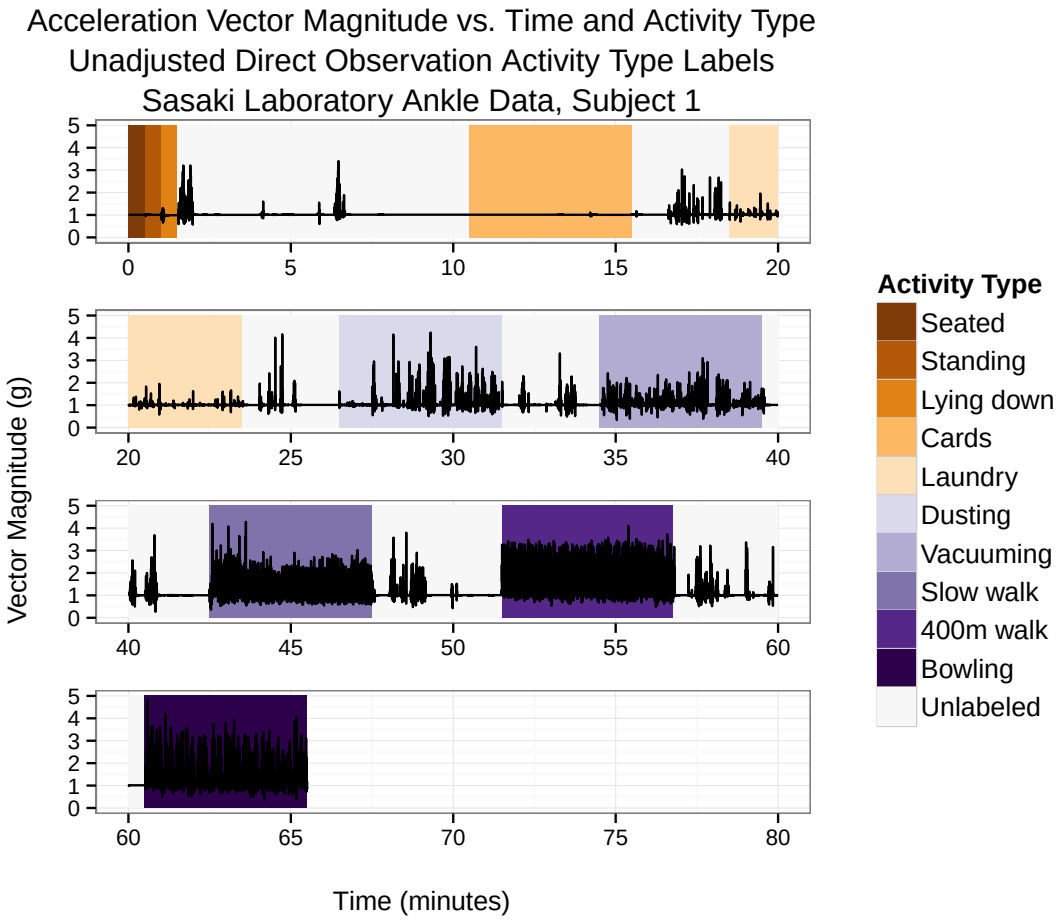


Figure 4. Plot of the acceleration vector magnitude over time for subject 1 in the laboratory component of the study by Sasaki [2013]. The background color indicates the originally recorded activity classification at each point in time before preprocessing.

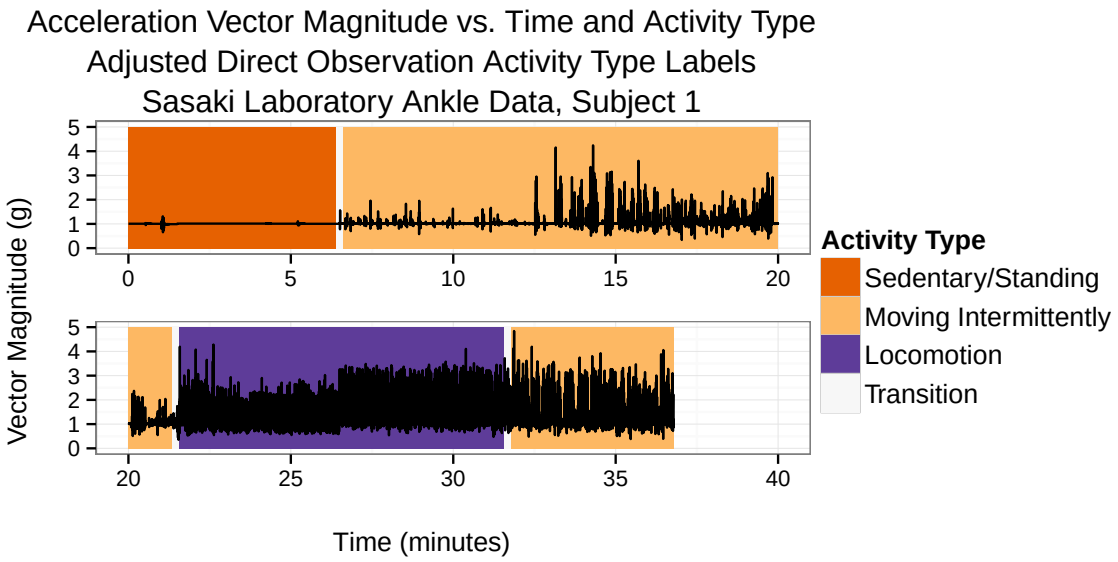


Figure 5. Plot of the acceleration vector magnitude over time for subject 1 in the laboratory component of the study by Sasaki [2013]. The background color indicates the adjusted activity classification at each point in time after preprocessing.

we use an expanded set of 77 features, listed in Table 3. We computed these features using the three individual axes of acceleration recordings and the vector magnitude, polar angle, and azimuth angle in a representation of the signal in spherical coordinates. Our choices of features are similar to those used in previous studies such as Mannini et al. [2013].

3.4 Plots and Discussion

In this Section we discuss some initial plots of the data from Mannini et al. [2013] and make two main observations: (1) there is temporal dependence in the activity type and intensity level, and (2) the features extracted from the acceleration data do not follow any simple parametric distribution. These properties of the data have important implications for our models: in order to represent the association between the accelerometer signal and the activity classification well, they should take account of the temporal dependence, and they should be flexible enough to handle features that have complex distributions. We will also demonstrate that there is a fair amount of variability between subjects in terms of both the patterns of the accelerometer signal that are associated with each activity category and the amount of time spent in each activity category in the free living setting. Ideally, our models would account for this subject-specific

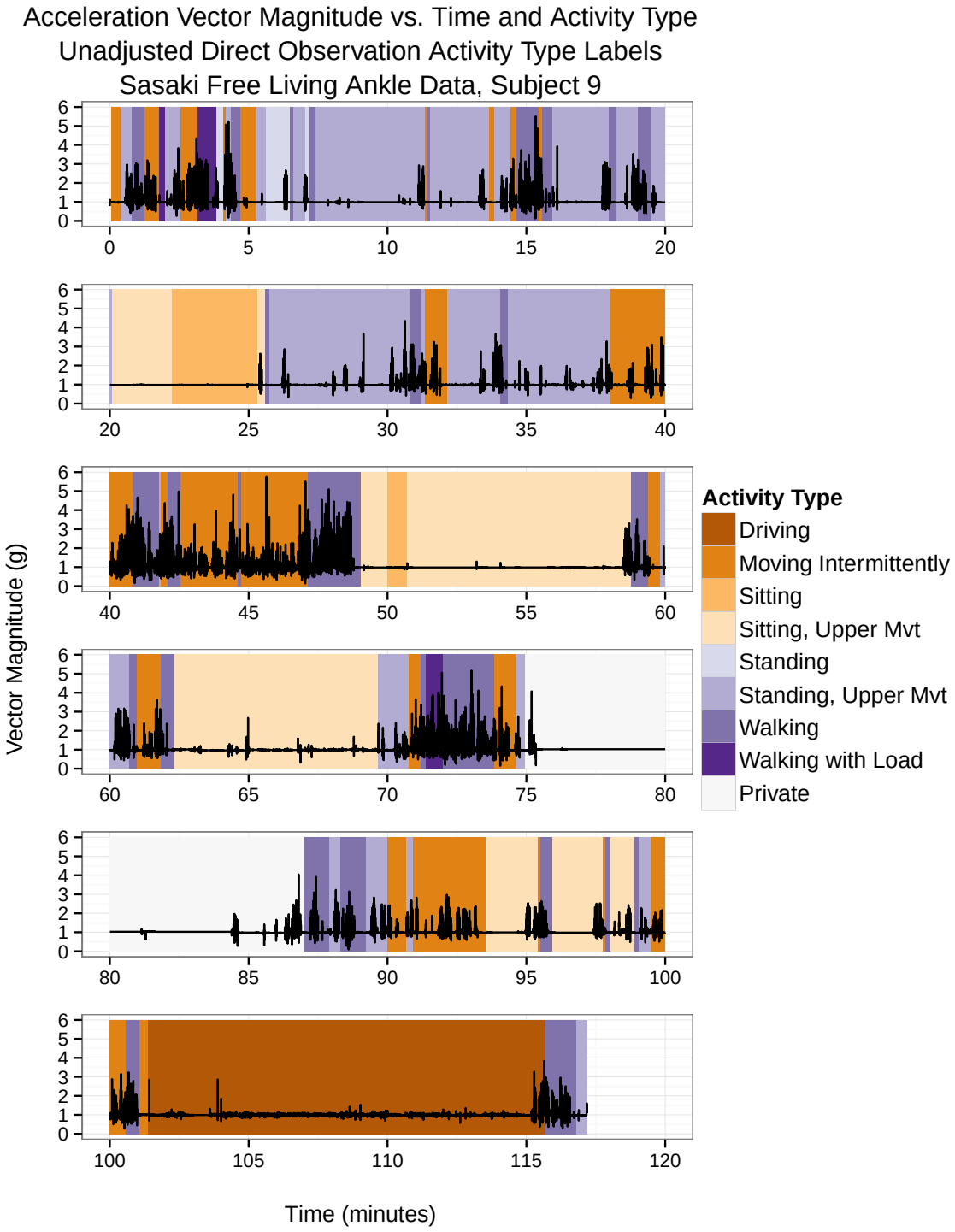


Figure 6. Plot of the acceleration vector magnitude over time for subject 9 in the free living component of the study by Sasaki [2013]. The background color indicates the originally recorded activity classification at each point in time before preprocessing.

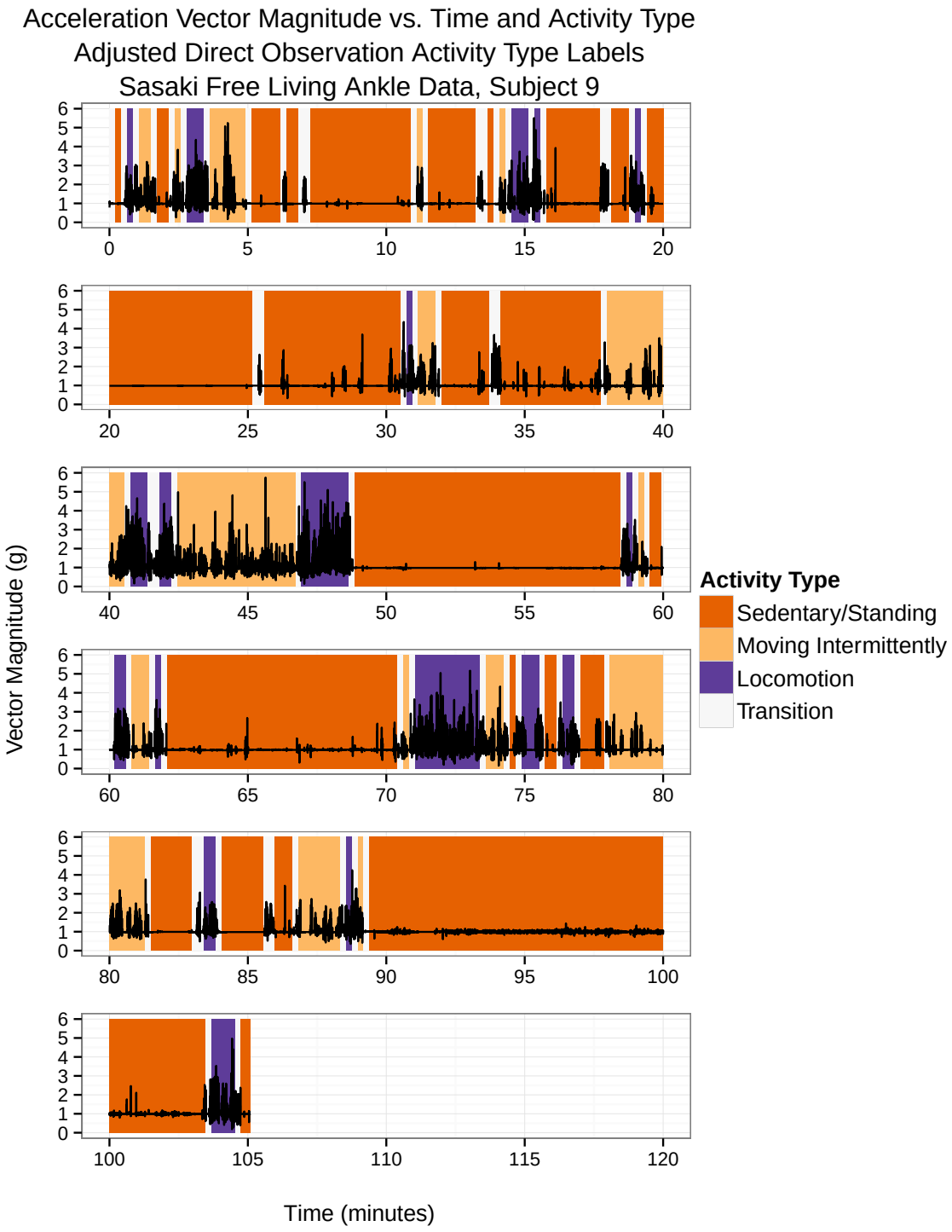


Figure 7. Plot of the acceleration vector magnitude over time for subject 9 in the free living component of the study by Sasaki [2013]. The background color indicates the adjusted activity classification at each point in time after preprocessing. Note that the time labeled as Private beginning at about minute 75 in Figure 6 has been removed.

Domain	Feature
Time Domain	Mean
	Standard deviation
	Minimum and maximum
Frequency Domain	Frequency and power of the first dominant frequency between 0.3 Hz and 15 Hz
	Frequency and power of the second dominant frequency between 0.3 Hz and 15 Hz
	Total power between 0.3 Hz and 15 Hz
	Ratio of the power of the first dominant frequency between 0.3 Hz and 15 Hz and the total power between 0.3 Hz and 15 Hz
	Frequency and power of the first dominant frequency between 0.3 Hz and 3 Hz
	Ratio of the frequency of the first dominant frequency between 0.3 Hz and 15 Hz in the current window and in the previous window

Table 2. Features extracted from the accelerometer signal in preprocessing the data from Mannini et al. [2013]. All features are computed using the acceleration vector magnitude.

variation as well; however, we will not pursue that line in this work. We will discuss this issue further in Chapters 7 and 9.

Our first observation, that there is temporal dependence in the activity type and intensity level, can be seen from the plots in Figures 4 through 7. The key point here is that if an individual is sedentary at one point in time then he or she is likely to remain sedentary in the near future, and similarly for the other activity classes. Thus, the activity type at a given time point is informative about the activity type at nearby time points.

Our second observation is that the features follow relatively complex distributions. This can be seen from the plots displayed in Figure 8. Note that these plots show only univariate and bivariate summaries of these distributions; the higher dimensional densities are even more complicated. It would be difficult to specify a simple parametric model for these data that modeled their distribution accurately. At the same time, the relatively large number of features we have extracted means that flexible density estimation methods such as kernel density estimation will struggle. This is a central challenge that we will seek to address in specifying our model.

Our final observation is that there is a fair amount of variation between subjects. We show this for the one pair of accelerometer features in Figure 9. This plot can be interpreted as indicat-

Domain	Feature	X	Y	Z	VM	θ	ρ
Time Domain	Mean	Y	Y	Y	Y	Y	Y
	Order Statistics: The 10th, 25th, 50th, 75th, and 90th percentiles	Y	Y	Y	Y	Y	Y
	Lag 1 autocorrelation	Y	Y	Y	Y	N	N
	Entropy: We first obtain a nonparametric estimate of the distribution of acceleration values by placing the observed VM values into 10 bins of equal size and calculating the proportion falling into each bin, $p_1, \ldots, p_{10}$. The estimated entropy is then $-\frac{1}{10}\sum_{i=1}^{10} p_i log(p_i)$	N	N	N	Y	N	N
Frequency Domain	Frequency and power of the first dominant frequency: The frequency and estimated power for the frequency with the highest estimated power	Y	Y	Y	Y	N	N
	Frequency and power of the second dominant frequency: Same as above, for the frequency with the second-highest estimated power	Y	Y	Y	Y	N	N
	Total power: The sum of the estimated power for all frequencies. Note that this is equal to $\frac{T-1}{2}$ times the sample variance of the observations if T is odd, with a slight adjustment if T is even.	Y	Y	Y	Y	N	N
	Frequency and power of the first dominant frequency in the band from 0.3 to 3 Hz	Y	Y	Y	Y	N	N
	Ratio of power of first dominant frequency in the band from 0.3 to 3Hz to power of first dominant frequency overall	Y	Y	Y	Y	N	N
	Entropy of the spectral density: After normalizing the estimated powers so that they sum to 1, we apply the entropy calculation above. This is a measure of how uniformly "distributed" the variance is among the frequencies considered.	Y	Y	Y	Y	N	N

Table 3. Features extracted from the accelerometer signal in preprocessing the data from Sasaki [2013]. The right-hand 6 columns indicate whether the listed feature was computed for the anteroposterior axis, mediolateral axis, vertical axis, vector magnitude, polar angle, and azimuthal angle.

ing that the distribution of the accelerometer features associated with each activity type varies between subjects. Alternatively, the decision boundaries between classes in the feature space are subject-dependent. In the free living data, there are also differences in the relative amounts of time the subjects spend in each activity category, and the probabilities of transitioning from one activity type to another. We depict these differences among the subjects in Figure 10. This variation among subjects represents an important aspect of the data that we will not account for in our models; we will discuss our reasons for this in Chapter 7. However, it is important to be aware of this between-subject variability.

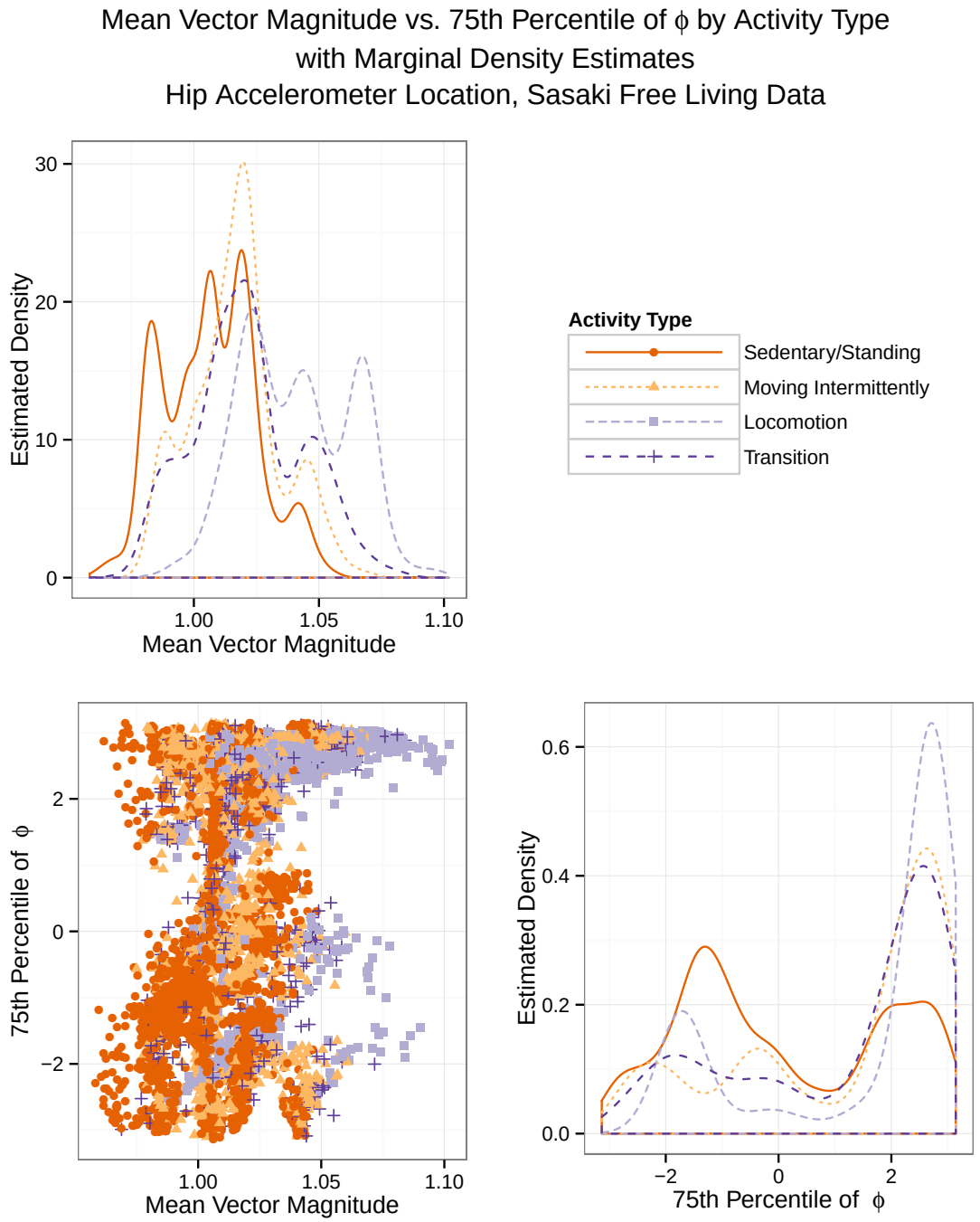


Figure 8. Plots showing the observed values and univariate kernel density estimates for one pair of features with the hip data from the free living component of the study by Sasaki [2013]. In the lower left panel, each point represents one window of length 12.8 seconds. The horizontal axis of the plot gives the average value of the acceleration vector magnitude within the given window. The vector magnitude includes acceleration due to gravity, so if a subject is stationary the vector magnitude is 1 g. The vertical axis of the plot gives the 75th percentile of the azimuthal angle, indicating the relative amounts of acceleration experienced by the accelerometer along the anteroposterior axis and the mediolateral axis.

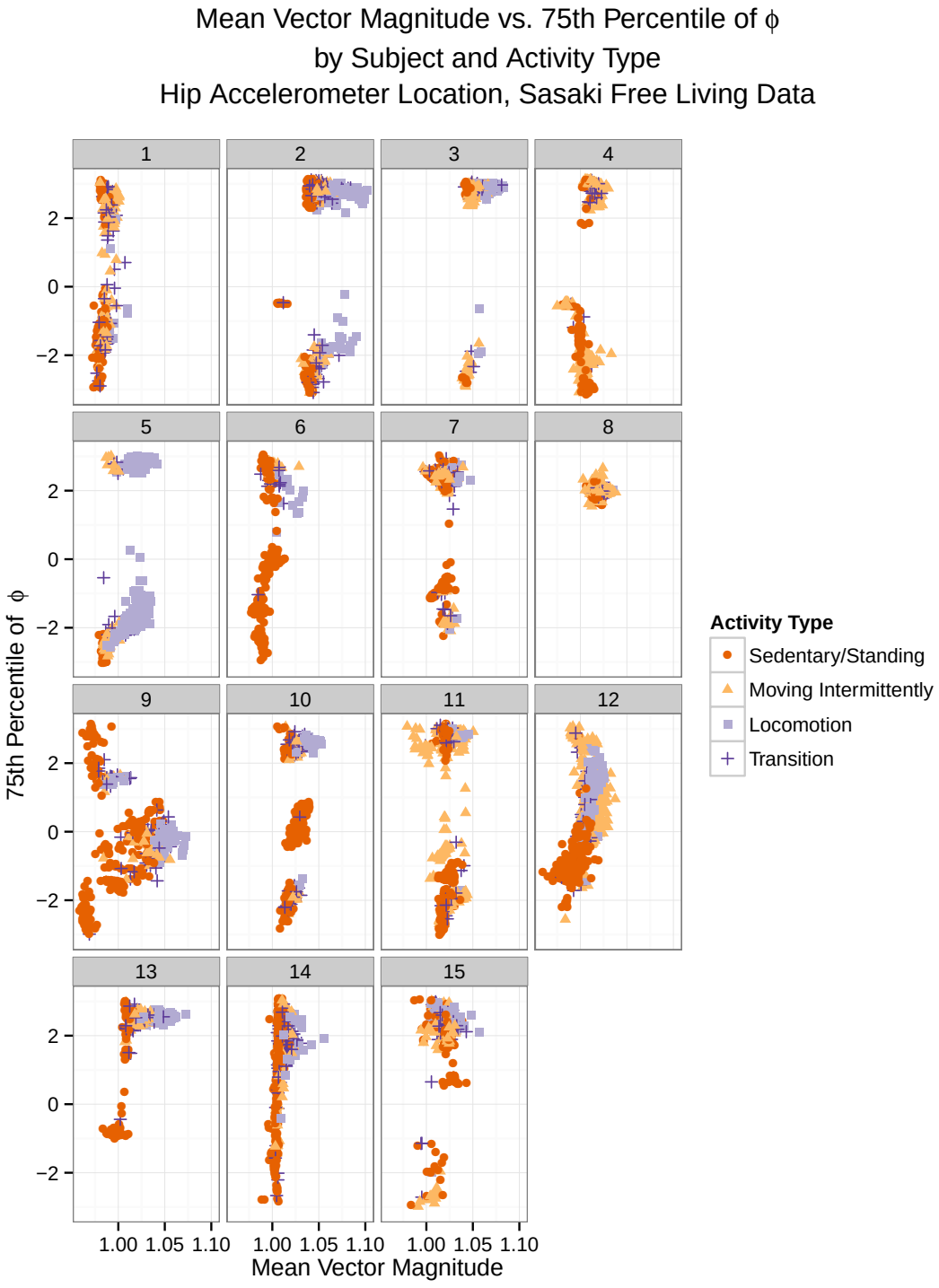


Figure 9. Plots showing the observed values for one pair of features in the hip data from the free living component of the study by Sasaki [2013]. Each point represents one window of length 12.8 seconds. Each panel shows the data for one subject. The features are as described in Figure 8.

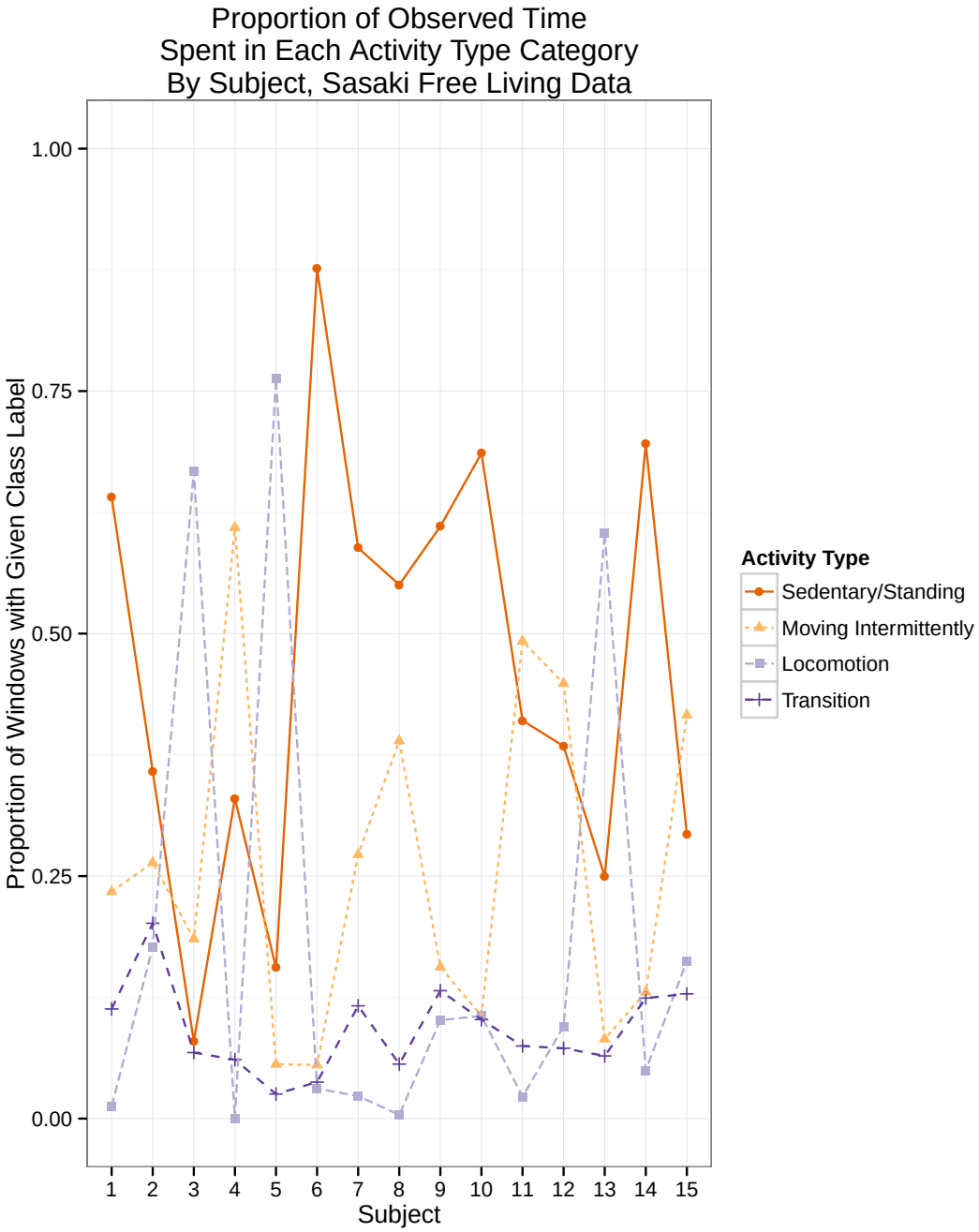


Figure 10. The proportion of time each subject spent in each activity category in the free living component of the study by Sasaki [2013].

CHAPTER 4

PRELIMINARY SIMULATION STUDIES

In this Section we conduct two preliminary simulation studies to motivate the models that we develop in Chapter 5. In Section 4.1, we demonstrate that models that account for temporal dependence are helpful when the data-generating mechanism includes temporal dependence. In Section 4.2, we show that discriminative methods can be superior to generative methods when the observation distributions associated with each state are not modeled well in the generative model. Variations on these results have already been proved or demonstrated through applications to real data sets in many different settings. Our simulation studies serve to confirm these prior results, as well as to offer specific insights related to our area of application.

4.1 Preliminary Simulation 1: Temporal Dependence

As we have seen in Chapter 3, physical activity data exhibit temporal dependence. In this Section, we describe a simulation study that demonstrates that when the data are temporally dependent, it is beneficial to account for that dependence in a classification model. Specifically, we show that when the data are generated from a HMM, HMMs yield higher classification accuracy than FMMs. We begin by describing the simulation study, then we present and discuss the results.

Our simulation study works with a simplified data generating mechanism so that we can focus on the effects of temporal dependence on classification accuracy. As we noted in Chapter 3, the strongest indication of temporal dependence in physical activity data is that individuals tend to engage in activity types for extended times. We capture this in our simulation study by simulating activity types from a Markov chain that has large entries on the diagonal of the

transition matrix. That is, the probability of transitioning from a given activity back to the same activity is large, and the probability of transitioning to a different activity is small.

The simulation study was conducted as follows. In each of 1000 simulations, we generate independent training and test data sets from the following model with $N = 5$ and $T = 1000$:

$Y_{i,t} \in \mathcal{S} = \{1, 2\}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$

$P(Y_{i,1} = 1) = P(Y_{i,1} = 2) = \frac{1}{2}$, $s \in \mathcal{S}$, $i = 1, \ldots, N$

$P(Y_{i,t} = s | Y_{i,t-1} = r) = q_{r,s}$, $r, s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 2, \ldots, T$, where

$$q_{r,s} = \begin{cases} 0.95 \text{ if } r = s, \\ 0.05 \text{ if } r \neq s \end{cases}$$

$X_{i,t} | Y_{i,t} = s \sim N(\mu_s, 1)$ $s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where

$$\mu_s = \begin{cases} -0.5 \text{ if } s = 1, \\ 0.5 \text{ if } s = 2 \end{cases}$$

The variable $Y_{i,t}$ corresponds to the activity type for subject i at window t, and $X_{i,t}$ corresponds to an observation made from an accelerometer for subject i at window t.

We use the training data set to estimate the parameters of the following two models:

Model 1 – FMM:

$P(Y_{i,t} = s; \boldsymbol{\pi}) = \pi_s$, $s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where

$0 \leq \pi_1, \pi_2 \leq 1$ and $\pi_1 + \pi_2 = 1$

$X_{i,t} | Y_{i,t} = s \sim N(\mu_s, 1)$, $s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$

The FMM has parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mu_1, \mu_2)$, where $\boldsymbol{\pi} = (\pi_1, \pi_2)$. These parameters are estimated using maximum likelihood. The estimate of π_s is the observed proportion of the training set with $y_{i,t} = s$, and the estimate of μ_s is the sample mean of the observed $x_{i,t}$ such that $y_{i,t} = s$.

Model 2 – HMM:

$P(Y_{i,1} = s; \boldsymbol{\pi}) = \pi_s$, $s \in \mathcal{S}$, $i = 1, \ldots, N$, where

$0 \leq \pi_1, \pi_2 \leq 1$ and $\pi_1 + \pi_2 = 1$

$P(Y_{i,t} = s | Y_{i,t-1} = r; Q) = q_{r,s}$, $r, s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where

$0 \leq q_{r,s} \leq 1 \, \forall r, s$ and $\sum_{s=1}^{2} q_{r,s} = 1 \, \forall r$

$X_{i,t} | Y_{i,t} = s \sim N(\mu_s, 1)$, $s \in \mathcal{S}$

The HMM has parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, Q, \mu_1, \mu_2)$, where $\boldsymbol{\pi} = (\pi_1, \pi_2)$ and $Q = [q_{r,s}]$. We use maximum likelihood estimates of $Q = [q_{r,s}]$, μ_1, and μ_2. The estimate of $q_{r,s}$ is

$$\frac{n_{r,s}}{n_{r,1} + n_{r,2}},$$

where $n_{r,s}$ is the number of observed transitions beginning in state r and ending in state s. The estimate of μ_s is the same as in the FMM. Also as in the FMM, we estimate $\boldsymbol{\pi}_s$ as the observed proportion of the sample with $y_{i,t} = s$; this is not the maximum likelihood estimate, but use of this estimate is a standard procedure to reduce sampling variance of the estimates.

Finally, we use the estimated model parameters to obtain a predicted value for each $y_{i,t}$ in the test data set. These predictions are made by using Bayes' rule to compute estimated class membership probabilities at each time point, and then selecting the class with highest estimated probability. This choice minimizes the 0-1 loss.

We now present the results of this simulation study. Figure 11 summarizes the overall proportion correct for the FMM and the HMM in each iteration of the simulation study. Tables 4 and 5 give confidence intervals for the proportion correct in each case as well as for the difference between these proportions. We see that the classification rates are much higher for the HMM than for the FMM.

Figure 12 shows how this higher classification rate is achieved. Whereas the FMM enforces a strict boundary between the classes in the space of values of $X_{i,t}$, the HMM allows this boundary to be crossed if nearby observations are likely to have come from one class or the other. This explains the horizontal bands of colors in the third column of the plot, where temporally adjacent observations are assigned to the same class.

These results suggest that when classifying physical activity that include temporal dependence, the statistical model used benefits from accounting for that dependence. We note that

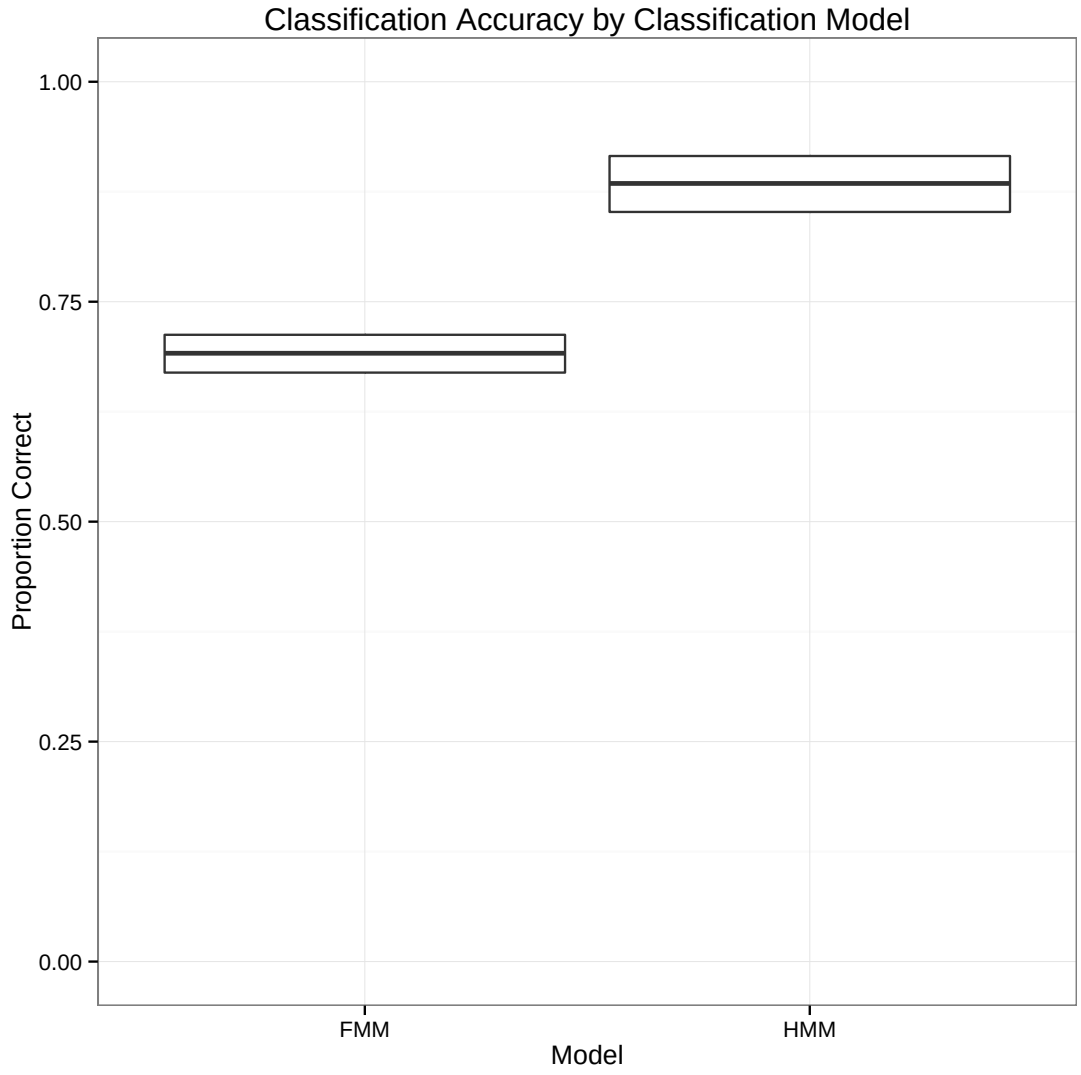


Figure 11. Preliminary Simulation 1: Boxplots showing the minimum, median, and maximum classification rates achieved by the FMM and HMM in 1000 simulations.

	Proportion Correct ($\widehat{p}$, approximate 95% CI)
FMM	0.6914 [0.6910, 0.6918]
HMM	0.8842 [0.8840, 0.8845]

Table 4. Preliminary Simulation 1: Estimated proportion correct for each model, with approximate confidence intervals. The confidence intervals are based on the normal approximation ignoring serial correlation, with a sample size of $N_{sims} \times N \times T = 1000 \times 5 \times 1000$.

	Difference in Proportion Correct ($\widehat{p}_{HMM} - \widehat{p}_{FMM}$, approximate 95% CI)
HMM - FMM	0.1928 [0.1923, 0.1934]

Table 5. Preliminary Simulation 1: Estimated difference in proportion correct for each model, with approximate confidence interval. The confidence interval is based on a paired t test, with a sample size of $N_{sims} = 1000$.

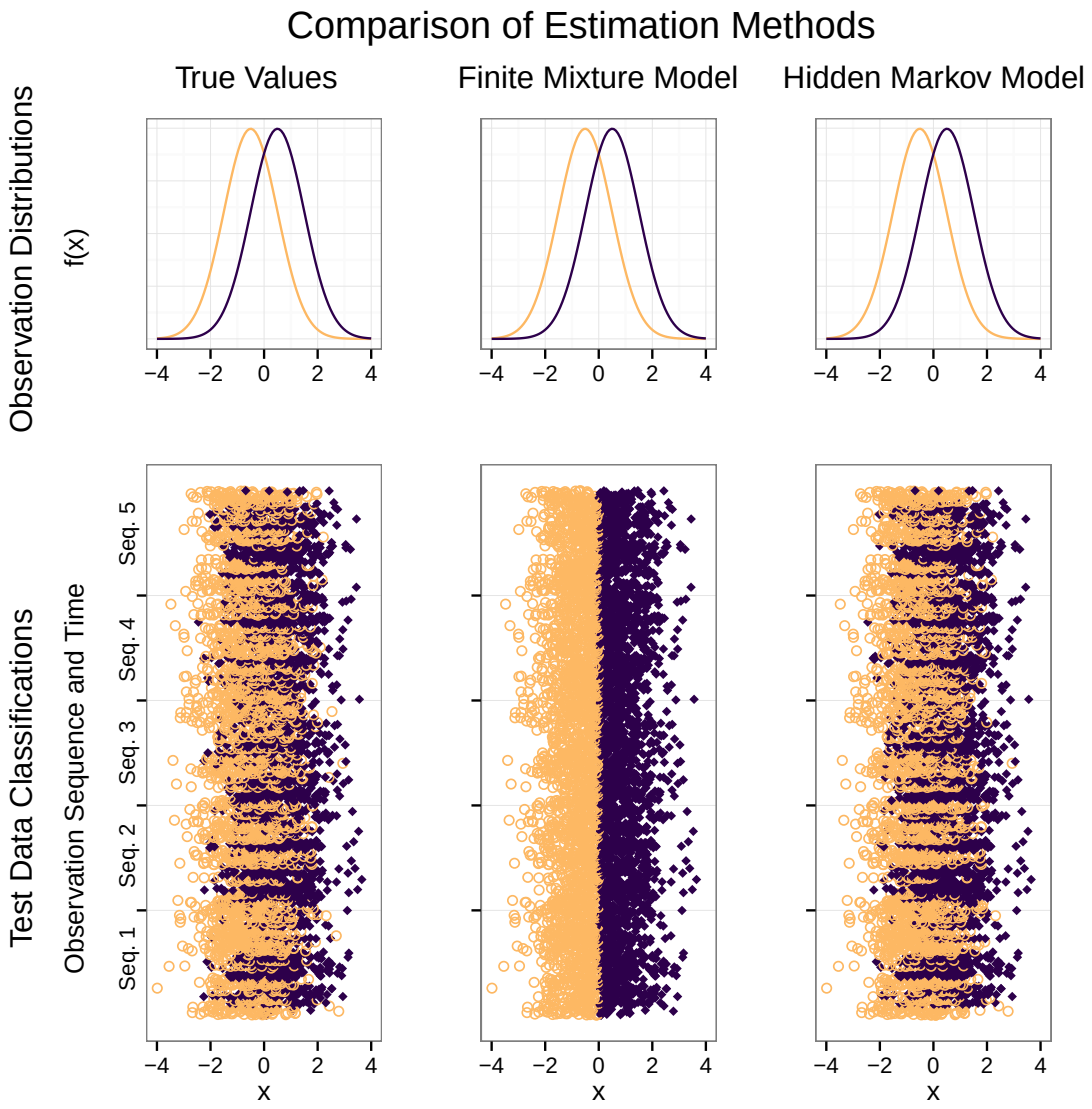


Figure 12. Preliminary Simulation 1: A detailed look at the results from one simulation. The top left panel shows the true observation distributions used in generating the data, and the lower left panel shows the true $x_{i,t}$ values and classifications for the simulated test data set. The center and right panels show the estimated observation distributions and predicted classes for the test data obtained from the FMM and HMM, respectively.

this simulation study has examined only the most obvious form of temporal dependence found in physical activity data, the high chance that an individual will continue doing their present activity type in the near future rather than switching to a new activity type. In fact, physical activity data also exhibit many other forms of temporal dependence. For example, some transitions between different activity types are more likely than others, and the dependence may have higher order dynamics. Classification models would probably also benefit from being able to model these forms of dependency.

4.2 Preliminary Simulation 2: Generative Model Misspecification

A second important characteristic of the physical activity data we displayed in Section 3.4 is that the distribution of feature values associated with each activity type is complex. It would be difficult to specify an accurate parametric model for these distributions, and flexible non-parametric approaches such as kernel density estimation perform poorly in the moderate to high dimensions present in these data. A common middle path used with generative HMMs approximates the observation distributions with mixtures of normal distributions. This can lead to an arbitrarily good approximation of the true density as the number of components in the mixture increases, but in practice with a finite sample size this approach entails some level of model misspecification. In this Section, we present a simulation study illustrating the fact that if the observation distribution is misspecified in the generative model, discriminative methods can provide superior classification performance.

The simulation study was conducted as follows. In each of 1000 simulations, we generate

independent training and test data sets from the following model with $N = 5$ and $T = 1000$:

$Y_{i,t} \in \mathcal{S} = \{1, 2\}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$

$P(Y_{i,1} = 1) = P(Y_{i,1} = 2) = \frac{1}{2}$, $s \in \mathcal{S}$, $i = 1, \ldots, N$

$P(Y_{i,t} = s | Y_{i,t-1} = r) = q_{r,s}$, $r, s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 2, \ldots, T$, where

$$q_{r,s} = \begin{cases} 0.95 \text{ if } r = s, \\ 0.05 \text{ if } r \neq s \end{cases}$$

$X_{i,t} | Y_{i,t} = 1 \sim N(0, 1)$

$X_{i,t} | Y_{i,t} = 2 \sim Exp(0.5)$

We use the training data set to estimate the parameters of the following two models:

Model 1 – HMM:

$P(Y_{i,1} = s; \boldsymbol{\pi}) = \pi_s$, $s \in \mathcal{S}$, $i = 1, \ldots, N$, where

$0 \leq \pi_1, \pi_2 \leq 1$ and $\pi_1 + \pi_2 = 1$

$P(Y_{i,t} = s | Y_{i,t-1} = r; Q) = q_{r,s}$, $r, s \in \mathcal{S}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where

$0 \leq q_{r,s} \leq 1 \, \forall r, s$ and $\sum_{s=1}^{2} q_{r,s} = 1 \, \forall r$

$X_{i,t} | Y_{i,t} = s \sim N(\mu_s, \sigma^2)$, $s \in \mathcal{S}$

This is the same model that was used in the simulation study of Section 4.1, except that we now treat the common variance σ^2 as an unknown parameter rather than fixing it to be equal to 1. As before, we estimate the model parameters via maximum likelihood, aside from the distribution of the initial state.

Model 2 – CRF:

$$P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i; \boldsymbol{\theta}) = \frac{\pi_{y_{i,1}} \prod_{t=2}^{T_i} q_{y_{i,t-1}, y_{i,t}} \prod_{t=1}^{T_i} exp(\beta_{y_{i,t},0} + \beta_{y_{i,t},1} x_{i,t})}{\sum_{\mathbf{y}_i^*} \pi_{y_{i,1}^*} \prod_{t=2}^{T_i} q_{y_{i,t-1}^*, y_{i,t}^*} \prod_{t=1}^{T_i} exp(\beta_{y_{i,t}^*,0} + \beta_{y_{i,t}^*,1} x_{i,t})}$$

We fix $\beta_{2,0} = \beta_{2,1} = 0$ so that the parameters are identifiable. As illustrated in Sutton and McCallum [2011], this model can be obtained from the HMM above by conditioning on $\mathbf{X}_i$. We estimate the parameters via maximum likelihood using numerical optimization methods. As with the HMM, in order to estimate the class memberships in the test data, we calculate the marginal class membership probabilities $P(Y_{i,t} = s | \mathbf{x}_i; \widehat{\boldsymbol{\theta}})$ and select the class with the highest probability.

	Proportion Correct ($\widehat{p}$, approximate 95% CI)
HMM	0.9211 [0.9208, 0.9213]
CRF	0.9415 [0.9413, 0.9417]

Table 6. Preliminary Simulation 2: Estimated proportion correct for each model, with approximate confidence intervals. The confidence intervals are based on the normal approximation ignoring serial correlation, with a sample size of $N_{sims} \times N \times T = 1000 \times 5 \times 1000$.

	Difference in Proportion Correct ($\widehat{p}_{CRF} - \widehat{p}_{HMM}$, approximate 95% CI)
CRF - HMM	0.0204 [0.0200, 0.0208]

Table 7. Preliminary Simulation 2: Estimated difference in proportion correct for each model, with approximate confidence interval. The confidence interval is based on a paired t test, with a sample size of $N_{sims} = 1000$.

Figure 13 and Tables 6 and 7 summarize the results of this simulation. We see here that on average, the classification accuracy of the discriminative model is about 2% higher than that of the generative model. As we discussed in the literature review, the difference in performance between the generative and discriminative approaches depends on relative quality of the generative and discriminative models. The simulation study we conducted in this Subsection is not an exhaustive study of this behavior; instead, it serves to motivate our modeling approach by giving a simple example demonstrating that discriminative methods can offer better performance than generative methods in some settings. We will conduct a more thorough simulation study in Chapter 6 and apply several methods to physical activity data in Chapters 7 and 8.

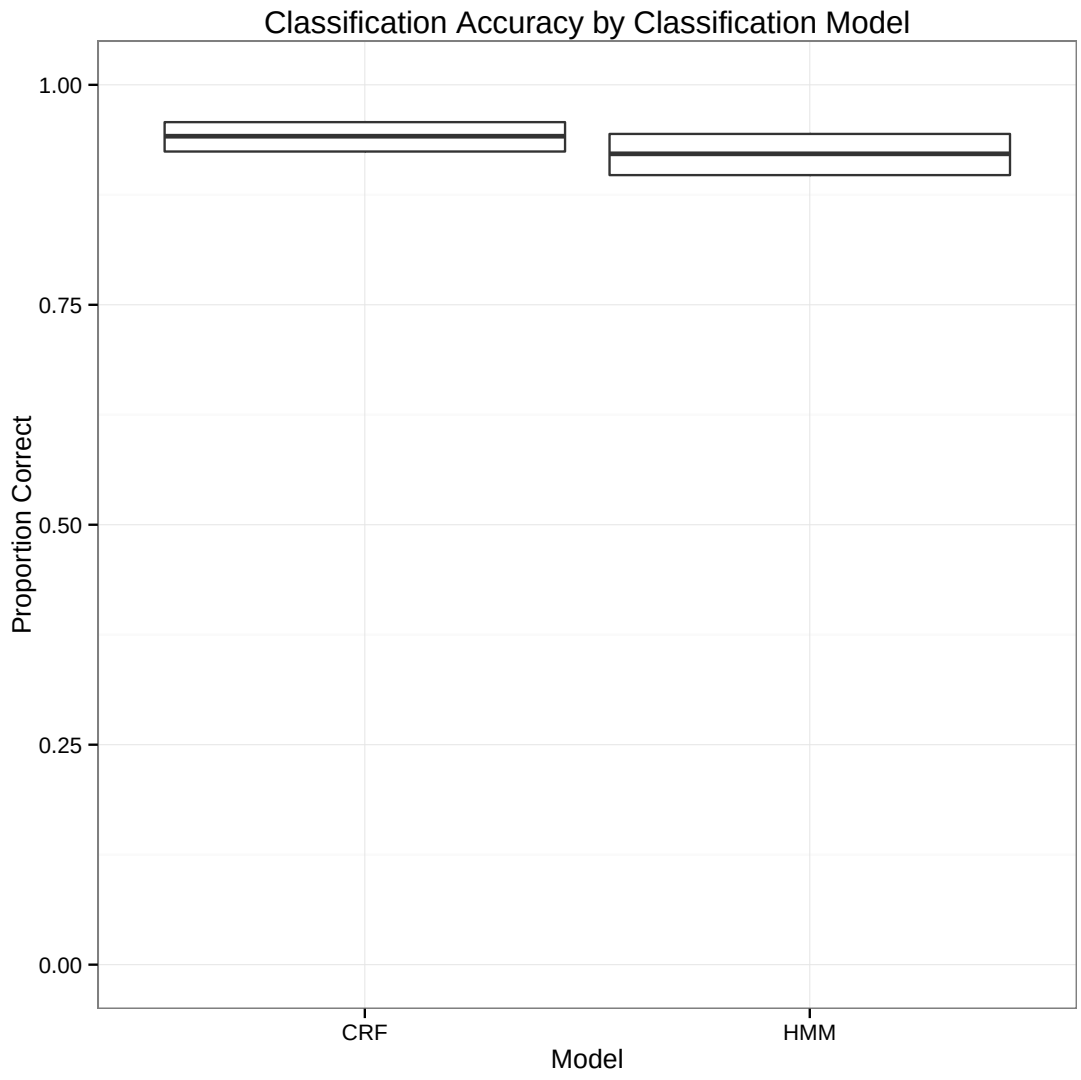


Figure 13. Preliminary Simulation 2: Boxplots showing the minimum, median, and maximum classification rates achieved by the CRF and HMM in 1000 simulations.

CHAPTER 5

CONDITIONAL RANDOM FIELD MODELS

In this Chapter we present three novel estimation strategies for the linear chain CRF model. We begin with a general introduction to our methods and a discussion of the elements they have in common in Section 5.1. We then discuss the methods in more detail in Sections 5.2, 5.3, and 5.4. In Section 5.5, we discuss how we perform classification given values for the model parameters. In Section 5.6, we review algorithms that have been developed in the literature for fast computation of the model likelihood and its gradient. These computational methods underlie Sections 5.2 through 5.5.

5.1 Introduction

The methods we present here are all based on the linear chain CRF. As we discussed in the literature review, the CRF specifies the conditional distribution of $\mathbf{Y}|\mathbf{X}$ using a graphical model. Formally, the model uses a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the conditional independence relationships among the variables $Y_{i,t}$ given $\mathbf{X} = \mathbf{x}$. The vertices of the graph are the components of the vector $\mathbf{Y}$ and the edges specify the conditional dependence structure of $\mathbf{Y}|\mathbf{X}$ as follows:

$$p(Y_{i,t}|\mathbf{X}, Y_{i^*,t^*}, (i,t) \neq (i^*,t^*)) = p(Y_{i,t}|\mathbf{X}, Y_{i^*,t^*}, Y_{i,t} \sim Y_{i^*,t^*}),$$

where $Y_{i,t} \sim Y_{i^*,t^*}$ indicates that the graph contains an edge connecting $Y_{i,t}$ and Y_{i^*,t^*} [Lafferty et al., 2001]. In words, $\mathbf{Y}|\mathbf{X}$ has the Markov property with respect to the graph: given $\mathbf{x}$ and the graph nodes that $Y_{i,t}$ is connected to, $Y_{i,t}$ is conditionally independent of all of the nodes it is not connected to. Although we condition on the realized values $\mathbf{x}$ and do not model their distribution, it is conceptually helpful to include the $\mathbf{x}_{i,t}$ in the graph to show that $Y_{i,t}$ depends on them.

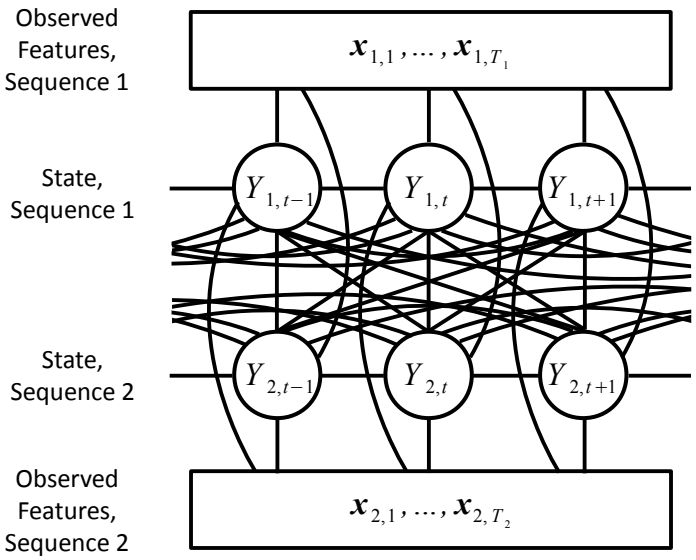


Figure 14. A diagram illustrating the CRF in the general case. Random variables that we model are shown in circles and random variables that we condition on are shown in rectangles. Conditional independence relationships are shown with node edges: conditional on the nodes that $Y_{i,t}$ is connected with, it is independent of the variables that it is not connected to. In the graph structure shown here each node is connected to all other nodes. Thus, the distribution of the state $Y_{i,t}$ depends on the observed features in all observation sequences and the states Y_{i^*,t^*} in all other observation sequences and time points.

In the general case, the activity type at each time t may depend on the activity types at all other times and the full vector $\mathbf{x}$. This is illustrated in Figure 14, where there are edges in the graph connecting each node $Y_{i,t}$ with each other Y_{i^*,t^*}, and with each $\mathbf{x}_{i^*}, i^* = 1, \ldots, N$.

We adopt a considerably less complex variation of the CRF model that makes several simplifying assumptions. First, we assume that the activity sequences for different subjects are conditionally independent given $\mathbf{x}$, and furthermore that $\mathbf{x}_{i^*}$ does not contain any information about $\mathbf{Y}_i$ for $i \neq i^*$. This allows us to express the conditional probability of the state sequences $\mathbf{Y}_i, i = 1, \ldots, N$ as a product of N separate CRFs:

$$p(\mathbf{Y}_1, \ldots, \mathbf{Y}_N | \mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{Y}_i | \mathbf{x}_i; \boldsymbol{\theta}), \tag{5.1.1}$$

where each term $p(\mathbf{Y}_i | \mathbf{x}_i; \boldsymbol{\theta})$ is modeled by a CRF with parameters $\boldsymbol{\theta}$. In terms of the graph, this assumption eliminates any connections between $Y_{i,t}$ and Y_{i^*,t^*} or $\mathbf{x}_{i^*,t^*}$ for $i \neq i^*$, as depicted in Figure 15.

The assumption of independence between the observation sequences seems plausible for

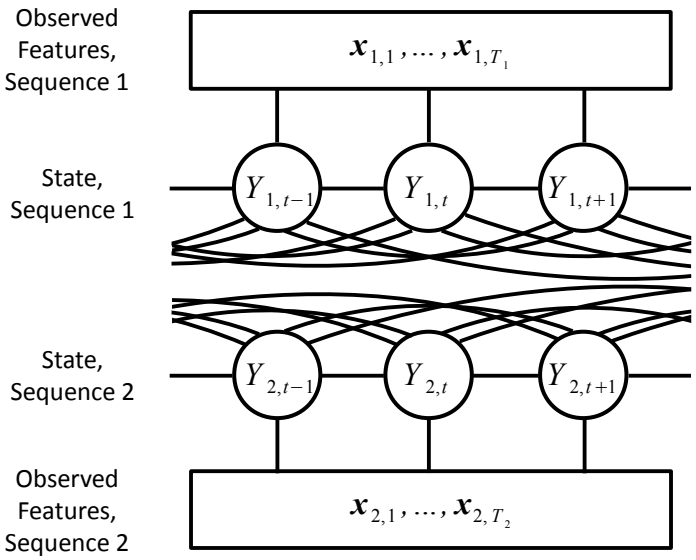


Figure 15. A diagram illustrating the use of a separate CRF for each sequence. Within each sequence, the distribution of the state $Y_{i,t}$ at each time depends on the states at all other times in that sequence and the feature values at all other times in that sequence.

the free living data from Sasaki [2013], where the subjects were observed separately and did not perform the same activities together, and there is only one observation sequence for each subject. In the laboratory setting, the subjects performed one of a small number of prescribed routines with a fixed order and duration of activity types. It is therefore conceivable that the observed accelerometer signal for one subject could contain information about the activity that another subject is performing at the same point in the routine. However, we do not wish to make use of this information in performing classification since this sort of connection between individuals is not generally available in the free living setting.

A second assumption is implicit in Equation (5.1.1): the CRF parameter vector $\boldsymbol{\theta}$ is the same for all subjects. As we will see, this means that the relative frequency of each activity class, the probabilities of transitioning from one activity type to another, and the compatibility of particular values of the accelerometer feature vector with each activity type do not vary with the subject. We saw evidence in Figures 9 and 10 in Chapter 3 that this assumption is not accurate: in fact, there appear to be variations between subjects in the frequencies of different activity types and the pattern of accelerometer features associated with each activity type. We do not model this variability between subjects because of the relatively limited size of our data sets. In particular,

for the free living setting, we have data for only 15 subjects performing a relatively limited range of activities for about 2 hours each. In Figure 9, it is not clear which of the apparent differences in the distributions of the observed features are due to differences in the movement patterns between subjects and which are due to the fact that the subjects performed different types of activities during the limited periods when they were observed. We believe that more data would be required to reliably model the variations between subjects.

We further simplify our model by assuming that within each activity sequence, the conditional dependence relations follow a first-order Markov structure:

$$p(Y_{i,t}|\mathbf{x}_i, Y_{i,t^*}, t \neq t^*) = p(Y_{i,t}|\mathbf{x}_i, Y_{i,t-1}, Y_{i,t+1}),$$

This simpler dependence structure corresponds to the linear chain graph depicted in Figure 16. We do not believe that changes in physical activity type over time really obey these first order Markov dynamics. However, we are wary of constructing complex models for the order and duration of activity types in the laboratory setting, since these were determined by the experimenters rather than the subjects. A complex model for time dependencies with the laboratory data might artificially inflate classification rates by capturing the specific order and duration of activities that were used for all subjects in the laboratory setups. In fact, we will consider an even further restricted parameterization of the model in our applications to the laboratory data in Chapter 7 in order to address this concern.

We have opted to focus on methods that are applicable to all three of our data sets, and have therefore used only models with the first-order dependence structure described above. However, we believe that relaxing this assumption could lead to improved classification performance in the free living setting. We will discuss possibilities for future work with more flexible dependence structures in Chapter 9.

Our final simplification to the structure of the CRF is depicted in Figure 17. Here, the conditional distribution for the state at time t depends only on the feature values at that time, rather than on the entire sequence of feature values. Our primary reason for making this restriction to the CRF model is to enable simpler comparison of the CRF estimation strategies we propose below with alternatives such as HMMs and static RFs, which typically only make use of the features at a single time point. However, we believe that estimation of a model where the distribution of the state at time t depends on the feature values in several adjacent windows

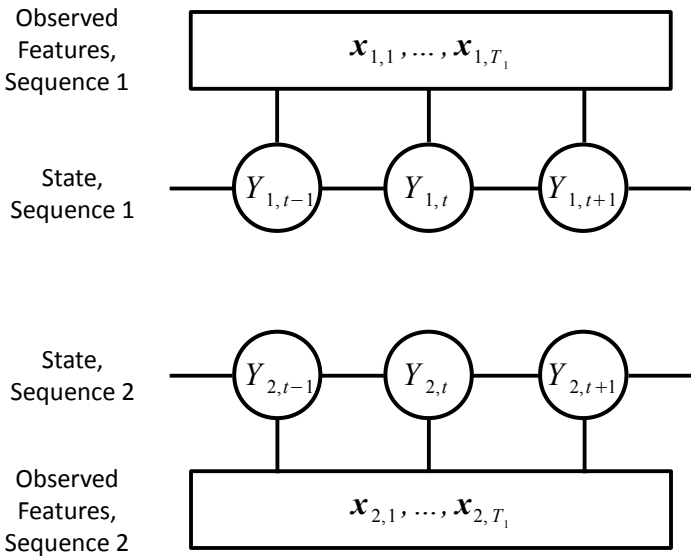


Figure 16. A diagram illustrating the use of a separate CRF for each sequence and first order Markov dependence. Within each sequence, the distribution of the state $Y_{i,t}$ at each time depends on the states at the adjacent two time points in the sequence and the feature values at all other times in that sequence.

$t - \delta, \ldots, t, \ldots, t + \delta$ would be feasible with our data and could lead to improvements in classification performance. Again, we will discuss this idea more in our conclusions in Chapter 9.

It is common to assume that the probability mass function for the linear chain CRF model illustrated in Figure 17 has the following form:

$$p(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} \exp\left\{ \sum_{t=2}^{T_i} \sum_{k=1}^{K} \Psi_k(y_{i,t-1}, y_{i,t}, \mathbf{x}_{i,t}; \boldsymbol{\theta}) \right\}. \tag{5.1.2}$$

Here the Ψ_k are functions depending on parameters $\boldsymbol{\theta}$, and $Z(\mathbf{x}_i; \boldsymbol{\theta})$ is a normalization constant ensuring that the conditional distribution of $\mathbf{Y}_i | \mathbf{X}_i = \mathbf{x}_i$ sums to 1. We work with a simplified variation of this model:

$$p(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} \exp\left\{ \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}) \zeta_s + \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1}) \mathbb{I}_{\{s\}}(y_{i,t}) \omega_{r,s} + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t}) \Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\xi}) \right\} \tag{5.1.3}$$

Here, $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_S]$, $\boldsymbol{\Omega} = [\omega_{r,s}]$, and $\boldsymbol{\xi}$ are parameters that are included in $\boldsymbol{\theta}$. Sutton and McCallum [2011] show that this CRF specification arises as the model for $\mathbf{Y}_i | \mathbf{X}_i$ obtained by conditioning on $\mathbf{X}_i$ in a first-order HMM. The term involving $\boldsymbol{\zeta}$ corresponds to the initial state distribution of the HMM, the term involving $\boldsymbol{\Omega}$ corresponds to the state transition probabilities of

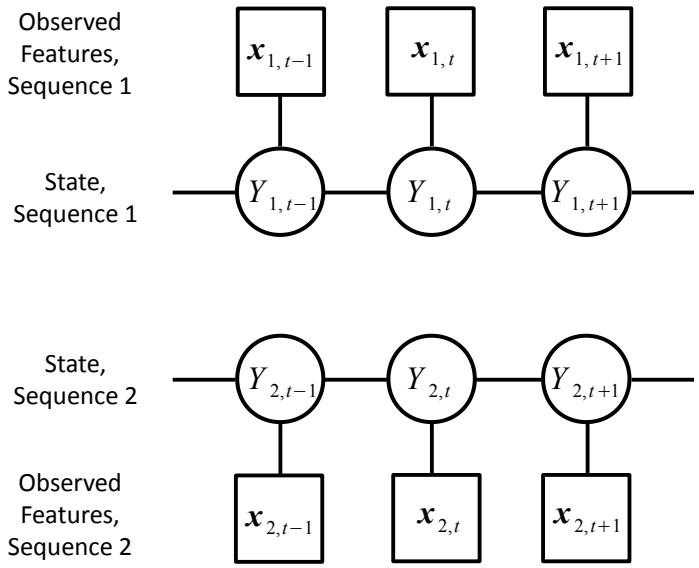


Figure 17. A diagram illustrating the use of a separate CRF for each sequence with fist order Markov dependence structure and dependence only on the observed features at a single time point. Within each sequence, the distribution of the state $Y_{i,t}$ at each time depends on the states at the adjacent two time points in the sequence and the feature values at that time.

the HMM, and the term involving Ψ_s corresponds to the observation distributions of the HMM. In the context of the CRF, these terms can be interpreted as giving the relative chances of observing an initial state of s, a transition between states r and s, and the feature vector $\mathbf{x}_{i,t}$ if the state at time t is s; however, unlike the HMM, they do not directly specify conditional probabilities for these events. We leave the functions Ψ_s unspecified for now; our primary contributions in the following Sections are new estimation strategies for Ψ_s.

In order to understand how our work fits into the existing literature, it is helpful to consider where each of our methods falls into the taxonomy of models and estimation strategies we outlined in the Chapter 1. We characterize the methods in general terms here, and discuss them in more detail in the following Sections.

All of the methods we propose share several qualities in common. As is clear from the discussion above, our methods are based on dynamic models that account for the sequential dependence in the activity types. We can also see from Equation (5.1.3) that our models are conditional models for the individuals' activity types $\mathbf{Y}_i$ over time given the observed features $\mathbf{X}_i$ from the accelerometer signal.

The decision to use dynamic models that condition on $\mathbf{X}_i$ is motivated by the discussion and

preliminary simulation study results outlined in Chapter 4. We saw there that dynamic models are helpful in the presence of sequential dependence, and conditional models are helpful when it is difficult to specify an accurate model for the distribution of $\mathbf{X}_i$ that can feasibly be estimated. The combination of the complexity of the distribution of accelerometer features as illustrated in Figure 8 and the fairly large number of features makes it difficult specify a model for the features that will perform well. The complexity of the feature distributions means that simple parametric models will be a poor representation of the true distribution, while the high dimension limits the effectiveness of more flexible approaches such as kernel density estimation.

Our methods are also similar to each other in that they are ensemble methods that use a logarithmic opinion pool to combine many component models and they make use of random feature subset selection in the parameter estimation process. However, the specific mechanics for obtaining the component model fits and selecting the feature subsets differ between our three approaches.

The method we discuss in Section 5.2 combines boosting and bagging at the level of observation sequences with a parametric specification of Ψ_s; we refer to this approach as **BB-Par-CRF**. The method we introduce in Section 5.3 also uses boosting and sequence bagging in the estimation process, but employs a non-parametric variation on the model by using regression trees for Ψ_s. We refer to this method as **BB-Nonpar-CRF**. Our third approach, discussed in Section 5.4, combines the estimation methods used for static random forests with the CRF model. We propose two variations on this third estimation strategy. In the first method, we perform bagging by resampling complete observation sequences. We refer to this approach as **RF-seq-CRF**. In the second method, we perform bagging at the level of individual time points within observation sequences. We refer to this method as **RF-CRF**. In each Section we give the model formulation, outline our proposed estimation strategy, and discuss the reasoning behind the decisions we have made in formulating the estimation algorithm.

5.2 CRF Model 1: BB-Par-CRF

In this Section we introduce the first of our three estimation strategies for the CRF model. We adopt a parametric specification for the Ψ_s functions in Equation (5.1.3); to connect with the common notation in generalized linear models, we allow Ψ_s to depend on parameters $\boldsymbol{\beta}$. The

full parameter vector is then $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\Omega}, \beta)$. Conditional on the observed features $\mathbf{x}_i$, the state sequences $\mathbf{Y}_i$ are assumed to be independent with the following distribution:

$$p(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} \exp \left\{ \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}) \sum_{m=1}^{M} \zeta_s^m + \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1}) \mathbb{I}_{\{s\}}(y_{i,t}) \sum_{m=1}^{M} \omega_{r,s}^m + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t}) \Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\beta}) \right\}, \text{ where} \tag{5.2.1}$$

$$\Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\beta}) = \sum_{m=1}^{M} \beta_{s,0}^m + \sum_{d=1}^{D} \left(\sum_{m=1}^{M} \beta_{s,d}^m \right) x_{i,t,d}, \text{ and} \tag{5.2.2}$$

$$Z(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{\mathbf{y}_i^*} \exp \left\{ \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}^*) \sum_{m=1}^{M} \zeta_s^m + \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1}^*) \mathbb{I}_{\{s\}}(y_{i,t}^*) \sum_{m=1}^{M} \omega_{r,s}^m + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t}^*) \Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\beta}) \right\}. \tag{5.2.3}$$

The index $\mathbf{y}_i^*$ of the summation in the expression for $Z(\mathbf{x}_i; \boldsymbol{\theta})$ varies over all possible sequences of states. $\mathbb{I}_A(x)$ is the indicator function, taking the value 1 if $x \in A$ and 0 otherwise.

The parameters in this model are not identifiable for two reasons. First, we can permute the parameter values among the m indices without affecting the value of the likelihood function. However, this is not a problem for the particular estimation algorithm we develop. Another issue is that if we add any vector $\boldsymbol{\delta}$ of length $D + 1$ to each $\boldsymbol{\beta}_s^m$, the conditional probability of $\mathbf{Y}_i | \mathbf{X}_i$ is unchanged since $\boldsymbol{\delta}$ cancels out of the numerator and denominator of Equation (5.2.1). Similar problems arise for the ζ_s^m and $\omega_{r,s}^m$ parameters. We resolve these problems by fixing $\zeta_S^m = 0$, $\omega_{S,S}^m = 0$, and $\boldsymbol{\beta}_S^m = \underline{0}$ for all m.

Equations (5.2.1) through (5.2.3) present the model as a single CRF, but it is also possible to view it as a combination of M separate CRFs in a LOP by rearranging the order of the summa-

tions:

$$
\begin{aligned}
p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}) &= \frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})} \exp\left[\sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}) \sum_{m=1}^{M} \zeta_s^m \right. \\
&+ \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1}) \mathbb{I}_{\{s\}}(y_{i,t}) \sum_{m=1}^{M} \omega_{r,s}^m \\
&\left. + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t}) \left\{ \sum_{m=1}^{M} \beta_{s,0}^m + \sum_{d=1}^{D} \left(\sum_{m=1}^{M} \beta_{s,d}^m \right) x_{i,t,d} \right\} \right] \\
&= \frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})} \exp\left[\sum_{m=1}^{M} \left\{ \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}) \zeta_s^m + \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1}) \mathbb{I}_{\{s\}}(y_{i,t}) \omega_{r,s}^m \right. \right. \\
&\left. \left. + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t}) \left(\beta_{s,0}^m + \sum_{d=1}^{D} \beta_{s,d}^m x_{i,t,d} \right) \right\} \right] \frac{\exp\{-\sum_{m=1}^{M} \log Z(\mathbf{x}_i;\boldsymbol{\theta}^m)\}}{\exp\{-\sum_{m=1}^{M} \log Z(\mathbf{x}_i;\boldsymbol{\theta}^m)\}} \\
&= \frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta}) \exp\{-\sum_{m=1}^{M} \log Z(\mathbf{x}_i;\boldsymbol{\theta}^m)\}} \exp\left\{ \sum_{m=1}^{M} \log p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}^m) \right\} \quad (5.2.4)
\end{aligned}
$$

Here, $p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}^m)$ is the conditional probability of the state sequence $\mathbf{y}_i$ from a CRF with parameters $\boldsymbol{\zeta}^m$, $\boldsymbol{\Omega}^m = [\omega_{r,s}^m]$, and $\boldsymbol{\beta}^m$, and $Z(\mathbf{x}_i;\boldsymbol{\theta}^m)$ is defined analogously to Equation (5.2.3). By performing the same steps in the denominator that we showed above for the numerator, we obtain

$$
p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}) = \frac{\exp\left\{\sum_{m=1}^{M} \log p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}^m)\right\}}{\sum_{\mathbf{y}_i^*} \exp\left\{\sum_{m=1}^{M} \log p(\mathbf{Y}_i = \mathbf{y}_i^*|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}^m)\right\}}, \quad (5.2.5)
$$

which is a combination of the conditional state sequence probabilities from M separate CRFs in the form of a LOP.

The log-likelihood for the model parameters $\boldsymbol{\theta}$ based on a training data set comprising N

observation sequences is:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) &= \sum_{i=1}^{N} \log\{p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})\} \\
&= \sum_{i=1}^{N}\left[\sum_{m=1}^{M}\left\{\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,1})\zeta_s^m + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S}\mathbb{I}_{\{r\}}(y_{i,t-1})\mathbb{I}_{\{s\}}(y_{i,t})\omega_{r,s}^m \right.\right. \\
&\qquad \left.\left. + \sum_{t=1}^{T_i}\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,1})\left(\beta_{s,0}^m + \sum_{d=1}^{D}\beta_{s,d}^m x_{i,t,d}\right)\right\} - \log\{Z(\mathbf{x}_i;\boldsymbol{\theta})\}\right] \\
&= \sum_{i=1}^{N}\left[\sum_{m=1}^{M}\left\{\zeta_{y_{i,1}}^m + \sum_{t=2}^{T_i}\omega_{y_{i,t-1},y_{i,t}}^m + \sum_{t=1}^{T_i}\left(\beta_{y_{i,t},0}^m + \sum_{d=1}^{D}\beta_{y_{i,t},d}^m x_{i,t,d}\right)\right\}\right. \\
&\qquad - \log\left\{\sum_{\mathbf{y}_i^*}\exp\left(\sum_{m=1}^{M}\left\{\zeta_{y_{i,1}^*}^m + \sum_{t=2}^{T_i}\omega_{y_{i,t-1}^*,y_{i,t}^*}^m \right.\right.\right. \\
&\qquad\qquad \left.\left.\left. + \sum_{t=1}^{T_i}\left(\beta_{y_{i,t}^*,0}^m + \sum_{d=1}^{D}\beta_{y_{i,t}^*,d}^m x_{i,t,d}\right)\right\}\right)\right\}\Bigg]
\end{aligned}
\tag{5.2.6}
$$

We now discuss our estimation strategy for this model. At a general level, the algorithm employs both bagging and boosting. In the bagging step we generate many different training data sets by drawing observation sequences with replacement from the full set of all observation sequences. The boosting step then produces a separate collection of component model fits for each of these training data sets. These collections of component models are independent in the sense that the parameter estimates for component models obtained using one bagged training data set do not depend on the parameter estimates for component models based on a different bagged training data set. However, within each of these collections, the component models are estimated sequentially: in each iteration of the boosting step, a new component model is estimated with parameter estimates that depend on the parameter estimates for previous component models. At the end of this process, all of the resulting component models are combined in a LOP. This overview of the algorithm is illustrated in Figure 18. Algorithm 5.1 gives a step-by-step description of the procedure.

We have two reasons for using bagging in our estimation procedure. First, as we saw in the literature review, bagging can reduce the generalization error of classifiers. In an application of parametric CRF models to text processing, Smith and Osborne [2007] explored the idea of using bagging and combining the resulting models with a LOP and found that this led to a small increase in classification performance. Our methods differ from theirs in that we use boosting to estimate the component models based on each bagged data set, while they used unpenalized

Algorithm 5.1. ***BB-Par-CRF*** *Estimation Algorithm*
Method: bag_boost_par_LCCRF
Inputs: *Labeled sequence data* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N\}$
Outputs: *Bagged and boosted model parameter estimates.*

1. *Initialize* `component_models` $= \{\}$.
2. *For* $b = 1, \ldots, M_{bag}$, *repeat the following:*
   (a) *Draw a sample of* N *observation sequences with replacement from the full set of all observation sequences. Collect these sequences in* $\mathcal{B}^b$. *Collect the unsampled sequences in* $\mathcal{O}^b$.
   (b) *Set* `new_component_models` $=$ ***boost_par_LCCRF***$(\mathcal{B}^b, \mathcal{O}^b)$.
   (c) *Set* `component_models` $=$ `component_models` $\bigcup$ `new_component_models`.
3. *For each component in* `component_models`, *divide the estimates* $\widehat{\zeta}_s^m$, $\widehat{\omega}_{r,s}^m$, *and* $\widehat{\boldsymbol{\beta}}_s^m$ *by* M_{bag}.
4. *Return* `component_models`.

Method: boost_par_LCCRF
Inputs: *Labeled sequence data* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N_{train}\}$ *and* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N_{validation}\}$.
Outputs: *Boosted model parameter estimates.*

1. *Initialize* `component_models` $= \{\}$, $m = 0$, *and* `val_prop_correct`$[0] = -\infty$.
2. *Repeat the following until the largest element of* `val_prop_correct` *is not within the last* `M_search_threshold` *values stored in* `val_prop_correct`*:*
   (a) *Set* $m = m + 1$, `attempt_num` $= 0$, *and* `val_prop_correct`$[m] = -\infty$.
   (b) *Repeat the following until* `val_prop_correct`$[m]$ $>$ `val_prop_correct`$[m-1]$ *or* `attempt_num` $=$ `max_attempts`*:*
      i. *Set* `attempt_num` $=$ `attempt_num` $+ 1$.
      ii. *Randomly select the set* $\mathcal{A}^m \subset \{1, \ldots, D\}$ *of active features for the* m*th component model. The coefficients* $\beta_{s,d}^m$ *are fixed equal to* 0 *for* $d \notin \mathcal{A}^m$.
      iii. *If* $m = 1$, *initialize* $\tilde{\zeta}_s^m = log(\frac{n_s}{n_S})$ *and* $\tilde{\omega}_{r,s}^m = log(\frac{n_{r,s}}{n_{S,S}})$ *for all* $r, s = 1, \ldots, S$. *Here,* n_s *is the number of occurrences of state* s *and* $n_{r,s}$ *is the number of transitions from state* r *to state* s *in the training data set. If* $m > 1$, *initialize* $\tilde{\zeta}_s^m = \tilde{\omega}_{r,s}^m = 0$ *for all* $r, s = 1, \ldots, S$.
      iv. *Initialize* $\tilde{\boldsymbol{\beta}}_{s,d}^m$ *for* $s = 1, \ldots, S-1$, $d \in \{0\} \bigcup \mathcal{A}^m$ *by fitting a multinomial logistic regression model with offsets equal to*

$$\left(\sum_{l=1}^{m-1} \zeta_{y_{i,1}}^l\right)^{\mathbb{I}_{\{1\}}(t)} \left(\sum_{l=1}^{m-1} \omega_{y_{i,t}}^l\right)^{(1-\mathbb{I}_{\{1\}}(t))} + \sum_{l=1}^{m-1} \left(\beta_{s,0}^l + \sum_{d=1}^{D} \beta_{s,d}^l x_{i,t,d}\right).$$

      v. *Using a numerical optimization routine, update* $\tilde{\omega}_{r,s}^m$ *and* $\tilde{\boldsymbol{\beta}}^m$ *to the constrained local maximum likelihood estimates based on the training data, holding the parameter estimates for previous components and elements of* $\widehat{\boldsymbol{\beta}}_s^m$ *not in the active set fixed.*
      vi. *Using all component fits up through component* m, *predict the values of* $\mathbf{y}_i$ *for the validation data set. If the proportion of time points at which the prediction was correct is greater than* `val_prop_correct`$[m]$, *store it in* `val_prop_correct`$[m]$ *and set* $\widehat{\boldsymbol{\zeta}}^m = \tilde{\boldsymbol{\zeta}}^m$, $\widehat{\boldsymbol{\Omega}}^m = \tilde{\boldsymbol{\Omega}}^m$, *and* $\widehat{\boldsymbol{\beta}}^m = \tilde{\boldsymbol{\beta}}^m$.
   (c) *Set* `component_models` $=$ `component_models` $\bigcup \{(\widehat{\boldsymbol{\zeta}}^m, \widehat{\boldsymbol{\Omega}}^m, \widehat{\boldsymbol{\beta}}^m)\}$.
3. *Return* `component_models`.

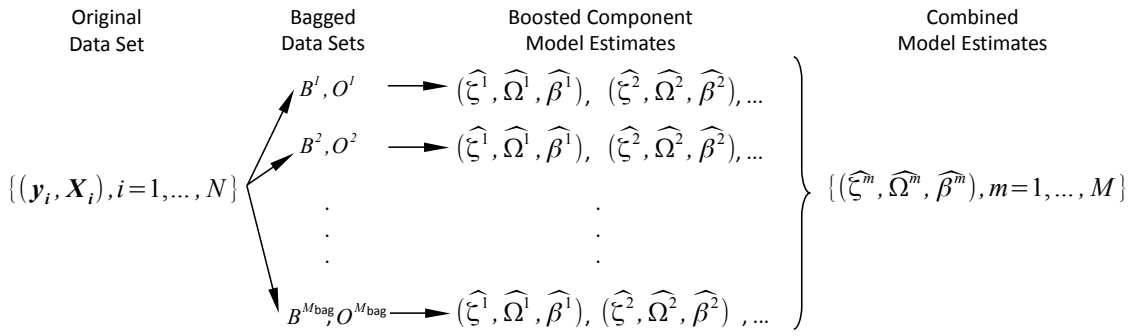


Figure 18. A diagram illustrating the estimation process for the BB-Par-CRF model. We begin by drawing the bagged data sets $\mathcal{B}^b$, $b = 1, \ldots, M_{bag}$; these define the corresponding out of bag data sets $\mathcal{O}^b$. Each bagged data set is used separately as the input to the boosting procedure, which produces a collection of component model estimates. These component models are then combined to obtain the final model estimate.

maximum likelihood.

A second benefit to bagging is that we can think of it as a mechanism for approximating an exhaustive leave-p-out cross-validation procedure in which all subsets of size p are used as the validation set in the boosting step. As we will discuss in more detail shortly, the boosting step relies on the use of a validation procedure to determine the number of boosting iterations to perform. Often, boosting methods use a single partition of the observed data into training and validation subsets [Hastie et al., 2009]. However, the performance of the resulting classifier can be sensitive to the choice of which observations are in the training and validation sets. We can mitigate this problem by combining the results from many different assignments of observations to the training and validation sets. Bagging and combining the results with a LOP is one way to achieve this since each bagged training data set naturally defines a corresponding validation data set: the collection of all observation sequences that were not included in the bagged training set.

The number M_{bag} of bagged training data sets is a user-supplied parameter. A common recommendation for bagging is that M_{bag} should be large enough that the resulting predictions are similar to what they would be if a value of $M_{bag}/2$ were used [e.g., Liaw and Wiener, 2002]. When a two-stage procedure is used in which the component model parameters are estimated separately before combining them with a LOP, it is beneficial to scale the component model parameters using non-negative weights that sum to 1. As we discussed in the literature review,

previous researchers have found that allowing for assignment of unequal weights to the component models does not result in appreciable gains in predictive performance. We save computation time by using equal weights of $1/M_{bag}$.

The boosting step of Algorithm 5.1 can be interpreted as a random block coordinate ascent algorithm converging to the maximum likelihood parameter estimates based on the given training data set, but with early stopping used to reduce overfitting. In each boosting iteration, we estimate the parameters of a new CRF with parameters $\boldsymbol{\theta}^m$. This new CRF uses only a randomly selected subset of the observed features; the coefficients $\beta_{s,d}^m$ for the remaining features are set to 0. Roughly, we estimate these parameters by maximizing the likelihood within the affine subspace of the parameter space that is spanned by the new parameter updates:

$$\widehat{\boldsymbol{\theta}}^m = \underset{\boldsymbol{\theta}^m}{\operatorname{argmax}}\, \ell\left(\sum_{l=1}^{m-1} \widehat{\boldsymbol{\theta}}^l + \boldsymbol{\theta}^m | (\mathbf{y}_i, \mathbf{x}_i), i = 1, \ldots, N_{train}\right) \tag{5.2.7}$$

Here, the likelihood function ℓ is based on a particular bagged training data set and the index l runs over the parameter estimates obtained in all previous boosting iterations using the same training set. We hold these previous parameter values fixed at the time that we estimate $\boldsymbol{\theta}^m$. The fact that some of the $\boldsymbol{\beta}_{s,d}^m$ coefficients are fixed to 0 means that parameter vectors of the form $\sum_{l=1}^{m-1} \widehat{\boldsymbol{\theta}}^l + \boldsymbol{\theta}^m$ do not include the full parameter space. The inclusion of the previous parameter estimates $\sum_{l=1}^{m-1} \widehat{\boldsymbol{\theta}}^l$ as an offset means that each update searches over an affine subspace of the parameter space.

Our algorithm deviates from the rule specified by Equation (5.2.7) in one way: we do not update the estimates of $\boldsymbol{\zeta}$ in the boosting process. Instead, we fix $\boldsymbol{\zeta} = log(\frac{n_s}{T_{train}}) - log(\frac{n_S}{T_{train}}) = log(\frac{n_s}{n_S})$, where n_s is the number of observations from state s in the training data set. We do not use the maximum likelihood estimates for $\boldsymbol{\zeta}$ because most of the information about $\boldsymbol{\zeta}$ in the likelihood function is contained in the first observation from each sequence. Since we do not have many observation sequences, the maximum likelihood estimates of these parameters would have high sampling variance. We therefore take a strategy that is commonly used in estimation of HMMs and base our estimates on the overall frequency of each observed class. The particular values we use correspond to the estimates that are often used for HMMs in the sense of the connection between HMMs and CRFs that we described in Section 5.1.

Random coordinate descent algorithms have been examined for a variety of optimization problems and shown to converge to local minima under regularity conditions on the objective

function [Shalev-Shwartz and Tewari, 2011, Richtárik and Takáč, 2014]. The motivation for using random coordinate descent in these papers is different from ours, however. It has been observed that in many applications of CRFs, the maximum likelihood parameter estimates overfit the training data, so that classification performance on a new data set is lower than it was on the training data. This problem is particularly common in settings where the dimension of the feature space is large. One solution to this is to perform estimation by maximizing a penalized likelihood function that effectively limits the magnitude of some of the coefficients.

Our algorithm represents an alternative approach in which we regularize the CRF model by gradually moving towards the maximum likelihood parameter estimates, but stopping the estimation process before we have begun to overfit the training data. The stopping point is selected by evaluation of the classifier's performance on a validation data set. When the proportion of time points classified correctly in the validation set starts to decrease, we stop the boosting procedure. The particular validation data set we use is the out-of-bag data that were not drawn in creating each particular training data set in the bagging step. In our applications in Chapters 7 and 8, we use leave-one-subject-out cross validation to evaluate performance of our estimation algorithm. When we do that, we will embed the bagging procedure within the leave-one-subject-out cross validation, so that separate validation data sets are used to evaluate model performance and select the stopping point in the boosting process. We will describe these procedures in more detail in the applications Chapters.

The use of randomized coordinate ascent to search the parameter space can be viewed as a mechanism for slowing down convergence to the maximum likelihood parameter estimates. This results in a larger number of potential stopping points that may yield good performance on the validation data set. Similar mechanisms for slowing down the convergence of the parameter estimates are common in boosting algorithms; examples include shrinkage of coefficient estimates or limiting the number of leaves in regression trees [Hastie et al., 2009].

The proportion correct in the validation set can fluctuate, particularly near the beginning of the boosting process. In order to handle this, Algorithm 5.1 allows for a user-specified search threshold, `M_search_threshold`. The boosting process continues as long as the first occurrence of the maximum proportion correct was within the last `M_search_threshold` iterations. The value of `M_search_threshold` can be selected by examining plots of the proportion correct as a function of the boosting iteration; in our applications of the method, we have set `M_search_threshold` $=$

100. Figure 19 illustrates this search strategy.

Within each boosting iteration, we obtain the estimates $\mathbf{\Omega}^m$ and $\boldsymbol{\beta}^m$ through numerical optimization of the log-likelihood function. If we are in the first boosting iteration for a particular training data set, we initialize $\widehat{\omega}_{r,s}^m = log(\frac{n_{r,s}}{T_{train}-N}) - log(\frac{n_{S,S}}{T_{train}-N}) = log(\frac{n_{r,s}}{n_{S,S}})$ for all $r, s = 1, \ldots, S$, where $n_{r,s}$ is the number of observed transitions from state r to state s in the training data set. These initial values correspond to the estimates that would be obtained from a HMM by maximum likelihood. If we are past the first boosting iteration, we initialize $\widehat{\omega}_{r,s}^m = 0$ for all $r, s = 1, \ldots, S$. Effectively, this amounts to re-using the value of $\mathbf{\Omega}$ that was obtained after the previous boosting iteration.

We initialize $\boldsymbol{\beta}^m$ by fitting a multinomial logistic regression model with offset terms for each observation. This procedure can be motivated by considering estimation for a simplified version of the model where all rows of $\mathbf{\Omega}^m$ are the same: $\omega_{r,s}^m = \omega_s^m \, \forall r = 1, \ldots, S$. With this simplification the model no longer accounts for temporal dependence in the labels $y_{i,t}$, and we can rewrite the

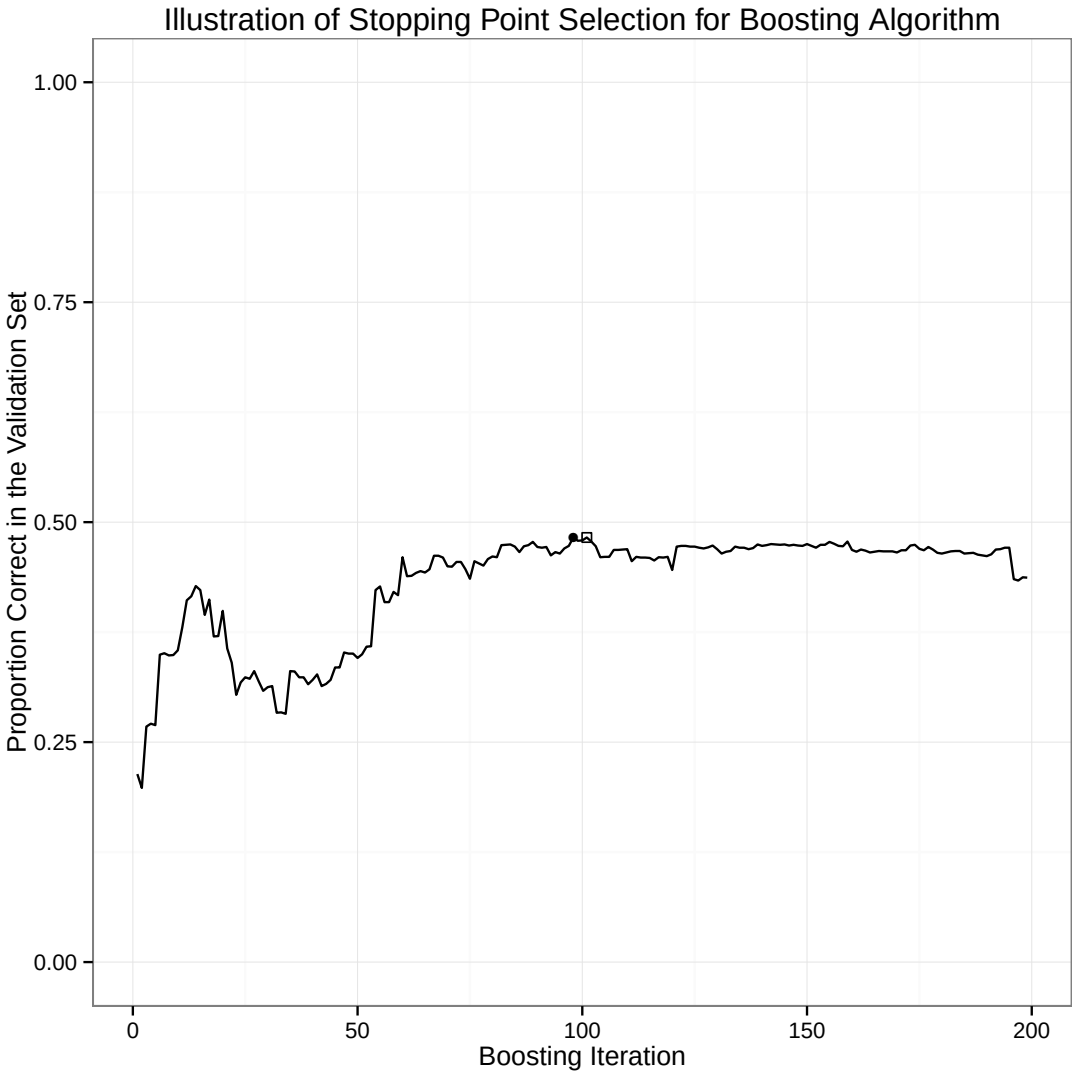


Figure 19. Selection of the stopping point in the boosting step. The horizontal axis represents the boosting iteration and the vertical axis shows the proportion of windows classified correctly in the validation data set. The search threshold is set to 100 iterations. The boosting process halts when the first occurrence of the maximum proportion correct in the validation set was not within the last 100 iterations. In this example, the first occurrence is on iteration 98, indicated in the plot with a solid circle. This maximum is reached again at iteration 101, indicated with an unfilled square. The boosting process halts after iteration 199. This example is taken from the fit to the hip data from subject 1 in the free living data from Sasaki [2013], which we describe in more detail in Chapter 7.

restricted log-likelihood for $\boldsymbol{\beta}^m$ (holding fixed $\boldsymbol{\zeta}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\beta}^l, l = 1, \ldots, m-1$) as follows:

$$
\begin{aligned}
&\ell(\boldsymbol{\beta}^m; \mathbf{y}, \mathbf{x}, \boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\beta}^l, l = 1, \ldots, m-1) \qquad (5.2.8)\\
&= \sum_{i=1}^{N} \left[\zeta_{y_{i,1}} + \sum_{t=2}^{T_i} \sum_{l=1}^{m} \omega^l_{y_{i,t-1}, y_{i,t}} + \sum_{t=1}^{T_i} \sum_{l=1}^{m} \left(\beta^l_{y_{i,t},0} + \sum_{d=1}^{D} \beta^l_{y_{i,t},d} x_{i,t,d} \right) \right.\\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}^*_i} \exp \left(\zeta_{y^*_{i,1}} + \sum_{t=2}^{T_i} \sum_{l=1}^{m} \omega^l_{y^*_{i,t-1}, y^*_{i,t}} + \sum_{t=1}^{T_i} \sum_{l=1}^{m} \left(\beta^l_{y^*_{i,t},0} + \sum_{d=1}^{D} \beta^l_{y^*_{i,t},d} x_{i,t,d} \right) \right) \right\} \right]\\
&= \sum_{i=1}^{N} \left[\zeta_{y_{i,1}} + \sum_{t=2}^{T_i} \sum_{l=1}^{m} \omega^l_{y_{i,t}} + \sum_{t=1}^{T_i} \sum_{l=1}^{m} \left(\beta^l_{y_{i,t},0} + \sum_{d=1}^{D} \beta^l_{y_{i,t},d} x_{i,t,d} \right) \right.\\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}^*_i} \exp \left(\zeta_{y^*_{i,1}} + \sum_{t=2}^{T_i} \sum_{l=1}^{m} \omega^l_{y^*_{i,t}} + \sum_{t=1}^{T_i} \sum_{l=1}^{m} \left(\beta^l_{y^*_{i,t},0} + \sum_{d=1}^{D} \beta^l_{y^*_{i,t},d} x_{i,t,d} \right) \right) \right\} \right]\\
&= \sum_{i=1}^{N} \left[\sum_{t=1}^{T_i} \left\{ \zeta^{\mathbb{I}_{\{1\}}(t)}_{y_{i,1}} \left(\sum_{l=1}^{m} \omega^l_{y_{i,t}} \right)^{(1-\mathbb{I}_{\{1\}}(t))} + \sum_{l=1}^{m} \left(\beta^l_{y_{i,t},0} + \sum_{d=1}^{D} \beta^l_{y_{i,t},d} x_{i,t,d} \right) \right\} \right.\\
&\qquad - \log \left\{ \sum_{\mathbf{y}^*_i} \exp \left(\sum_{t=1}^{T_i} \left\{ \zeta^{\mathbb{I}_{\{1\}}(t)}_{y^*_{i,1}} \left(\sum_{l=1}^{m} \omega^l_{y^*_{i,t}} \right)^{(1-\mathbb{I}_{\{1\}}(t))} \right.\right.\right.\\
&\qquad\qquad \left.\left.\left. + \sum_{l=1}^{m} \left(\beta^l_{y^*_{i,t},0} + \sum_{d=1}^{D} \beta^l_{y^*_{i,t},d} x_{i,t,d} \right) \right\} \right) \right\} \Bigg]\\
&= \sum_{i=1}^{N} \left[\sum_{t=1}^{T_i} \left\{ F^{m-1}(y_{i,t}, \mathbf{x}_{i,t}, t) + \beta^m_{y_{i,t},0} + \sum_{d=1}^{D} \beta^m_{y_{i,t},d} x_{i,t,d} \right\} \right.\\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}^*_i} \prod_{t=1}^{T_i} \exp \left(F^{m-1}(y^*_{i,t}, \mathbf{x}_{i,t}, t) + \beta^m_{y^*_{i,t},0} + \sum_{d=1}^{D} \beta^m_{y^*_{i,t},d} x_{i,t,d} \right) \right\} \right]\\
&= \sum_{i=1}^{N} \left[\sum_{t=1}^{T_i} \left\{ F^{m-1}(y_{i,t}, \mathbf{x}_{i,t}, t) + \beta^m_{y_{i,t},0} + \sum_{d=1}^{D} \beta^m_{y_{i,t},d} x_{i,t,d} \right\} \right.\\
&\qquad \left. - \log \left\{ \prod_{t=1}^{T_i} \sum_{\mathbf{y}^*_{i,t}=1}^{S} \exp \left(F^{m-1}(y^*_{i,t}, \mathbf{x}_{i,t}, t) + \beta^m_{y^*_{i,t},0} + \sum_{d=1}^{D} \beta^m_{y^*_{i,t},d} x_{i,t,d} \right) \right\} \right]\\
&= \sum_{i=1}^{N} \sum_{t=1}^{T_i} \left[F^{m-1}(y_{i,t}, \mathbf{x}_{i,t}, t) + \beta^m_{y_{i,t},0} + \sum_{d=1}^{D} \beta^m_{y_{i,t},d} x_{i,t,d} \right.\\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}^*_{i,t}=1}^{S} \exp \left(F^{m-1}(y^*_{i,t}, \mathbf{x}_{i,t}, t) + \beta^m_{y^*_{i,t},0} + \sum_{d=1}^{D} \beta^m_{y^*_{i,t},d} x_{i,t,d} \right) \right\} \right] \qquad (5.2.9)
\end{aligned}
$$

Equation (5.2.9) is the log-likelihood of a multinomial logistic regression with offsets of

$$
F^{m-1}(s, \mathbf{x}_{i,t}, t) = \zeta^{\mathbb{I}_{\{1\}}(t)}_{y_{i,1}} \left(\sum_{l=1}^{m} \omega^l_{y_{i,t}} \right)^{(1-\mathbb{I}_{\{1\}}(t))} + \sum_{l=1}^{m-1} \left(\beta^l_{s,0} + \sum_{d=1}^{D} \beta^l_{s,d} x_{i,t,d} \right)
$$

for each observation index (i,t) and class $s = 1,\ldots,S$. Using these initial values saves computation time by reducing the number of times that the full CRF likelihood must be evaluated.

We use the "L-BFGS-B" method of Byrd et al. [1995] to perform the optimization. This method takes the model log-likelihood and its gradient as inputs. We now show that the partial derivatives of the log-likelihood with respect to $\omega^{m^*}_{r^*,s^*}$ and $\beta^{m^*}_{s^*,d^*}$ for particular values of r^*, s^*, d^*, and m^* can be expressed in terms of the marginal class probabilities $p(Y_{i,t} = s^*|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})$ and $p(Y_{i,t-1} = r^*, Y_{i,t} = s^*|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})$. These relationships have been derived previously [e.g., Sutton and McCallum, 2011], and are helpful because techniques for relatively fast computation of these marginal class probabilities are available. We will discuss these methods in Section 5.6.

We consider the partial derivatives of the log-likelihood at stage m^* of the boosting process based on one of the bagged training samples. For convenience, we re-index the observation sequences in the bagged sample by $i = 1,\ldots,N_{train}$, as in the **boost_par_LCCRF** method of Algorithm 5.1. To simplify the notation, we set

$$A(\mathbf{x}_i,\mathbf{y}_i;\boldsymbol{\theta}) = \zeta_{y_{i,1}} + \sum_{t=2}^{T_i}\sum_{l=1}^{m^*}\omega^l_{y_{i,t-1},y_{i,t}} + \sum_{t=1}^{T_i}\sum_{l=1}^{m^*}\left(\beta^l_{y_{i,t},0} + \sum_{d=1}^{D}\beta^l_{y_{i,t},d}x_{i,t,d}\right).$$

The partial derivative with respect to $\omega^{m^*}_{r^*,s^*}$ can then be rewritten as follows:

$$\begin{aligned}
\frac{\partial}{\partial\omega^{m^*}_{r^*,s^*}}\ell(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) &= \sum_{i=1}^{N}\frac{\partial}{\partial\omega^{m^*}_{r^*,s^*}}\log\{p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})\} \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial\omega^{m^*}_{r^*,s^*}}\left[A(\mathbf{x}_i,\mathbf{y}_i;\boldsymbol{\theta}) - \log\{Z(\mathbf{x}_i;\boldsymbol{\theta})\}\right] \\
&= \sum_{i=1}^{N}\left[\sum_{t=2}^{T_i}\mathbb{I}_{\{r^*\}}(y_{i,t-1})\mathbb{I}_{\{s^*\}}(y_{i,t})\right. \\
&\qquad \left. -\frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})}\sum_{\mathbf{y}^*_i}\exp\left\{A(\mathbf{x}_i,\mathbf{y}^*_i;\boldsymbol{\theta})\right\}\sum_{t=2}^{T_i}\mathbb{I}_{\{r^*\}}(y^*_{i,t-1})\mathbb{I}_{\{s^*\}}(y^*_{i,t})\right] \\
&= \sum_{i=1}^{N}\sum_{t=2}^{T_i}\left\{\mathbb{I}_{\{r^*\}}(y_{i,t-1})\mathbb{I}_{\{s^*\}}(y_{i,t}) - p(Y_{i,t-1} = r^*, Y_{i,t} = s^*|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})\right\} \qquad (5.2.10)
\end{aligned}$$

For $\beta^{m^*}_{s^*,d^*}$, we have:

$$\begin{aligned}
\frac{\partial}{\partial \beta^{m^*}_{s^*,d^*}} \ell(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) &= \sum_{i=1}^{N} \frac{\partial}{\partial \beta^{m^*}_{s^*,d^*}} \log\{p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})\} \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \beta^{m^*}_{s^*,d^*}} \left[A(\mathbf{x}_i,\mathbf{y}_i) - \log\{Z(\mathbf{x}_i;\boldsymbol{\theta})\}\right] \\
&= \sum_{i=1}^{N} \left[\sum_{t=1}^{T_i} \mathbb{I}_{\{s^*\}}(y_{i,t}) x_{i,t,d^*} \right. \\
&\qquad \left. - \frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})} \sum_{\mathbf{y}_i^*} \exp\{A(\mathbf{x}_i,\mathbf{y}_i^*)\} \sum_{t=1}^{T_i} \mathbb{I}_{\{s^*\}}(y^*_{i,t}) x_{i,t,d^*}\right] \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T_i} x_{i,t,d^*} \left\{\mathbb{I}_{\{s^*\}}(y_{i,t}) - p(Y_{i,t} = s^*|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta})\right\}
\end{aligned}$$

In order to simplify this expression, we have augmented the observed feature vector $\mathbf{x}_{i,t} = (x_{i,t,1},\ldots,x_{i,t,D})$ with the value $x_{i,t,0} = 1$, corresponding to the intercept terms $\beta^m_{s,0}$.

5.3 CRF Model 2: BB-Nonpar-CRF

Our second model and estimation strategy are very similar to the methods we discussed in Section 5.2, but here we use regression trees for Ψ_s. To fit with the common notation for regression trees, we now allow the functions Ψ_s to depend on parameters $\boldsymbol{\tau}$; the full parameter vector is now $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\tau})$. Our revised model is as follows:

$$\begin{aligned}
p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})} \exp\left\{\sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}) \sum_{m=1}^{M} \zeta_s^m \right. \\
\left. + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1})\mathbb{I}_{\{s\}}(y_{i,t}) \sum_{m=1}^{M} \omega^m_{r,s} + \sum_{t=1}^{T_i}\sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t})\Psi_s(\mathbf{x}_{i,t};\boldsymbol{\tau})\right\}, \text{ where}
\end{aligned} \tag{5.3.1}$$

$$\Psi_s(\mathbf{x}_{i,t};\boldsymbol{\tau}) = \sum_{m=1}^{M}\sum_{j=1}^{J_s^m} \rho^m_{s,j} \mathbb{I}_{R^m_{s,j}}(\mathbf{x}_{i,t}), \text{ and} \tag{5.3.2}$$

$$\begin{aligned}
Z(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{\mathbf{y}_i^*} \exp\left\{\sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y^*_{i,1})\zeta_s + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y^*_{i,t-1})\mathbb{I}_{\{s\}}(y^*_{i,t}) \sum_{m=1}^{M} \omega^m_{r,s} \right. \\
\left. + \sum_{t=1}^{T_i}\sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y^*_{i,t})\Psi_s(\mathbf{x}_{i,t};\boldsymbol{\tau})\right\}.
\end{aligned} \tag{5.3.3}$$

In the expression for Ψ_s, the $R^m_{s,j}$ are the terminal regions of a binary regression tree that partitions the domain of $\mathbf{X}_{i,t}$ into disjoint regions and assigns a regression constant to each region.

The parameters for this tree are $\boldsymbol{\tau}_s^m = (\boldsymbol{\phi}_s^m, \boldsymbol{\rho}_s^m)$, where $\boldsymbol{\phi}_s^m$ specifies the variables and split points used in forming the regions $R_{s,j}^m$ and $\boldsymbol{\rho}_s^m$ gives the regression constant associated with each region. J_s^m denotes the number of leaf nodes in the regression tree for component model m and state s. We collect all of the model parameters in the vector $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\Omega}, \boldsymbol{\tau})$ This model has the same interpretation as a LOP of CRFs that we discussed for the parametric formulation in Section 5.2. The model also suffers from the same problems with identifiability that the parametric version of the model had. We resolve these problems by fixing $\zeta_S^m = 0$, $\omega_{S,S}^m = 0$, and $\boldsymbol{\rho}_S^m = \underline{0}$ for all m.

The log-likelihood for the model parameters based on N observation sequences is:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) &= \sum_{i=1}^{N} \log\{p(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta})\} \\
&= \sum_{i=1}^{N} \left[\sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1}) \sum_{m=1}^{M} \zeta_s^m + \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1}) \mathbb{I}_{\{s\}}(y_{i,t}) \sum_{m=1}^{M} \omega_{r,s}^m \right. \\
&\qquad \left. + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,t}) \Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\tau}) - \log\{Z(\mathbf{x}_i; \boldsymbol{\theta})\} \right] \\
&= \sum_{i=1}^{N} \left[\sum_{m=1}^{M} \left\{ \zeta_{y_{i,1}}^m + \sum_{t=2}^{T_i} \omega_{y_{i,t-1}, y_{i,t}}^m + \sum_{t=1}^{T_i} \sum_{j=1}^{J_{y_{i,t}}^m} \rho_{y_{i,t},j}^m \mathbb{I}_{R_{y_{i,t},j}^m}(\mathbf{x}_{i,t}) \right\} \right. \\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}_i^*} \exp \left(\sum_{m=1}^{M} \left\{ \zeta_{y_{i,1}^*}^m + \sum_{t=2}^{T_i} \omega_{y_{i,t-1}^*, y_{i,t}^*}^m + \sum_{t=1}^{T_i} \sum_{j=1}^{J_{y_{i,t}^*}^m} \rho_{y_{i,t}^*,j}^m \mathbb{I}_{R_{y_{i,t}^*,j}^m}(\mathbf{x}_{i,t}) \right\} \right) \right\} \right]
\end{aligned}
\tag{5.3.4}
$$

We use this model as the foundation for the approaches discussed in this Section and the next; the difference between the two methods lies in the estimation strategies we use. Our first estimation algorithm for this model is given in Algorithm 5.2. It is very similar to the algorithm we presented for the parametric model in Section 5.2 in that it combines bagging and boosting. The bagging step is essentially unchanged from the procedure in Algorithm 5.1: we draw M_{bag} samples of size N with replacement from the set of all observation sequences $\{(\mathbf{y}_i, \mathbf{X}_i), i = 1, \ldots, N\}$ and fit a collection of component models using each bagged data set. With the nonparametric formulation, rather than dividing estimated coefficients $\widehat{\beta}_s^m$ by M_{bag} as in step 3 of the **bag_boost_par_CRF** method, we now divide the estimated regression constant $\rho_{s,j}^m$ in each region by M_{bag}.

The boosting step requires more substantial changes to adapt to the nonparametric model specification. Each iteration of our boosting method proceeds in two stages. First we use a gra-

Algorithm 5.2. ***BB-Nonpar-CRF*** *Estimation Algorithm*
Method: bag_boost_nonpar_CRF
Inputs: *Labeled sequence data* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N\}$
Outputs: *Bagged and boosted model parameter estimates.*

1. *Perform 10-fold cross-validation to select the tree depth.*
2. *Initialize* `component_models` $= \{\}$.
3. *For* $b = 1, \ldots, M_{bag}$*, repeat the following:*
   (a) *Draw a sample of* N *observation sequences with replacement from the full set of all observation sequences. Collect these sequences in* $\mathcal{B}^b$*. Collect the unsampled sequences in* $\mathcal{O}^b$*.*
   (b) *Set* `new_component_models` $=$ ***boost_nonpar_CRF***$(\mathcal{B}^b, \mathcal{O}^b)$.
   (c) *Set* `component_models` $=$ `component_models` $\bigcup$ `new_component_models`.
4. *For each component in* `component_models`*, divide the estimates* $\widehat{\zeta}_s^m$*,* $\widehat{\omega}_{r,s}^m$*, and* $\widehat{\boldsymbol{\rho}}_{s,j}^m$ *by* M_{bag}.
5. *Return* `component_models`.

Method: boost_nonpar_CRF
Inputs: *Labeled sequence data* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N_{train}\}$ *and* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N_{validation}\}$.
Outputs: *Boosted model parameter estimates.*

1. *Initialize* `component_models` $= \{\}$*,* $m = 0$*, and* `val_prop_correct`$[0] = -\infty$.
2. *Repeat the following until the largest element of* `val_prop_correct` *is not within the last* `M_search_threshold` *values stored in* `val_prop_correct`*:*
   (a) *Set* $m = m + 1$*,* `attempt_num` $= 0$*, and* `val_prop_correct`$[m] = -\infty$.
   (b) *For each state* s*, create the set of regression examples*

$$\left\{ \left(\mathbf{x}_{i,t},\ \frac{\partial \ell(\boldsymbol{\Psi}|\mathbf{y}, \mathbf{x}; \boldsymbol{\zeta}, \boldsymbol{\Omega})}{\partial \Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\tau})} \bigg|_{\boldsymbol{\Psi} = \widehat{\boldsymbol{\Psi}}^{m-1}} \right), i = 1, \ldots, N, t = 1, \ldots, T_i \right\}.$$

   (c) *Repeat the following until* `val_prop_correct`$[m] >$ `val_prop_correct`$[m-1]$ *or* `attempt_num` $=$ `max_attempts`*:*
      i. *Set* `attempt_num` $=$ `attempt_num` $+ 1$.
      ii. *Randomly select the set* $\mathcal{A}^m \subset \{1, \ldots, D\}$ *of active features for the* m*th component model. Only splits using these features will be used in the regression trees.*
      iii. *If* $m = 1$*, initialize* $\tilde{\zeta}_s^m = log(\frac{n_s}{n_S})$ *and* $\tilde{\omega}_{r,s}^m = log(\frac{n_{r,s}}{n_{S,S}})$ *for all* $r, s = 1, \ldots, S$*. If* $m > 1$*, initialize* $\tilde{\zeta}_s^m = \tilde{\omega}_{r,s}^m = 0$ *for all* $r, s = 1, \ldots, S$.
      iv. *For each state* s*, use the CART algorithm to grow a regression tree with parameters* $(\tilde{\boldsymbol{\phi}}^m, \tilde{\boldsymbol{\rho}}^m)$ *using the data set for that state that was created in step 2b.*
      v. *Update* $\tilde{\omega}_{r,s}^m$ *to the constrained local maximum likelihood estimates based on the training data, holding fixed all other parameter estimates.*
      vi. *Using all component fits up through component* m*, predict the values of* $\mathbf{y}_i$ *for the validation data set. If the proportion of time points at which the prediction was correct is greater than* `val_prop_correct`$[m]$*, store it in* `val_prop_correct`$[m]$ *and set* $\widehat{\boldsymbol{\zeta}}^m = \tilde{\boldsymbol{\zeta}}^m$*,* $\widehat{\boldsymbol{\Omega}}^m = \tilde{\boldsymbol{\Omega}}^m$*,* $\widehat{\boldsymbol{\phi}}^m = \tilde{\boldsymbol{\phi}}^m$*, and* $\widehat{\boldsymbol{\rho}}^m = \tilde{\boldsymbol{\rho}}^m$.
   (d) *Set* `component_models` $=$ `component_models` $\bigcup \{(\widehat{\boldsymbol{\zeta}}^m, \widehat{\boldsymbol{\Omega}}^m, \widehat{\boldsymbol{\phi}}^m, \widehat{\boldsymbol{\rho}}^m)\}$.
3. *Return* `component_models`.

dient tree boosting approach based on the work of Dietterich et al. [2004] to update our estimates of Ψ_s, $s = 1, \ldots, S$. Then we find maximum likelihood updates for $\mathbf{\Omega}$, holding the Ψ_s fixed at their current estimates.

It will be helpful to introduce some notation in order to facilitate our discussion of the gradient tree boosting method. We use the bold $\mathbf{\Psi}$ to denote the vector of values $\Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\tau})$ for all combinations of states s and observations $\mathbf{x}_{i,t}$. We denote the vector of estimated values after stage m of the boosting process by $\widehat{\mathbf{\Psi}}^m$; a single element of this vector is $\widehat{\Psi}_s^m(\mathbf{x}_{i,t}; \boldsymbol{\tau}^l, l = 1, \ldots, m) = \sum_{l=1}^m \sum_{j=1}^{J_s^l} \rho_{s,j}^l \mathbb{I}_{R_{s,j}^l}(\mathbf{x}_{i,t})$.

The main idea behind gradient tree boosting is that the component regression trees are obtained as a series of approximate functional gradient ascent steps. In estimating the regression trees, the objective that we seek to maximize is the log-likelihood regarded as a function of the values $\mathbf{\Psi}$, holding fixed the values of $\boldsymbol{\zeta}^l$ and $\mathbf{\Omega}^l$ obtained from earlier components:

$$\begin{aligned}
\ell(\mathbf{\Psi}|\mathbf{y}, \mathbf{x}; \boldsymbol{\zeta}^l, \mathbf{\Omega}^l, l = 1, \ldots, m-1) &= \sum_{i=1}^N \log\{p(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \mathbf{\Psi}, \boldsymbol{\zeta}^l, \mathbf{\Omega}^l, l = 1, \ldots, m-1)\} \\
&= \sum_{i=1}^N \left[\sum_{l=1}^{m-1} \zeta_{y_{i,1}}^l + \sum_{t=2}^{T_i} \sum_{l=1}^{m-1} \omega_{y_{i,t-1}, y_{i,t}}^l + \sum_{t=1}^{T_i} \Psi_{y_{i,t}}(\mathbf{x}_{i,t}; \boldsymbol{\tau}) \right. \\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}_i^*} \exp \left(\sum_{l=1}^{m-1} \zeta_{y_{i,1}^*}^l + \sum_{t=2}^{T_i} \sum_{l=1}^{m-1} \omega_{y_{i,t-1}^*, y_{i,t}^*}^l + \sum_{t=1}^{T_i} \Psi_{y_{i,t}^*}(\mathbf{x}_{i,t}; \boldsymbol{\tau}) \right) \right\} \right]
\end{aligned}$$

In stage m of the boosting process, we wish to increase this log-likelihood function by updating our current estimates $\widehat{\Psi}_s^{m-1}(\mathbf{x}_{i,t}; \boldsymbol{\tau}^l, l = 1, \ldots, m-1) = \sum_{l=1}^{m-1} \sum_{j=1}^{J_s^l} \rho_{s,j}^l \mathbb{I}_{R_{s,j}^l}(\mathbf{x}_{i,t})$ with a new regression tree for each $s = 1, \ldots, S$. In order to do this, we note that the gradient of the log-likelihood with respect to $\mathbf{\Psi}$ at the current estimate $\widehat{\mathbf{\Psi}}^{m-1}$ indicates how much the log-likelihood can be improved by making a small change in the value that $\widehat{\Psi}_s^{m-1}$ takes at each observation. Where the gradient is positive, we can increase the likelihood by increasing the value of $\widehat{\Psi}_s^{m-1}(\mathbf{x}_{i,t}; \boldsymbol{\theta})$, and similarly when the gradient is negative. Our strategy for obtaining the new regression tree for state s is therefore to fit the tree to the pairs

$$\left\{ \left(\mathbf{x}_{i,t}, \left. \frac{\partial \ell(\mathbf{\Psi}|\mathbf{y}, \mathbf{x}; \boldsymbol{\zeta}, \mathbf{\Omega})}{\partial \Psi_s(\mathbf{x}_{i,t}; \boldsymbol{\tau})} \right|_{\mathbf{\Psi} = \widehat{\mathbf{\Psi}}^{m-1}} \right), i = 1, \ldots, N, t = 1, \ldots, T_i \right\}, \tag{5.3.5}$$

treating the partial derivative of the log-likelihood as the dependent variable to be estimated as a function of $\mathbf{x}_{i,t}$. In order to estimate the tree parameters, we use the CART algorithm of Breiman et al. [1984], as suggested by Friedman [2001] and Dietterich et al. [2004].

The partial derivatives in Equation (5.3.5) can be expressed in terms of the estimated class membership probabilities after stage $m-1$. This relationship can be shown as follows, where we differentiate with respect to the term for a particular state s^*, subject i^*, and time point t^*:

$$
\begin{aligned}
&\frac{\partial \ell(\boldsymbol{\Psi}|\mathbf{y},\mathbf{x};\boldsymbol{\zeta},\boldsymbol{\Omega})}{\partial \Psi_{s^*}(\mathbf{x}_{i^*,t^*};\boldsymbol{\tau})}\bigg|_{\boldsymbol{\Psi}=\widehat{\boldsymbol{\Psi}}^{m-1}} \\
&\quad= \sum_{i=1}^{N} \frac{\partial}{\partial \Psi_{s^*}(\mathbf{x}_{i^*,t^*};\boldsymbol{\tau})} \left[\sum_{l=1}^{m-1} \zeta^l_{y_{i,1}} + \sum_{t=2}^{T_i} \sum_{l=1}^{m-1} \omega^l_{y_{i,t-1},y_{i,t}} + \sum_{t=1}^{T_i} \Psi_{y_{i,t}}(\mathbf{x}_{i,t};\boldsymbol{\tau}) \right. \\
&\qquad \left. - \log \left\{ \sum_{\mathbf{y}^*_i} \exp \left(\sum_{l=1}^{m-1} \zeta^l_{y^*_{i,1}} + \sum_{t=2}^{T_i} \sum_{l=1}^{m-1} \omega^l_{y^*_{i,t-1},y^*_{i,t}} + \sum_{t=1}^{T_i} \Psi_{y^*_{i,t}}(\mathbf{x}_{i,t};\boldsymbol{\tau}) \right) \right\} \right] \Bigg|_{\boldsymbol{\Psi}=\widehat{\boldsymbol{\Psi}}^{m-1}} \\
&\quad= \mathbb{I}_{\{s^*\}}(y_{i^*,t^*}) - \frac{1}{Z(\mathbf{x}_{i^*};\boldsymbol{\theta})} \sum_{\{\mathbf{y}^*_{i^*}: y^*_{i^*,t^*}=s^*\}} \exp \left(\sum_{l=1}^{m-1} \zeta^l_{y^*_{i^*,1}} + \sum_{t=2}^{T_{i^*}} \sum_{l=1}^{m-1} \omega^l_{y^*_{i^*,t-1},y^*_{i^*,t}} \right. \\
&\qquad \left. + \sum_{t=1}^{T_i} \Psi_{y^*_{i^*,t}}(\mathbf{x}_{i^*,t};\boldsymbol{\tau}) \right) \Bigg|_{\boldsymbol{\Psi}=\widehat{\boldsymbol{\Psi}}^{m-1}} \\
&\quad= \mathbb{I}_{\{s^*\}}(y_{i^*,t^*}) - \widehat{p}(Y_{i^*,t^*}=s^*|\mathbf{x}_{i^*};\boldsymbol{\theta}^l, l=1,\ldots,m-1)
\end{aligned}
$$

A similar derivation is given in Dietterich et al. [2004]. This expression in terms of the estimated class membership probabilities is convenient because fast computation techniques are available to calculate these probabilities, as we will see in Section 5.6. It also offers an intuitively appealing interpretation of the boosting process: in each stage, the new regression trees are estimated using the residuals from the previous stage.

As with other boosting algorithms, it can be helpful to regularize the individual component models in order to reduce the potential to overfit the training data. Friedman [2001] and Dietterich et al. [2004] recommend achieving this by limiting the number of leaves in each tree. We follow a similar idea by limiting the depth of each tree. We use 10-fold cross-validation to choose the tree depth, selecting the depth that yields the highest proportion of time points classified correctly in the validation data sets. However, it is too computationally expensive to reproduce the entire bagging and boosting process within each fold of the cross-validation procedure. Instead, we further sub-partition the cross-validation training data sets into training and validation subsets which are passed to the boosting routine. This does not give us a completely accurate picture of how the performance of our bagging and boosting method varies with tree depth, but it offers a large reduction in computation time. Figure 20 illustrates this process.

As in the parametric boosting algorithm, we also regularize the component models by al-

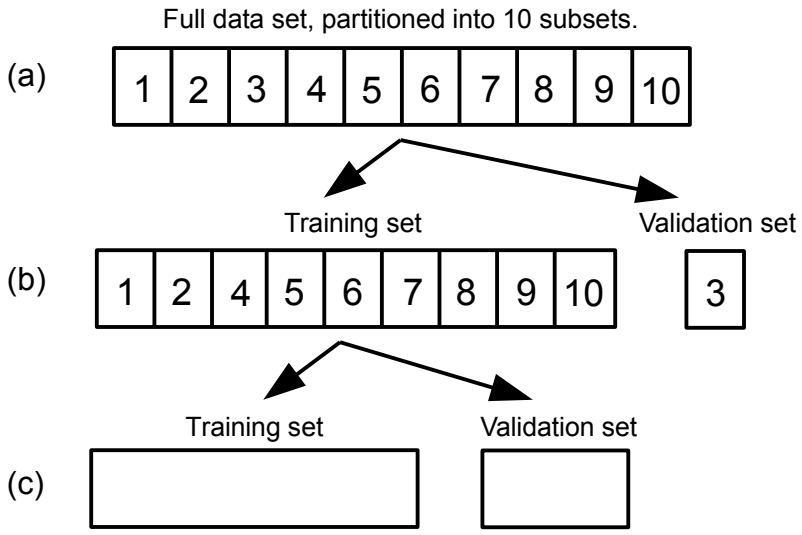


Figure 20. A diagram illustrating the cross-validation process used to select the tree depth in estimation of the BB-Nonpar-CRF model. There are two layers of cross validation: one to select the tree depth, and one to select the number of iterations performed in the boosting step when training the models. In the figure, step (a) shows the full labeled training data set, with the observation sequences partitioned into 10 disjoint subsets of approximately equal size. For each candidate value of the tree depth, we train 10 separate CRFs. In estimating the kth CRF, we hold the kth partition from step (a) out of the training data set. Step (b) shows this for $k = 3$. We evaluate the classification performance of the resulting model using the held-out observation sequences. The boosting process requires the use of a validation set to determine the number of boosting iterations. In step (c), the training data in step (b) are further partitioned into training and validation sets for the boosting procedure. A full replication of the model training procedure would use bagging at step (c) to obtain many different training and validation sets, but this is time consuming.

lowing each component model to use only a random subset of the features. In theory, we could also use cross-validation to select the number of features used in each component model. Again, this would be computationally expensive. Instead, we have simply fixed the size of the feature subsets to 3 in our applications of the method.

In addition to slowing the rate of convergence of the boosting procedure by restricting the flexibility of the regression trees in each step, this also regularizes the final model by eliminating high-order interactions between the features. Following the discussion in Chapter 6 of Hastie et al. [2009], consider the ANOVA-style decomposition of $\Psi_s(\mathbf{x}_{i,t})$:

$$\Psi_s(\mathbf{x}_{i,t}) = \alpha + \sum_{d_1} \Psi_{s,d_1}(\mathbf{x}_{i,t,d_1}) + \sum_{d_1<d_2} \Psi_{s,d_1,d_2}(\mathbf{x}_{i,t,d_1}, \mathbf{x}_{i,t,d_2}) + \cdots$$

Here α is an overall offset, Ψ_{s,d_1} is a function capturing the "main effects" of component $d_1 \in \{1, \ldots, D\}$ of the feature vector $\mathbf{x}_{i,t}$, Ψ_{s,d_1,d_2} is a function capturing pairwise interactions, and so on. Our restriction that each component model uses only three of the observed features eliminates all interactions of order 4 or higher from this decomposition. This reduces the flexibility of the overall model, and in turn its potential for overfitting the training data.

After fitting the new regression trees we estimate $\mathbf{\Omega}^m$ through constrained maximum likelihood, holding fixed the regression tree parameters as well as all parameters from earlier boosting iterations. This estimation is performed through numerical optimization of the likelihood function. The gradient calculations are the same as those we presented in Section 5.2. As with the parametric boosting algorithm in Section 5.2, we use the out of bag sample to select the number of boosting iterations to perform. The procedure is the same as we described for Algorithm 5.1, choosing the value of m that corresponds to the highest proportion of correctly classified time points in the out of bag sample.

The method we have described here for estimating Ψ_s follows the gradient tree boosting algorithm of Dietterich et al. [2004] closely, with three changes. First, we have integrated the method with bagging by fitting a separate collection of component models to each bagged data set and using the out of bag data to select the number of boosting iterations to perform. Dietterich et al. [2004] suggested using cross-validation to choose the number of boosting iterations, although they did not specify the details of how this would be done. In similar procedures, Hastie et al. [2009] suggest using a single partition of the observed data into training and validation subsets. As we discussed in Section 5.2, the resulting classifier can be sensitive to which

observation sequences are in the training and validation sets. Bagging and combining the resulting model estimates in a LOP can reduce this dependence on the selection of the validation set. A second change from the methods of Dietterich et al. [2004] is our use of a randomly selected subset of features in each component model. This offers another mechanism for regularizing the component models in addition to limiting the number of leaves in the regression trees.

Finally, Dietterich et al. [2004] based their method on a more flexible model specification in which the Ψ functions include not only $\mathbf{x}_{i,t}$, but also the values of $\mathbf{x}_{i,t^*}$ in a window of time points t^* around t and the state $y_{i,t-1}$ at the previous time point. Our primary reason for using a less flexible specification for Ψ_s is simply to put the method on a more level footing with the other approaches we will compare our methods to in Chapters 6, 7 and 8. This allows us to compare more directly how different treatments of Ψ_s affect the classification results. We note that our fundamental contributions of integrating bagging with boosting and regularizing the component models through variable subset selection could be applied equally as well with a more flexible specification of Ψ_s than we have used.

5.4 CRF Model 3: RF-CRF

Our third estimation strategy for CRFs is based on essentially the same non-parametric specification for Ψ_s that we used for the **BB-Nonpar-CRF** model in the previous Section. We restate the model here with some minor modifications to the notation to reflect changes in the estimation strategy that we will discuss below:

$$p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})}\exp\left\{\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,1})\zeta_s + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S}\mathbb{I}_{\{r\}}(y_{i,t-1})\mathbb{I}_{\{s\}}(y_{i,t})\omega_{r,s} + \sum_{t=1}^{T_i}\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,t})\Psi_s(\mathbf{x}_{i,t};\boldsymbol{\tau})\right\}, \text{ where} \tag{5.4.1}$$

$$\Psi_s(\mathbf{x}_{i,t};\boldsymbol{\tau}) = \sum_{m=1}^{M}\sum_{j=1}^{J^m}\rho_{s,j}^m\mathbb{I}_{R_j^m}(\mathbf{x}_{i,t}), \text{ and} \tag{5.4.2}$$

$$Z(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{\mathbf{y}_i^*}\exp\left\{\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,1}^*)\zeta_s + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S}\mathbb{I}_{\{r\}}(y_{i,t-1}^*)\mathbb{I}_{\{s\}}(y_{i,t}^*)\omega_{r,s} + \sum_{t=1}^{T_i}\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,t}^*)\Psi_s(\mathbf{x}_{i,t};\boldsymbol{\tau})\right\}.$$

This model statement differs in two ways from that given in Equations (5.3.1) through (5.3.3). First, we no longer have separate values of $\boldsymbol{\zeta}$ and $\boldsymbol{\Omega}$ for each component model m. Second, the regions R_j^m no longer have a subscript for the state s. This reflects the fact that in the algorithm for parameter estimation that we develop below, the same variables and split points will be used for the trees for all states within each component model. The regression constant $\rho_{s,j}^m$ in each region does still depend on the state, however.

The estimation methods we present in this Section combine the static random forest algorithm of Breiman [2001] with the CRF model. As we discussed in the literature review, random forests are static classification or regression models that have been successful in many applications. The random forest comprises a collection of classification or regression trees that are estimated in a process that includes one or more sources of randomness. The two most common methods for introducing randomness to the estimation procedure are through bagging and random selection of the variables used to split the tree nodes. Algorithms 5.3 and 5.4 give step-by-step descriptions of methods that incorporate these ideas for estimating random forests with the dynamic CRF model.

One challenge in adapting the RF estimation method for use with CRFs is how to handle bagging. The bagging procedure used in estimation of static RFs entails resampling individual observations. In our case, these individual observations are the time points within sequences of interdependent values. This poses a problem if a particular observation is not included in the bagged sample, but another observation in the same sequence is: we cannot drop the observation that was not sampled from the likelihood calculations without affecting the contribution to the likelihood from the observation that was sampled.

We propose two variations on the bagging algorithm to address this problem. In the first variation, we perform bagging by resampling complete observation sequences with replacement from the set of all observed sequences. This is the approach to bagging that we used in the previous two Sections, and which has been used in the literature previously. We refer to the method using bagging at the level of complete sequences as **RF-seq-CRF**.

In our second bagging strategy, we perform the sampling in two stages. First, we draw complete observation sequences with replacement to obtain the set $\mathcal{B}_{seq}^m$ of bagged sequences that will be used in estimating the mth component model. Second, we select a subset of the observations in each sampled sequence that will count "more heavily" toward the likelihood computation for

Algorithm 5.3. ***RF-seq-CRF*** *Estimation Algorithm*
Method: RF_seq_CRF
Inputs: *Labeled sequence data* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N\}$
Outputs: *Model parameter estimates.*

1. *Initialize* `component_models` $= \{\}$.
2. *Initialize* $\widehat{\zeta}_s = log(\frac{n_s}{n_S})$ *and* $\widehat{\omega}_{r,s} = log(\frac{n_{r,s}}{n_{S,S}})$ *for all* $r, s = 1, \ldots, S$.
3. *For* $m = 1, \ldots, M$*, repeat the following:*
   (a) *Draw a sample of* η *observation sequences with replacement from the full set of all observation sequences. Collect these sequences in* $\mathcal{B}^m_{seq}$.
   (b) *Fit a static random classification tree using all of the individual observations in* $\mathcal{B}^m_{seq}$ *as the training data set. This tree is grown using the mechanism for random selection of split variables in the algorithm of Breiman [2001], but without any additional bagging of the training data set. The tree has parameters* $(\boldsymbol{\phi}^m_{class}, \boldsymbol{\rho}^m_{class})$*, where* $\boldsymbol{\phi}^m_{class}$ *specifies the variables and split points used in splitting the tree nodes and* $\boldsymbol{\rho}^m_{class}$ *specifies the estimated class in each leaf node.*
   (c) *For each state* $s = 1, \ldots, S$*, set* $\widehat{\boldsymbol{\phi}}^m_s = \boldsymbol{\phi}^m_{class}$ *and* $\widehat{\rho}_{s,j} = \max[-100, \log\{\mathbb{I}_{\{s\}}(\rho^m_{class,j})\}] - \max[-100, \log\{\mathbb{I}_{\{S\}}(\rho^m_{class,j})\}]$.
   (d) *Set* `component_models` $=$ `component_models` $\bigcup\{(\widehat{\boldsymbol{\phi}}^m, \widehat{\boldsymbol{\rho}}^m)\}$.
4. *Repeat the following until the largest change between consecutive iterations in the class membership probabilities across all observations in the original data set based on the combined CRF with all component models is less than a specified tolerance, or a maximum number of parameter update iterations has been reached:*
   (a) *For* $m = 1, \ldots, M$*, update the values of* $\widehat{\rho}^m_{s,j}$ *assigned to each leaf node by numerically maximizing the log-likelihood based on the observation sequences in* $\mathcal{B}^m_{seq}$.
   (b) *Update* $\widehat{\Omega}$ *to the constrained local maximum likelihood estimates based on the original (unbagged) training data, holding fixed all other parameter estimates.*
5. *For each component in* `component_models`*, divide the estimates* $\widehat{\rho}^m_{s,j}$ *by* M.
6. *Return* $\widehat{\boldsymbol{\zeta}}$*,* $\widehat{\boldsymbol{\omega}}$*, and* `component_models`.

Algorithm 5.4. ***RF-CRF*** *Estimation Algorithm*
Method: RF_CRF
Inputs: *Labeled sequence data* $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N\}$
Outputs: *Model parameter estimates.*

1. *Initialize* `component_models` $= \{\}$.
2. *Initialize* $\widehat{\zeta}_s = log(\frac{n_s}{n_S})$ *and* $\widehat{\omega}_{r,s} = log(\frac{n_{r,s}}{n_{S,S}})$ *for all* $r, s = 1, \ldots, S$.
3. *For* $m = 1, \ldots, M$, *repeat the following:*
   (a) *Draw a sample of* η *observation sequences with replacement from the full set of all observation sequences. Collect these sequences in* $\mathcal{B}^m_{seq}$.
   (b) *Draw a sample of* ν *individual observations* $(y_{i,t}, \mathbf{x}_{i,t})$ *without replacement from the set of all observations included in* $\mathcal{B}^m_{seq}$. *Collect these observations in* $\mathcal{B}^m_{point}$ *and the remaining observations that were included in* $\mathcal{B}^m_{seq}$ *but not* $\mathcal{B}^m_{point}$ *in* $\mathcal{O}^m_{point}$.
   (c) *Fit a static random classification tree using* $\mathcal{B}^m_{point}$ *as the training data set. This tree is grown using the mechanism for random selection of split variables in the algorithm of Breiman [2001], but without any additional bagging of the training data set. The tree has parameters* $(\boldsymbol{\phi}^m_{class}, \boldsymbol{\rho}^m_{class})$, *where* $\boldsymbol{\phi}^m_{class}$ *specifies the variables and split points used in splitting the tree nodes and* $\boldsymbol{\rho}^m_{class}$ *specifies the estimated class in each leaf node.*
   (d) *For each state* $s = 1, \ldots, S$, *set* $\widehat{\boldsymbol{\phi}}^m_s = \boldsymbol{\phi}^m_{class}$ *and* $\widehat{\rho}_{s,j} = \max[-100, \log\{\mathbb{I}_{\{s\}}(\rho^m_{class,j})\}] - \max[-100, \log\{\mathbb{I}_{\{S\}}(\rho^m_{class,j})\}]$.
   (e) *Set* `component_models` $=$ `component_models` $\bigcup\{(\widehat{\boldsymbol{\phi}}^m, \widehat{\boldsymbol{\rho}}^m)\}$.
4. *Repeat the following until the largest change between consecutive iterations in the class membership probabilities across all observations in the original data set based on the combined CRF with all component models is less than a specified tolerance, or a maximum number of parameter update iterations has been reached:*
   (a) *For* $m = 1, \ldots, M$, *update the values of* $\widehat{\rho}^m_{s,j}$ *assigned to each leaf node by numerically maximizing a modified version of the log-likelihood based on the observation sequences in* $\mathcal{B}^m_{seq}$ *where the value of* Ψ_s *is taken to be the score from the* m*th tree if the observation is in* $\mathcal{B}^m_{point}$, *and the average score across all trees other than the* m*th if the observation is not in* $\mathcal{B}^m_{point}$.
   (b) *Update* $\widehat{\Omega}$ *to the constrained local maximum likelihood estimates based on the original (unbagged) training data, holding fixed all other parameter estimates.*
5. *For each component in* `component_models`, *divide the estimates* $\widehat{\rho}^m_{s,j}$ *by* M.
6. *Return* $\widehat{\boldsymbol{\zeta}}$, $\widehat{\boldsymbol{\omega}}$, *and* `component_models`.

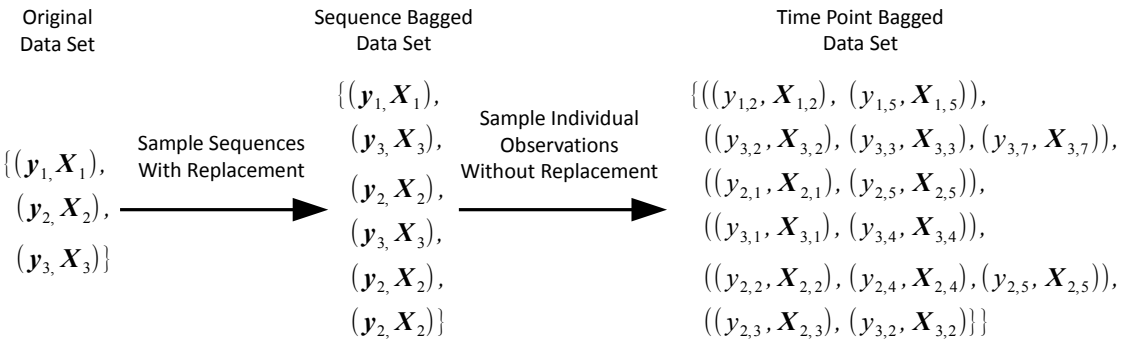


Figure 21. A diagram illustrating the two stage bootstrap sampling process for a single bagging index m in the RF-CRF model. We begin by drawing a sample of size η with replacement from the set of all observed sequences; $\eta = 6$ in this example. Within each sampled sequence, we then select a subset of the time points which will contribute the most to the likelihood for the parameters in the mth component model. The parameter ν specifies the total number of observations in the time point bagged data set, which may be distributed unequally among the bagged sequences; $\nu = 14$ in this example.

the parameters in the mth component model. We discuss our method for computing the likelihood in order to achieve this below. We collect the observations at these time points in $\mathcal{B}^m_{point}$. If a particular sequence was drawn multiple times in the first stage of sampling, each copy of that sequence in $\mathcal{B}^m_{seq}$ may yield a different set of individual observations in the second stage of sampling. This is illustrated in Figure 21. With this procedure, the final set of observations that that will contribute the most to our likelihood computations may include duplicates, since the first stage involves sampling with replacement. We refer to the model using this two stage bagging procedure as **RF-CRF**.

This sampling procedure requires us to specify the number η of observation sequences that will be drawn in the first stage and the number ν of individual time points that will be drawn in the second stage. In our applications of the method, we have used $\eta = 2 * N$ and $\nu = |\mathcal{B}^m_{seq}|/2$, where $|\mathcal{B}^m_{seq}|$ denotes the number of individual observations in $\mathcal{B}^m_{seq}$ across all sequences. We selected these values because when the number of time points in each observation sequence is balanced, the resulting marginal distribution for the number of times each observation is included in the bag is similar to the marginal distribution that arises from the usual method of sampling individual observations with replacement, which has proven effective in many classification problems. This is illustrated in Figure 22. We have not explored the effects of using other

values for η and ν or sampling from some sequences with higher probability than others.

When we use the two stage bagging procedure, we also use an altered scheme for computing the CRF likelihood so that the in-bag observations contribute more to the likelihood for the parameters than the out-of-bag observations. In order to motivate our method, consider the usual method for performing bagging with static models. After drawing the bootstrap sample $\mathcal{B}^m$ for the mth bag, we form the log-likelihood based on only the observations in the bag:

$$\ell(\boldsymbol{\tau}^m|\mathcal{B}^m) = \sum_{(y_{i,t},\mathbf{x}_{i,t})\in\mathcal{B}^m} log\{p(Y_{i,t} = y_{i,t}|\mathbf{x}_{i,t};\boldsymbol{\tau}^m)\}$$

For the purposes of estimation of the parameters for the mth tree, the out-of-bag observations do not enter the likelihood. However, we could add the terms for the out-of-bag observations back into the likelihood computation without affecting estimation of $\boldsymbol{\tau}^m$, as long as those terms depend on a different parameter vector $\boldsymbol{\tau}^* \neq \boldsymbol{\tau}^m$:

$$\ell(\boldsymbol{\tau}^m|\mathcal{B}^m) = \sum_{(y_{i,t},\mathbf{x}_{i,t})\in\mathcal{B}^m} log\{p(Y_{i,t} = y_{i,t}|\mathbf{x}_{i,t};\boldsymbol{\tau}^m\} + \sum_{(y_{i,t},\mathbf{x}_{i,t})\notin\mathcal{B}^m} log\{p(Y_{i,t} = y_{i,t}|\mathbf{x}_{i,t};\boldsymbol{\tau}^*\}$$

We apply this idea to estimation of the regression trees in our CRF specification. We estimate the parameters for the trees in the mth model by maximizing a likelihood function that is based on the observation sequences that were sampled in the first bag sampling stage. Terms corresponding to observations that were selected in the second stage of bag sampling depend on the parameters $\boldsymbol{\tau}^m$ that are currently being estimated. Terms for out-of-bag observations do not depend on $\boldsymbol{\tau}^m$; instead, we use the average value from all other component models $l \neq m$. Holding the remaining model parameters fixed, the log-likelihood we maximize in order to estimate $\boldsymbol{\tau}^m$ is as follows:

$$\begin{aligned}
\ell(\boldsymbol{\tau}^m|\mathbf{y},\mathbf{x};\boldsymbol{\zeta},\boldsymbol{\Omega},\boldsymbol{\tau}^l,l\neq m) &= \sum_{i=1}^{\eta}\log\{p(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\zeta},\boldsymbol{\Omega},\boldsymbol{\tau})\} \\
&= \sum_{i=1}^{\eta}\Bigg[\sum_{s=1}^{S}\mathbb{I}_{\{s\}}(y_{i,1})\zeta_s + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S}\mathbb{I}_{\{r\}}(y_{i,t-1})\mathbb{I}_{\{s\}}(y_{i,t})\omega_{r,s} \\
&\quad + \sum_{\{t:(y_{i,t},\mathbf{x}_{i,t})\in\mathcal{B}^m_{point}\}}\Psi_{y_{i,t}}(\mathbf{x}_{i,t};\boldsymbol{\tau}^m) + \sum_{\{t:(y_{i,t},\mathbf{x}_{i,t})\notin\mathcal{B}^m_{point}\}}\Psi_{y_{i,t}}(\mathbf{x}_{i,t};\boldsymbol{\tau}^l,l\neq m) \\
&\quad - \log\{Z(\mathbf{x}_i;\boldsymbol{\theta})\}\Bigg]
\end{aligned} \tag{5.4.3}$$

In this computation, i indexes sequences in $\mathcal{B}^m_{seq}$. We perform the computations of the terms in Z in a similar manner, so that only components corresponding to observations in $\mathcal{B}^m_{point}$ depend on $\boldsymbol{\tau}^m$.

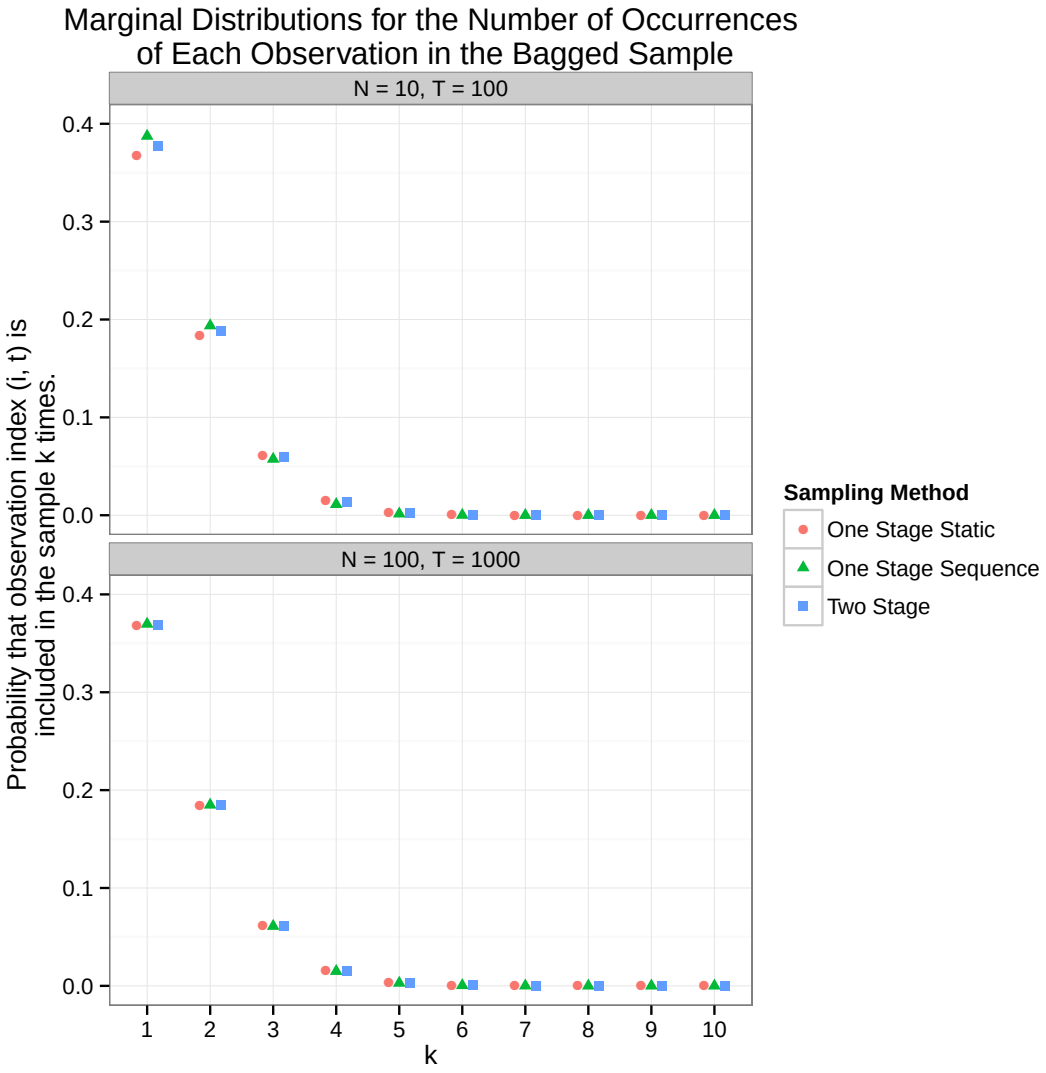


Figure 22. Marginal distributions for the number of times each individual observation is included in the bagged data set. The two stage sampling method is the method we describe in this Subsection, in which we first sample observation sequences with replacement and then select a subset of observations in each sequence that was sampled in the first stage. This distribution was calculated using a sample size of $\eta = 2 * N$ for the first stage and $\nu = \eta * T/2$ in the second stage. Note that the maximum value for k is $2 * N$; we show only the first 10 values in these plots. The one stage sampling method is the usual boostrapping method used for static models, where a sample of size $N * T$ is drawn with replacement from the set of all individual time point observations. We show the distributions for two sample sizes for the original data set. In both cases, all sequences have the same length T. Note that although the marginal distributions for the number of times each individual observation is sampled are similar, the joint distributions for the number of times all observations are sampled are different. In the two stage procedure, observations in the same sequence are more likely to be sampled together.

Unlike the example we discussed above with a static model, the parameter estimates obtained by maximizing (5.4.3) do depend on the terms corresponding to the out-of-bag observations because of the sequential dependence captured by $\boldsymbol{\Omega}$. This means that the estimate for $\boldsymbol{\tau}^m$ also depends on the parameter values for $\boldsymbol{\zeta}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\tau}^l$, $l \neq m$ used in computing the likelihood. Our solution to this is to use an iterative parameter estimation procedure, alternately updating our estimates for $\boldsymbol{\Omega}$ and $\boldsymbol{\tau}^m$, $m = 1, \ldots, M$. Although it would be possible to re-estimate the split variables and split points in the regression trees on each iteration of this parameter updating procedure, that would be very slow. Instead, we fix the split points for the trees and only update the regression constants $\boldsymbol{\rho}$. As with our earlier models, we hold $\boldsymbol{\zeta}$ fixed at its initial estimate.

We initialize this iterative process using the estimates for Ω corresponding to the maximum likelihood estimates from a HMM, which we also used for the models discussed in earlier Sections. We obtain initial estimates for the regression trees by fitting a classification random forest with our bagged data. For each component model m, the regression trees for all states use the same split variables and split points as the mth tree in the classification random forest. The initial estimates for the regression constants are $\widehat{\rho}^m_{s,j} = \max[-100, \log\{\mathbb{I}_{\{s\}}(\rho^m_{class,j})\}] - \max[-100, \log\{\mathbb{I}_{\{S\}}(\rho^m_{class,j})\}]$, where $\rho^m_{class,j}$ is the estimated class in leaf j of the classification tree. These values yield an estimated class probability close to 1 for the class that was selected by the mth classification tree. Our initial values for $\boldsymbol{\rho}$ are bounded below by -100 so that the probabilities of the other classes are non-zero; this is necessary for the gradient calculations used in updating these parameter values.

As with the estimation algorithms we introduced in Sections 5.2 and 5.3, the initial values of the parameter estimates are updated using numerical optimization routines. The gradient of the log-likelihood with respect to the elements of $\boldsymbol{\Omega}$ can be expressed in terms of the class probabilities $p(Y_{i,t-1} = r, Y_{i,t} = s|\mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta})$ just as we saw in Section 5.2. The partial derivatives with respect to $\rho^m_{s,j}$ can be rewritten in terms of the marginal class probabilities. To simplify the notation, we set

$$A(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta}) = \sum_{s=1}^{S} \mathbb{I}_{\{s\}}(y_{i,1})\zeta_s + \sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S} \mathbb{I}_{\{r\}}(y_{i,t-1})\mathbb{I}_{\{s\}}(y_{i,t})\omega_{r,s}$$
$$+ \sum_{\{t:(y_{i,t},\mathbf{x}_{i,t})\in\mathcal{B}^m_{point}\}} \Psi_{y_{i,t}}(\mathbf{x}_{i,t}; \boldsymbol{\tau}^m) + \sum_{\{t:(y_{i,t},\mathbf{x}_{i,t})\notin\mathcal{B}^m_{point}\}} \Psi_{y_{i,t}}(\mathbf{x}_{i,t}; \boldsymbol{\tau}^l, l \neq m).$$

The partial derivative of the likelihood in Equation (5.4.3) with respect to $\rho^m_{s,j}$ can then be rewrit-

ten as follows:

$$\begin{aligned}
\frac{\partial}{\partial \rho_{s,j}^m}\ell(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}) &= \sum_{i=1}^{N}\frac{\partial}{\partial \rho_{s,j}^m}\log\{p(\mathbf{Y}_i=\mathbf{y}_i|\mathbf{X}_i=\mathbf{x}_i;\boldsymbol{\theta})\} \\
&= \sum_{i=1}^{\eta}\frac{\partial}{\partial \rho_{s,j}^m}\left[A(\mathbf{x}_i,\mathbf{y}_i;\boldsymbol{\theta})-\log\{Z(\mathbf{x}_i;\boldsymbol{\theta})\}\right] \\
&= \sum_{i=1}^{\eta}\left[\sum_{t=2}^{T_i}\mathbb{I}_{\{s\}}(y_{i,t})\mathbb{I}_{R_j^m}(\mathbf{x}_{i,t})\right. \\
&\qquad \left. -\frac{1}{Z(\mathbf{x}_i;\boldsymbol{\theta})}\sum_{\mathbf{y}_i^*}\exp\{A(\mathbf{x}_i,\mathbf{y}_i^*;\boldsymbol{\theta})\}\sum_{t=2}^{T_i}\mathbb{I}_{\{s\}}(y_{i,t})\mathbb{I}_{R_j^m}(\mathbf{x}_{i,t})\right] \\
&= \sum_{i=1}^{\eta}\sum_{t=2}^{T_i}\left\{\mathbb{I}_{\{s\}}(y_{i,t})\mathbb{I}_{R_j^m}(\mathbf{x}_{i,t})-p(Y_{i,t}=s|\mathbf{X}_i=\mathbf{x}_i;\boldsymbol{\theta})\right\} \qquad (5.4.4)
\end{aligned}$$

This result is the same for the partial derivatives of the ordinary log-likelihood used in the **RF-seq-CRF** algorithm.

5.5 Classification

The previous three Sections described estimation procedures for models for the conditional distribution of the sequence of activity types $\mathbf{Y}_i$ given the feature values $\mathbf{X}_i$. These estimation procedures take as input a set of "labeled" sequences $(\mathbf{y}_i, \mathbf{x}_i)$ where the true activity type is observed. Our ultimate objective is to use the estimated models to perform classification in sequences where the activity type is not known. In the boosting algorithms of Sections 5.2 and 5.3, classification is also performed as an intermediate step in the estimation procedure.

Each set of parameter values for the CRF model determines a distribution on the space of all state sequences $\mathbf{y}_i \in \{1,\ldots,S\}^{T-i}$. However, this distribution is too complex to be of direct use in most cases. Several simpler summaries of this distribution are in common use.

In many settings, such as natural language processing, it is appropriate to summarize this distribution using the modal state sequence or the top few most probable sequences. Using these estimates ensures that the estimated classes at consecutive time points are compatible. In our setting, it is less critical that consecutive labels are compatible. Instead, we seek to maximize the expected number of windows that are labeled correctly, or equivalently, minimize the expected number of windows that are labeled incorrectly. More formally, we wish to select the values of

$\widehat{\mathbf{y}}_i$ that minimize the following loss function:

$$\begin{aligned} L(\widehat{\mathbf{y}}_i; \mathbf{x}_i) &= E_{\mathbf{Y}_i|\mathbf{x}_i} \left\{ \sum_{t=1}^{T_i} 1 - \mathbb{I}_{\{Y_{i,t}\}}(\widehat{\mathbf{y}}_{i,t}) \right\} \\ &= T_i - \sum_{t=1}^{T_i} E_{\mathbf{Y}_i|\mathbf{x}_i} \left\{ \mathbb{I}_{\{Y_{i,t}\}}(\widehat{\mathbf{y}}_{i,t}) \right\} \\ &= T_i - \sum_{t=1}^{T_i} P\left(\mathbf{Y}_{i,t} = \widehat{\mathbf{y}}_{i,t} | \mathbf{x}_i\right) \\ &\approx T_i - \sum_{t=1}^{T_i} \widehat{P}\left(\mathbf{Y}_{i,t} = \widehat{\mathbf{y}}_{i,t} | \mathbf{x}_i\right). \end{aligned} \tag{5.5.1}$$

In the last line, we have replaced the true conditional probability for the state at time t with the estimated probability based on a CRF with parameters $\boldsymbol{\theta}$. The contribution to the loss at each time t is equal to negative the estimated conditional probability that the estimated class $\widehat{y}_{i,t}$ is the true state. This term is minimized by selecting the class with the largest estimated conditional probability at time t:

$$\widehat{y}_{i,t} = \underset{s}{\text{argmax}}\, \widehat{P}\left(\mathbf{Y}_{i,t} = s | \mathbf{x}_i\right) \tag{5.5.2}$$

We will discuss methods for computing these conditional probabilities in Section 5.6.

Although we will perform classification according to Equation (5.5.2) throughout the majority of our work, other estimates may be more appropriate in some applications. For example, the set of estimated probabilities for all classes $s = 1, \ldots, S$ at each time point may be of interest, and we will make use of these quantities in Chapter 8. In other studies, researchers may be more interested an estimate of the total time spent in each physical activity category in a given time span, such as a day.

5.6 Computation

Throughout the previous four Sections, we have omitted some computational details which we now present. The estimation algorithms in Sections 5.2, 5.3, and 5.4 require repeated computation of the likelihood and its gradient. We have shown that the gradient computations can be expressed in terms of the marginal class probabilities at each time t and each pair of consecutive times $(t-1, t)$. These marginal class probabilities are also used in the classification methods we discussed in Section 5.5. We begin by demonstrating that direct computation of the likelihood

and the marginal class membership probabilities is computationally expensive. Then we present a recursion procedure that can be used to calculate the likelihood and the marginal class probabilities much more quickly than the direct computations. All of the results we discuss in this Section have been developed in the literature previously [e.g., Sutton and McCallum, 2011]; we present them here for the sake of completeness.

For the presentation in this Section, we focus on the computations for a single observation sequence. This is justified because of the assumption made in our model formulation that there is no dependence between different observation sequences. As a consequence of this assumption, the likelihood is a product of separate terms for each observation sequence and the computations of marginal class probabilities required for classification can be performed separately for each sequence. It will be helpful to rewrite the log-likelihood function for the parameters based on one observation sequence in the following general form, which applies for all of the linear chain CRF models we presented in Sections 5.2, 5.3, and 5.4:

$$L(\boldsymbol{\theta}|\mathbf{y}_i, \mathbf{x}_i) = \frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} A_1(y_{i,1}, \mathbf{x}_{i,1}) \prod_{t=2}^{T_i} A_t(y_{i,t-1}, y_{i,t}, \mathbf{x}_{i,t}), \text{ where} \tag{5.6.1}$$

$$A_1(y_{i,1}, \mathbf{x}_{i,1}) = \exp\{\zeta_{y_{i,1}} + \Psi_{y_{i,1}}(\mathbf{x}_{i,1}; \boldsymbol{\theta})\},$$

$$A_t(y_{i,t-1}, y_{i,t}, \mathbf{x}_{i,t}) = \exp\{\omega_{y_{i,t-1}, y_{i,t}} + \Psi_{y_{i,t}}(\mathbf{x}_{i,t}; \boldsymbol{\theta})\} \text{ for } t \geq 2, \text{ and}$$

$$Z(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{\mathbf{y}_i^*} A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t=2}^{T_i} A_t(y_{i,t-1}^*, y_{i,t}^*, \mathbf{x}_{i,t}).$$

In Equation (5.6.1), the numerator is straightforward to compute. However, the normalizing term $Z(\mathbf{x}_i; \boldsymbol{\theta})$ consists of a sum over all possible sequences of classes. Direct computation of this normalizing term is of order $O(T_i S^{T_i})$, which is not computationally feasible except for small values of S and T_i.

The conditional probability that $Y_{i,t} = s$ can be obtained by marginalizing over the states at all other time points:

$$\begin{aligned} P(Y_{i,t} = s|\mathbf{x}_i; \boldsymbol{\theta}) &= \sum_{\{\mathbf{y}_i^* : y_{i,t}^* = s\}} P(\mathbf{Y}_i = \mathbf{y}_i^*|\mathbf{x}_i; \boldsymbol{\theta}) \\ &= \sum_{\{\mathbf{y}_i^* : y_{i,t}^* = s\}} \left[\frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t^*=2}^{T_i} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \right]. \end{aligned} \tag{5.6.2}$$

Similarly, the conditional probability that $Y_{i,t-1} = r$ and $Y_{i,t} = s$ can be found as follows:

$$P(Y_{i,t-1} = r, Y_{i,t} = s|\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{\{\mathbf{y}_i^*: y_{i,t-1}^* = r, y_{i,t}^* = s\}} P(\mathbf{Y}_i = \mathbf{y}_i^*|\mathbf{x}_i; \boldsymbol{\theta})$$

$$= \sum_{\{\mathbf{y}_i^*: y_{i,t-1}^* = r, y_{i,t}^* = s\}} \left[\frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t^*=2}^{T_i} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \right]. \tag{5.6.3}$$

Again, it is not computationally feasible to calculate these quantities directly; now we have to contend with calculation of not only $Z(\mathbf{x}_i; \boldsymbol{\theta})$, but also the summations in the numerators.

A set of recursions has been developed in the literature that can be used to reduce the computational complexity of the likelihood and marginal class probability calculations to $O(S^2 T_i)$ for each sequence. The recursions have two steps. First we obtain the forward variables $\boldsymbol{a}_{i,t} = [a_{i,t,1}, \ldots, a_{i,t,S}]$, where

$$a_{i,t,s} = \sum_{y_{i,1:t-1}^*} \left[A_t(y_{i,t-1}^*, s, \mathbf{x}_{i,t-1}) A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t^*=2}^{t-1} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \right] \tag{5.6.4}$$

$$= \sum_{y_{i,t-1}^*} \Bigg[A_t(y_{i,t-1}^*, s, \mathbf{x}_{i,t-1}) \times \sum_{y_{i,1:t-2}^*} A_{t-1}(y_{i,t-2}^*, y_{i,t-1}^*, \mathbf{x}_{i,t-1}) A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t^*=2}^{t-2} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \Bigg] \tag{5.6.5}$$

$$= \sum_{y_{i,t-1}^*} A_t(y_{i,t-1}^*, s, \mathbf{x}_{i,t}) a_{i,t-1,y_{i,t-1}^*}.$$

The summation in line (5.6.4) is over all sequences of states for the first $t-1$ time points. In line (5.6.5), we factor that summation into two pieces: one summation over all possible values for the state at time $t-1$, and the second summation over all sequences of states for the first $t-2$ time points. We initialize these recursions by setting $a_{i,1,s} = A_1(s, \mathbf{x}_{i,1})$ for all $s = 1, \ldots, S$.

Separately, we calculate the backward variables $\mathbf{b}_{i,t} = [b_{i,t,1}, \ldots, b_{i,t,S}]$, where

$$b_{i,t,s} = \sum_{y^*_{i,t+1:T_i}} A_{t+1}(s, y^*_{i,t+1}, \mathbf{x}_{i,t+1}) \prod_{t^*=t+2}^{T_i} A_{t^*}(y^*_{i,t^*-1}, y^*_{i,t^*}, \mathbf{x}_{i,t^*}) \tag{5.6.6}$$

$$= \sum_{y^*_{i,t+1}} \Bigg[A_{t+1}(s, y^*_{i,t+1}, \mathbf{x}_{i,t+1}) \times \sum_{y^*_{i,t+2:T_i}} A_{t+2}(y^*_{i,t+1}, y^*_{i,t+2}, \mathbf{x}_{i,t+2}) \prod_{t^*=t+3}^{T_i} A_{t^*}(y^*_{i,t^*-1}, y^*_{i,t^*}, \mathbf{x}_{i,t^*}) \Bigg] \tag{5.6.7}$$

$$= \sum_{y^*_{i,t+1}} A_{t+1}(s, y^*_{i,t+1}, \mathbf{x}_{i,t+1}) b_{i,t+1,y^*_{i,t+1}}. \tag{5.6.8}$$

We initialize these recursions by setting $b_{i,T_i,s} = 1$ for all $s = 1, \ldots, S$.

These recursions have exactly the same form as those used for inference with HMMs, but the interpretation of the forward and backward variables in terms of conditional probabilities is not available in the context of CRFs. In order to avoid numerical difficulties, in practice we work with the logarithms of the forward and backward variables. Once the forward and backward variables have been calculated, we can easily compute the desired quantities. In the formula for the likelihood of Equation (5.6.1), the numerator can be calculated directly and the normalizing factor can be computed as

$$\begin{aligned} Z(\mathbf{x}_i; \boldsymbol{\theta}) &= \sum_{\mathbf{y}^*_i} A_1(y^*_{i,1}, \mathbf{x}_{i,1}) \prod_{t=2}^{T_i} A_t(y^*_{i,t-1}, y^*_{i,t}, \mathbf{x}_{i,t}) \\ &= \sum_{y^*_{i,T_i}} \left[\sum_{\mathbf{y}^*_{i,1:T_i-1}} \left\{ A_{T_i}(y^*_{i,T_i-1}, y^*_{i,T_i}, \mathbf{x}_{i,T_i}) A_1(y^*_{i,1}, \mathbf{x}_{i,1}) \prod_{t=2}^{T_i-1} A_t(y^*_{i,t-1}, y^*_{i,t}, \mathbf{x}_{i,t}) \right\} \right] \\ &= \sum_{s=1}^{S} a_{i,T_i,s}. \end{aligned}$$

For the marginal class probabilities in Equation (5.6.2), the numerator is given by

$$\begin{aligned} &\sum_{\{\mathbf{y}^*_i : y^*_{i,t} = s\}} A_1(y^*_{i,1}, \mathbf{x}_{i,1}) \prod_{t^*=2}^{T_i} A_{t^*}(y^*_{i,t^*-1}, y^*_{i,t^*}, \mathbf{x}_{i,t^*}) \\ &\quad = \sum_{\mathbf{y}^*_{i,1:t-1}} \left\{ A_t(y^*_{i,t-1}, s, \mathbf{x}_{i,t}) A_1(y^*_{i,1}, \mathbf{x}_{i,1}) \prod_{t^*=2}^{t-1} A_{t^*}(y^*_{i,t^*-1}, y^*_{i,t^*}, \mathbf{x}_{i,t^*}) \right\} \\ &\qquad \times \sum_{\mathbf{y}^*_{i,t+1:T_i}} \left\{ A_{t+1}(s, y^*_{i,t+1}, \mathbf{x}_{i,t+1}) \prod_{t^*=t+2}^{T_i} A_{t^*}(y^*_{i,t^*-1}, y^*_{i,t^*}, \mathbf{x}_{i,t^*}) \right\} \\ &\quad = a_{i,t,s}\, b_{i,t,s}. \end{aligned}$$

For the pairwise marginal class probabilities in Equation (5.6.3), we have

$$
\begin{aligned}
&\sum_{\{\mathbf{y}_i^*: y_{i,t-1}^*=r, y_{i,t}^*=s\}} A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t^*=2}^{T_i} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \\
&\quad = \sum_{\mathbf{y}_{i,1:t-2}^*} \left\{ A_{t-1}(y_{i,t-2}^*, r, \mathbf{x}_{i,t-1}) A_1(y_{i,1}^*, \mathbf{x}_{i,1}) \prod_{t^*=2}^{t-2} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \right\} \\
&\qquad \times A_t(r, s, \mathbf{x}_{i,t}) \\
&\qquad \times \sum_{\mathbf{y}_{i,t+1:T_i}^*} \left\{ A_{t+1}(s, y_{i,t+1}^*, \mathbf{x}_{i,t+1}) \prod_{t^*=t+2}^{T_i} A_{t^*}(y_{i,t^*-1}^*, y_{i,t^*}^*, \mathbf{x}_{i,t^*}) \right\} \\
&\quad = a_{i,t-1,r}\, A_t(r, s, \mathbf{x}_{i,t})\, b_{i,t,s}.
\end{aligned}
$$

CHAPTER 6

SIMULATION STUDY

6.1 Introduction

We conducted a simulation study to compare the performance of our new classification methods to that of existing methods. We describe the design of the simulation study in Section 6.2 and present the results in Section 6.3.

6.2 Methods

The objective of the simulation study we describe in this Section is to understand how the three new classification methods developed in Chapter 5 perform relative to several existing methods that are commonly used when there is sequential dependence in the class labels. The relative performance of these methods depends on the characteristics of the particular classification task at hand. In our simulation study we focus on just two of these characteristics. We begin this Section by briefly outlining the previous approaches that we compare our methods to. We then discuss the characteristics of the classification problem that vary in the simulation study design and state the model used to generate the data in the simulation study.

As discussed in the literature review, many methods have been applied to perform classification in the presence of sequential dependence. We include a representative subset of these methods in our simulation study. In addition to the three new methods we developed in Chapter 5, we evaluate the following classification methods:

1. **RF**: A static random forest. We use the implementation in the `randomForest` package [Liaw and Wiener, 2002] for `R` [R Core Team, 2013], with the default options for number of

trees, node size, and number of variables considered for each split. This is the only classifier in our simulation study that does not explicitly account for sequential dependence.

2. **Gaussian-mixt-HMM**: A HMM with a mixture of Gaussians for the observation distributions. We estimate the model parameters by maximum likelihood. For the observation distributions, we use `R`'s `mclust` package [Fraley et al., 2012] to estimate the Gaussian mixture component parameters. This package allows for a number of possible restrictions on the parameterizations of the Gaussian component covariance matrices. It uses an EM algorithm to obtain local maximum likelihood estimates of the Gaussian component parameters, and BIC to select the covariance parameterization and number of components. Before fitting the **Gaussian-mixt-HMM** model, we apply the Yeo-Johnson transformation [Yeo and Johnson, 2000] to each covariate. This is a transformation to approximate normality similar to the Box-Cox transformation. In estimating the parameters of the transformation, we work with the residuals of a regression of the observed feature values in the training data on indicators of class membership. This effectively allows the transformation to account for a class-specific mean for the feature values. We use the implementation of this transformation that is available in the `car` package [Fox and Weisberg, 2011] in `R`. Informal experiments revealed that using this data transformation led to a small improvement in the proportion of windows classified correctly by the **Gaussian-mixt-HMM** method.

3. **Par-CRF**: A parametric specification of the linear chain CRF. We use the following model:

$$p(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} \exp\left\{ \sum_{s=1}^{S} \Psi_1(s, y_{i,1}) + \sum_{t=2}^{T_i} \sum_{r=1}^{S} \sum_{s=1}^{S} \Psi_2(r, s, y_{i,t-1}, y_{i,t}) + \sum_{t=1}^{T_i} \sum_{s=1}^{S} \Psi_3(s, y_{i,t}, \mathbf{x}_{i,t}) \right\}, \text{ where} \tag{6.2.1}$$

$$\Psi_1(s, y_{i,1}) = \mathbb{I}_{\{s\}}(y_{i,1}) \zeta_s,$$

$$\Psi_2(r, s, y_{i,t-1}, y_{i,t}) = \mathbb{I}_{\{r\}}(y_{i,t-1}) \mathbb{I}_{\{s\}}(y_{i,t}) \omega_{r,s},$$

$$\Psi_3(s, y_{i,t}, \mathbf{x}_{i,t}) = \beta_{s,0} + \sum_{d=1}^{D} \beta_{s,d} x_{i,t,d}.$$

Here $Z(\mathbf{x}_i; \boldsymbol{\theta})$ is a normalizing factor. In order to resolve problems with identifiability, we fix $\zeta_S = 0$, $\omega_{S,S} = 0$, and $\beta_{S,d'} = 0 \, \forall d' = 0, \ldots, D$. We estimate the model parameters via maximum likelihood.

4. **gradient-tree-CRF**: A linear chain CRF where the component of the model specifying compatibility between the class label at a time point t and the features at that time is estimated via gradient tree boosting. This method is similar to both the gradient tree boosting method of Dietterich et al. [2004] and the **BB-Nonpar-CRF** method we developed in Chapter 5. It differs from the method of Dietterich et al. [2004] in two ways. We introduced both of these changes to their methods to facilitate comparisons with the other approaches in the simulation study. Our first change is that we use maximum likelihood to estimate the parametric forms for Ψ_1 and Ψ_2 as in Equation (6.2.1). We only use gradient tree boosting to replace the estimation of Ψ_3. This is less flexible than the implementation of Dietterich et al. [2004]; their method can be viewed as using gradient tree boosting to estimate a more general non-parametric function combining Ψ_1, Ψ_2, and Ψ_3. The second difference between our implementation of gradient tree boosting and that of Dietterich et al. [2004] is that we only allow Ψ_3 to depend on the class labels and feature values at a single time point t. Dietterich et al. [2004] allow Ψ_3 to make use of the feature values occurring within $\pm w$ time points of t. Their approach is more flexible, but means that classification at the beginning and end of the observation sequence cannot be performed easily. Our implementation allows us to perform classification at all time points, as with the other classification methods in our comparison. The **gradient-tree-CRF** method is also similar to the **BB-Nonpar-CRF** method we developed in Chapter 5, but does not use a LOP to combine CRF models obtained from bagging the training data set or regularize the regression trees by restricting the set of variables they can use in node splits. Dietterich et al. [2004] recommend using internal cross-validation to select the number of boosting iterations, but do not detail an exact procedure for this. In similar boosting algorithms, Hastie et al. [2009] suggest the use of a single partition of the data into training and validation subsets, with classification performance on the validation subset used to select the number of boosting iterations. We follow this recommendation, using the same procedure we described for the **BB-Nonpar-CRF** method in Chapter 5 to select the stopping point. In order to evaluate the extent to which performance of the method depends on how the data are partitioned when forming these training and validation subsets, we partitioned the training data into 10 disjoint subsets and fit 10 models, holding out one subset at a time for use as the validation data. We report the median score and the worst score among the ten resulting model fits un-

der the headings of **gradient-tree-CRF-median** and **gradient-tree-CRF-worst** respectively. This gives an indication of the central tendency and worst case performance of the method with respect to how the data are partitioned into training and validation sets.

5. **MLR-HMM** and **RF-HMM**: McShane et al. [2013] propose a method for combining a static classification model that does not account for sequential dependence with a HMM. We implement this idea using multinomial logistic regression (MLR) or a random forest (RF) for the static classification model and a first-order HMM.

We examine how the performance of the above classification methods varies as a function of the following two characteristics of the classification task:

1. whether the distributions for the feature vectors at each time point are relatively simple or more complex, and

2. the Bayes error rate.

There are many other characteristics of classification problems that likely affect the performance of the methods under consideration, such as the sample size, the dimension of the observed feature vectors at each time point, the number of classes, the relative frequencies of each class, the frequency of mislabeled observations in the training data, and so on. We focus on the complexity of the emission distributions and the Bayes error rate because we believe they are the most useful in helping to explain variation in the performance of classification methods when applied to real physical activity data.

We use a factorial design to study the relationship between the problem characteristics listed above and the performance of the classification methods. For each combination of factor level settings in the design, we conduct 50 simulations with training and test data sets generated with parameter values specific to that cell of the design. Each training and test data set is generated independently, and consists of $N = 50$ sequences of length $T = 200$. We fix the number of classes to $S = 3$ and the dimension of the observed feature vectors to $D = 50$.

For each combination of factor levels, the data $(\mathbf{y}_i, \mathbf{x}_i) \in \{1, \ldots, S\}^T \times \mathbb{R}^{D \cdot T}$, $i = 1, \ldots, N$, are

generated from a first-order HMM as follows:

$$p(Y_{i,1} = s|\boldsymbol{\pi}) = \pi_s \text{ for } s \in \mathcal{S}, \text{ where } 0 \leq \pi_s \leq 1 \, \forall s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} \pi_s = 1, \tag{6.2.2}$$

$$\begin{aligned} p(Y_{i,t+1} = y_{i,t+1}|Y_{i,1:t} = y_{i,1:t}; Q) &= p(Y_{i,t+1} = y_{i,t+1}|Y_{i,t} = y_{i,t}; Q) \\ &= q_{y_{i,t}, y_{i,t+1}}, \text{ where } Q = [q_{r,s}], \, 0 \leq q_{r,s} \leq 1 \, \forall r, s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} q_{r,s} = 1, \end{aligned} \tag{6.2.3}$$

$$\begin{aligned} f(\mathbf{x}_{i,t}|Y_{i,t} = s) = \sum_{m=1}^{M_s} w_{s,m} f_{s,m}(\mathbf{x}_{i,t}; \boldsymbol{\theta}_{s,m}), \text{ where} \\ 0 \leq w_{s,m} \leq 1 \, \forall m \text{ and } \sum_{m=1}^{M_s} w_{s,m} = 1. \end{aligned} \tag{6.2.4}$$

The form of the distribution $f_{s,m}(\mathbf{x}_{i,t}; \boldsymbol{\theta}_{s,m})$ depends on the complexity level of the emission distributions.

For the parameters governing transitions among the classes, we set

$$\boldsymbol{\pi} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \text{ and}$$

$$Q = \begin{bmatrix} \frac{4}{5} & \frac{1}{10} & \frac{1}{10} \\ \frac{1}{10} & \frac{4}{5} & \frac{1}{10} \\ \frac{1}{10} & \frac{1}{10} & \frac{4}{5} \end{bmatrix}.$$

With these parameter values, the stationary distribution of the Markov chain for the class at each time point is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

For the distributions of the observed features we partition the feature vector, which has dimension $D = 50$, into several smaller groups and specify the distribution for each of these groups. We partition the feature vector as follows:

- $(X_{i,t,1}, X_{i,t,2}, X_{i,t,3})'$ are dependent on each other, and are independent of all other $X_{i,t,d}$, $d \leq 20$.
- $(X_{i,t,4}, X_{i,t,5}, X_{i,t,6})'$ are dependent on each other, and are independent of all other $X_{i,t,d}$, $d \leq 20$.
- $(X_{i,t,7}, X_{i,t,8})'$ are dependent on each other, and are independent of all other $X_{i,t,d}$, $d \leq 20$.
- $(X_{i,t,9}, X_{i,t,10})'$ are dependent on each other, and are independent of all other $X_{i,t,d}$, $d \leq 20$.
- $(X_{i,t,11}, \ldots, X_{i,t,20})'$ follow the same distribution as $(X_{i,t,1}, \ldots, X_{i,t,10})'$, but are independent of $(X_{i,t,1}, \ldots, X_{i,t,10})'$.

- $(X_{i,t,21}, \ldots, X_{i,t,50})' = A(X_{i,t,11}, \ldots, X_{i,t,20})' + \varepsilon$, where $\varepsilon \sim Gaussian(\underline{0}, I)$ and A is a fixed matrix.

In the cases with simple emission distributions, each smaller group follows a mixture of Gaussian distributions. Thus, in those cases the data are sampled from the **Gaussian-mixt-HMM** model with specified parameter values. In the cases with more complex emission distributions, we draw each feature from a location family of a Gamma distribution where the location, shape, and scale parameters are all obtained as a linear combination of the draws for the lower dimensions. We illustrate these distributions for the first 10 dimensions in the cases with low Bayes error rate in Figures 23 through 30. The distributions for the cases with high Bayes error rate are similar, but have less separation between the distributions associated with different classes.

We computed three statistics summarizing the performance of the classification methods in each trial of the simulation study. First, we computed the proportion of time points classified correctly. This quantity varies between 0 and 1, with 1 indicating that all time points were classified correctly. Second, we computed the macro F_1 score [Sokolova and Lapalme, 2009]. The macro F_1 score is defined as the mean of the F_1 scores for each individual class. This statistic combines information about the precision and recall for all classes. It varies between 0 and 1, with 1 indicating perfect classification performance. The third summary statistic we computed is the MSE of the class probability estimates relative to the indicators of the true class labels. More formally, we define

$$MSE = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{s=1}^{S} \left\{ \widehat{P}(Y_{i,t} = s|\mathbf{x}_i) - \mathbb{I}_{\{s\}}(y_{i,t}) \right\}^2 / T_i,$$

where $\widehat{P}(Y_{i,t} = s|\mathbf{x}_i)$ is the estimated probability that the state at time t is s given the observed features at all times. This measure is bounded below by 0; a value of 0 indicates that all of the estimated class probabilities matched the true class memberships exactly.

6.3 Results and Discussion

We now present the results of our simulation study. Figures 31 through 33 display box plots summarizing the results. As expected, in every case the **Bayes-Rule** offers the best classification performance as measured by the proportion of time points classified correctly, the macro F_1

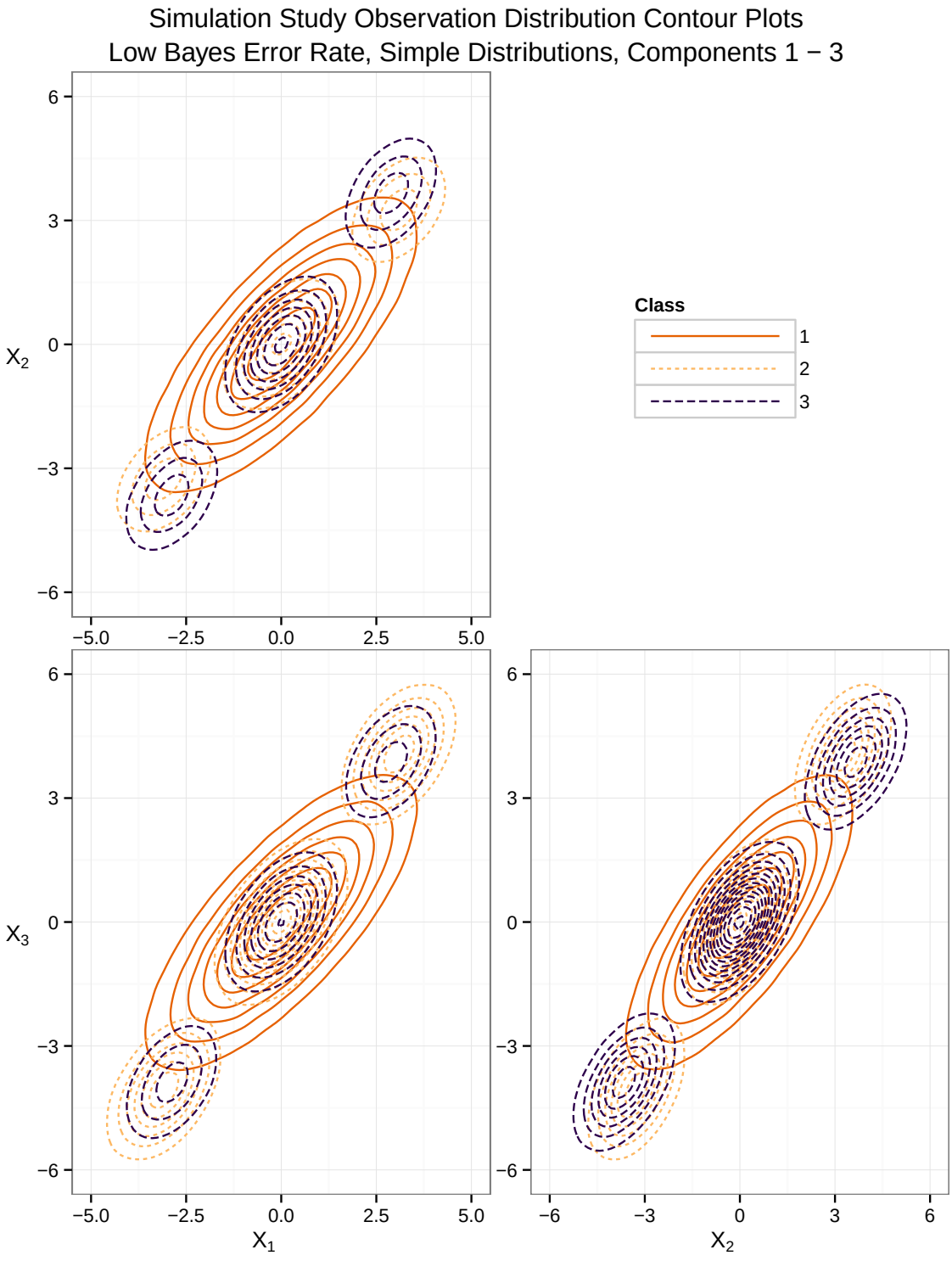


Figure 23. Pairwise contour plots of components 1 through 3 of the observation distributions used in the simulation study for the case with simple observation distributions and low Bayes error rate.

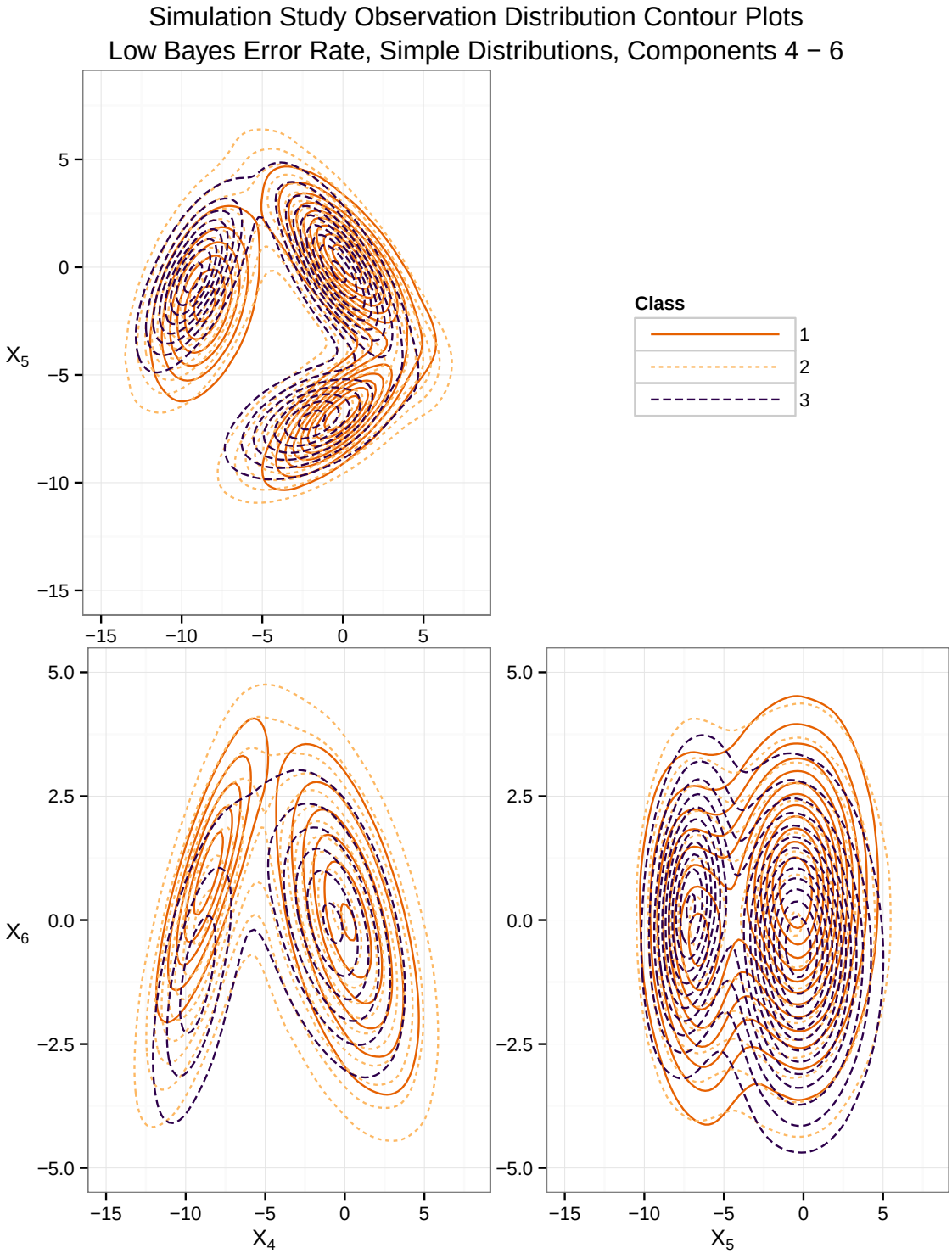


Figure 24. Pairwise contour plots of components 4 through 6 of the observation distributions used in the simulation study for the case with simple observation distributions and low Bayes error rate.

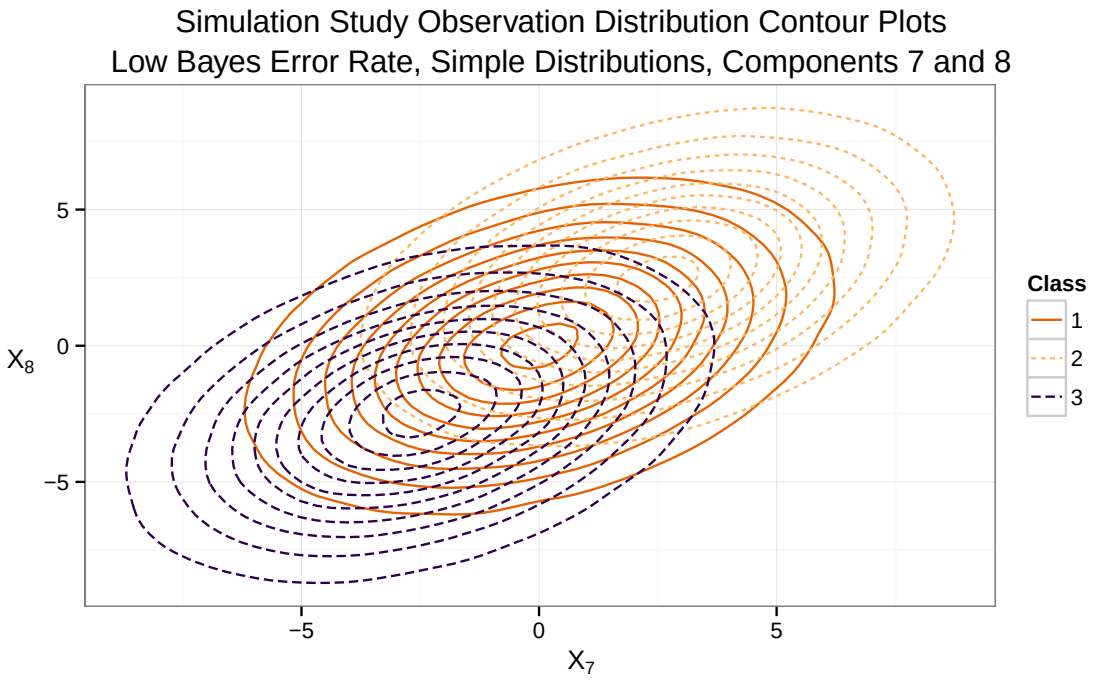


Figure 25. Pairwise contour plots of components 7 and 8 of the observation distributions used in the simulation study for the case with simple observation distributions and low Bayes error rate.

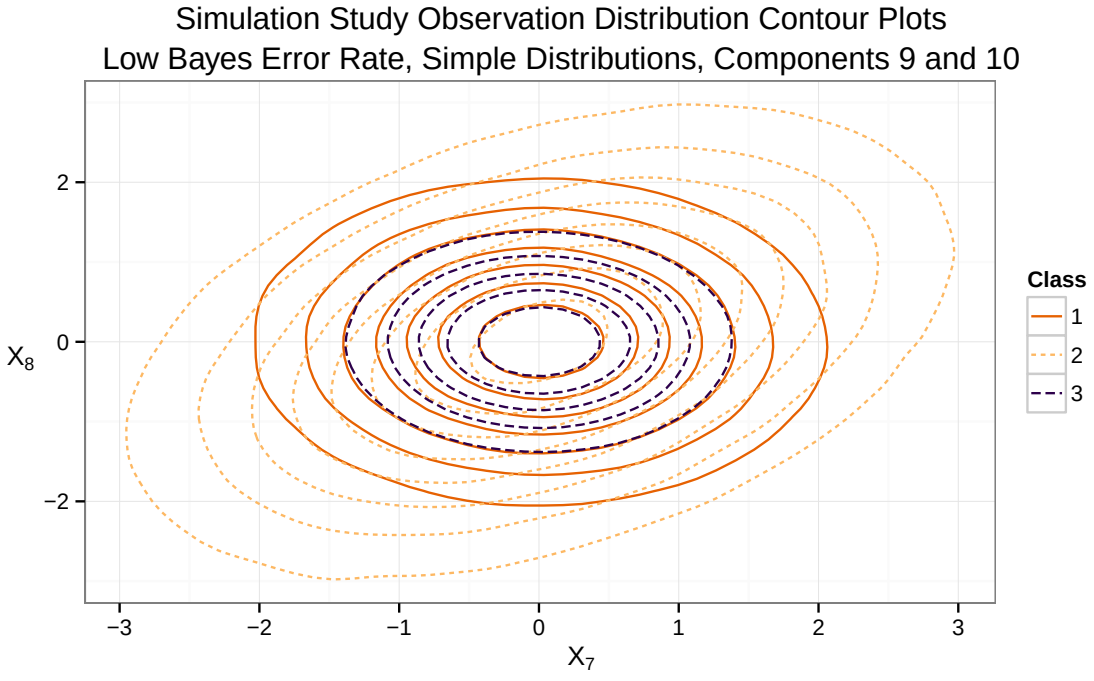


Figure 26. Pairwise contour plots of components 9 and 10 of the observation distributions used in the simulation study for the case with simple observation distributions and low Bayes error rate.

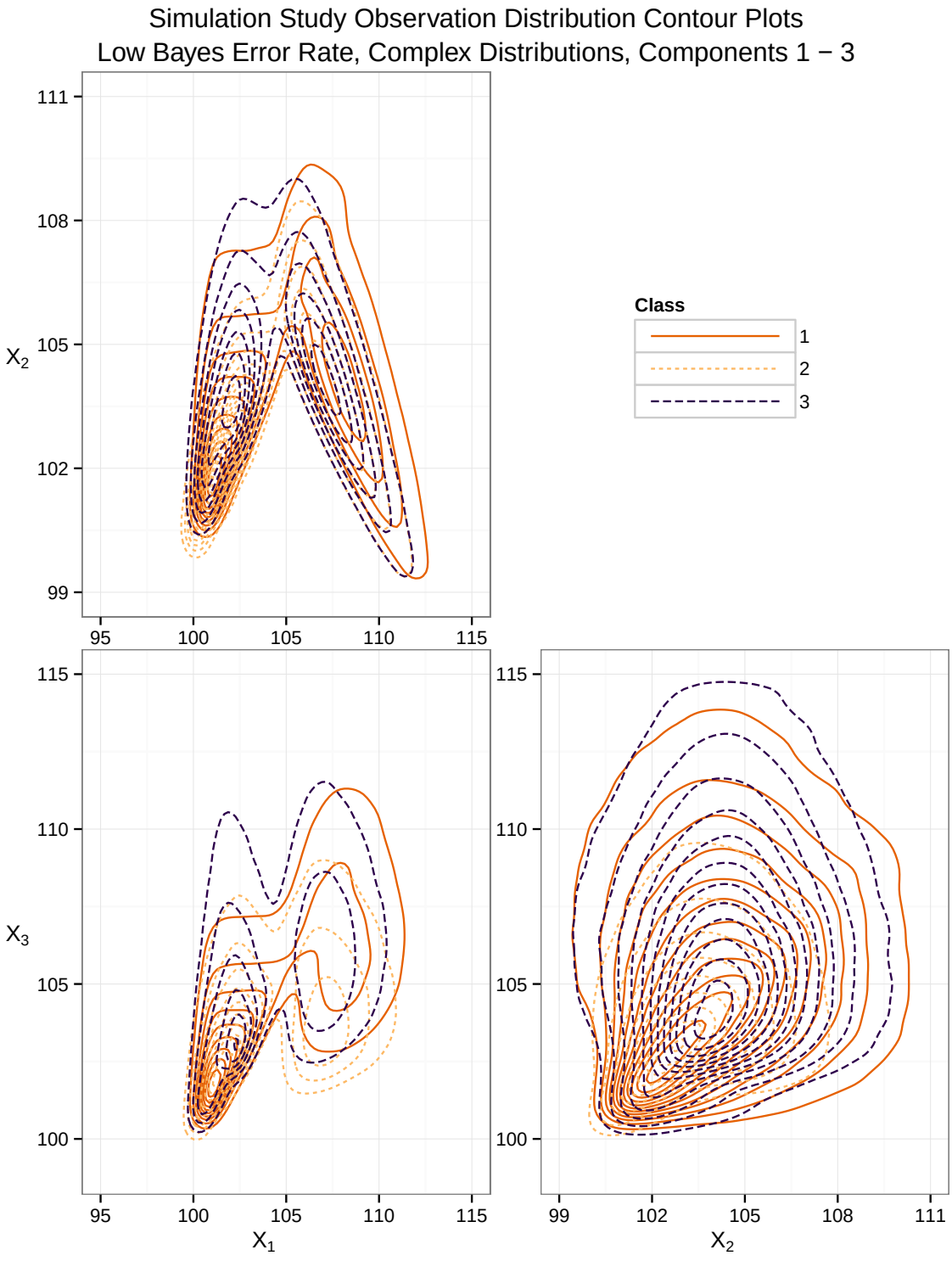


Figure 27. Pairwise contour plots of components 1 through 3 of the observation distributions used in the simulation study for the case with complex observation distributions and low Bayes error rate.

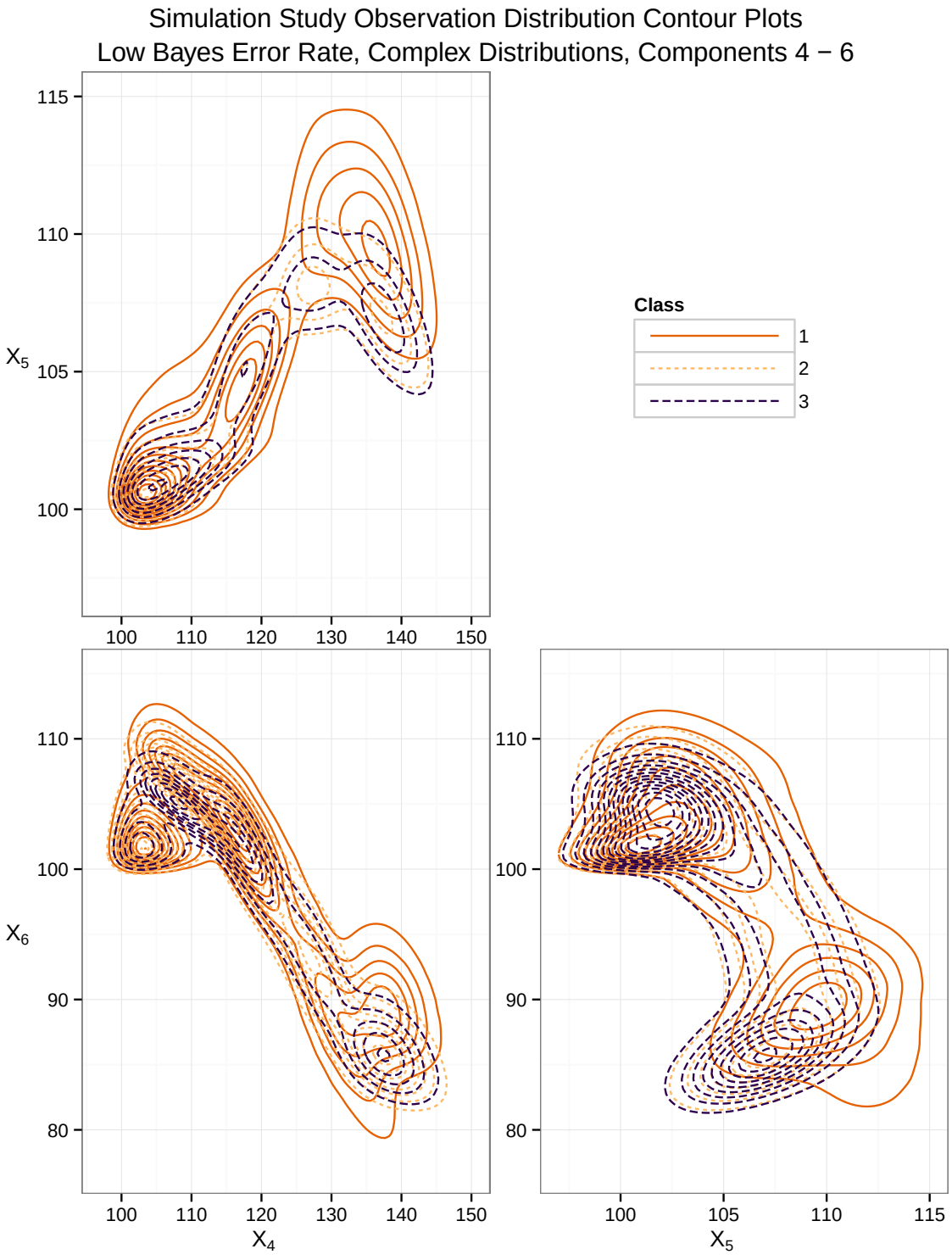


Figure 28. Pairwise contour plots of components 4 through 6 of the observation distributions used in the simulation study for the case with complex observation distributions and low Bayes error rate.

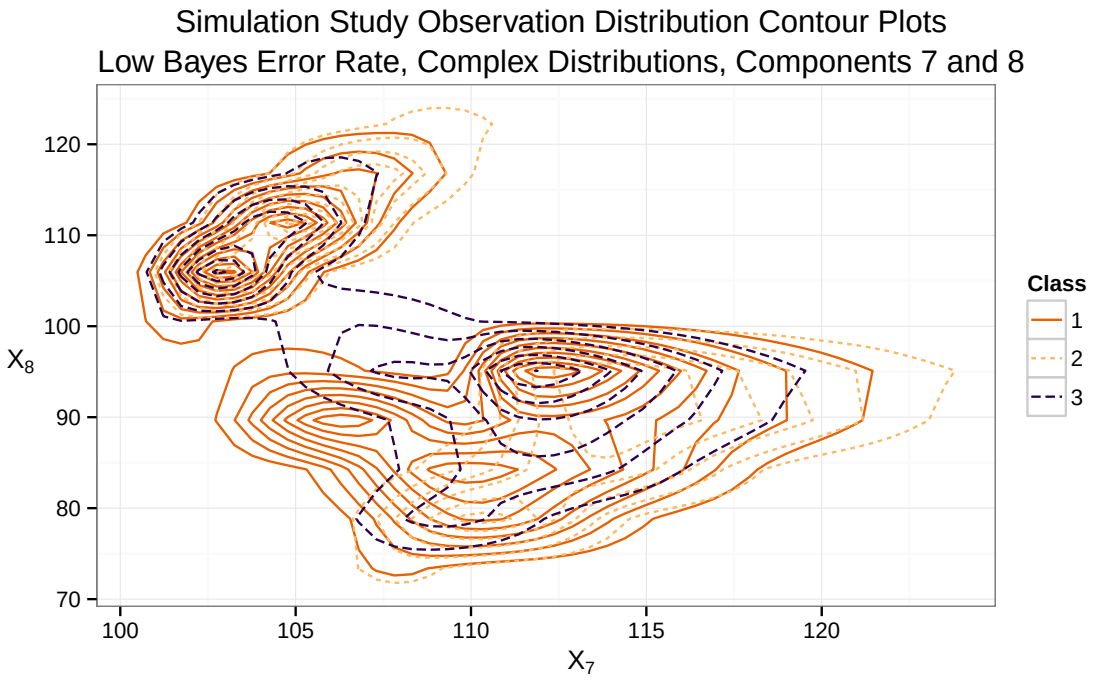


Figure 29. Pairwise contour plots of components 7 and 8 of the observation distributions used in the simulation study for the case with complex observation distributions and low Bayes error rate.

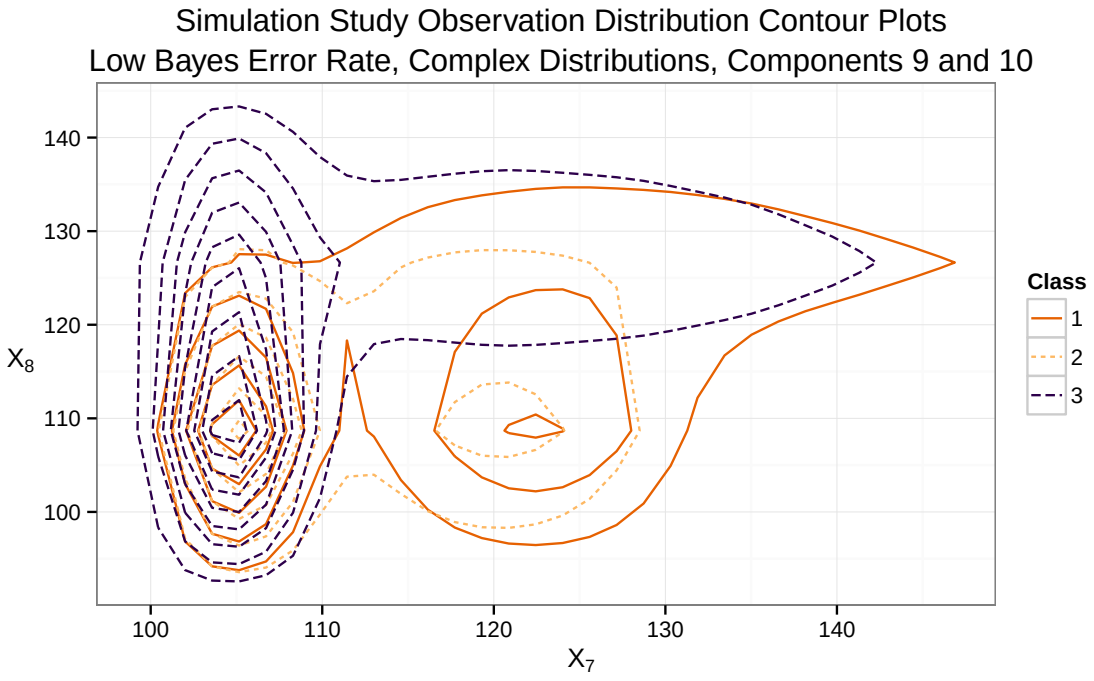


Figure 30. Pairwise contour plots of components 9 and 10 of the observation distributions used in the simulation study for the case with complex observation distributions and low Bayes error rate.

score, and the MSE of the estimated class probabilities relative to the true class membership indicators.

When the data are drawn from the **Gaussian-mixt-HMM** model as in the cases with simple observation distributions, the **Gaussian-mixt-HMM** model offers the best classification performance as measured by the proportion of time points classified correctly and the macro F_1 score. The **RF-HMM** has a lower MSE when the Bayes error rate is high, but the **Gaussian-mixt-HMM** still does fairly well by this measure. However, when the emission distribution is more complex, the **Gaussian-mixt-HMM** method has close to the worst performance among all of the methods we have compared. In the case with complex emission distributions and a low Bayes error rate, only the static **RF** method has a lower proportion correct or macro F_1 score, or a higher MSE than the **Gaussian-mixt-HMM**. In the case with complex emission distributions and a high Bayes error rate, the only alternative method with a higher MSE than the **Gaussian-mixt-HMM** is the **RF-seq-CRF**, and the only method with a lower proportion correct or macro F_1 score than the **Gaussian-mixt-HMM** is the worst-case performance of the **gradient-tree-CRF** method with respect to the partition of the data into training and validation subsets.

Together, these results confirm the general results we discussed the literature review regarding the relative performance of generative methods that specify a joint model for the class labels and features and discriminative methods that specify a conditional model for the class given the features. In short, the relative performance of these methods depends on how accurate the joint model for the class labels and the features specified by the generative model is. When the joint model is accurate, the generative approach is superior; when the joint model is misspecified, the discriminative approach is superior. Of course, the picture is a little more complex than this, and our simulation study has not examined all of the relevant factors. Misspecification is not an all or nothing property of a model; instead, a model may be misspecified to varying extents. For instance, it is likely that if the feature emission distributions exhibited only minor departures from a mixture of Gaussians with a small number of components, the **Gaussian-mixt-HMM** would still offer superior performance relative to the CRF-based models. With flexible models such as a mixture of Gaussians, another important factor is likely the size of the sample relative to the dimension of the feature vector. If the sample size were large in relation to the dimension, the **Gaussian-mixt-HMM** might not suffer as much in settings where the feature emission distributions are complex, since the method would be able to add more Gaussian components to

the mixture and thereby model the emission distributions more accurately. We believe that the settings in our simulation study with complex emission distributions are a reasonably accurate representation of the conditions that we face with real physical activity data, and indicate that methods that condition on the accelerometer features may have an advantage over methods that attempt to model them.

Our simulation study also offers insights into the performance of the static **RF** method. When the Bayes error rate is low, so that most time points are classified correctly, the **RF** method has the worst classification performance as measured by the proportion of time points classified correctly and the macro F_1 score, and close to the worst performance as measured by the MSE. However, when the Bayes error rate is large, so that more time points are classified incorrectly, the **RF** method actually outperforms several of the dynamic methods and the difference from the other dynamic methods is reduced. One possible explanation explanation for this behavior is that when the Bayes error rate is large, we are less certain about the estimated class at each individual time point. Thus, dynamic methods benefit less from borrowing information about the class at nearby times. However, even in this case several of the dynamic methods outperform the **RF**.

Among the remaining methods, all of which are dynamic models with discriminative estimation of at least some of the model parameters, the **RF-HMM** consistently achieves the best classification performance. Performance of the remaining methods is less consistent. For example, the **RF-CRF** method does better than the **BB-Nonpar-CRF** method in the cases with a high Bayes error rate, worse than the **BB-Nonpar-CRF** method in the case with low Bayes error rate and simple feature emission distributions, and about the same as the **BB-Nonpar-CRF** method in the case with low Bayes error rate and complex feature emission distributions.

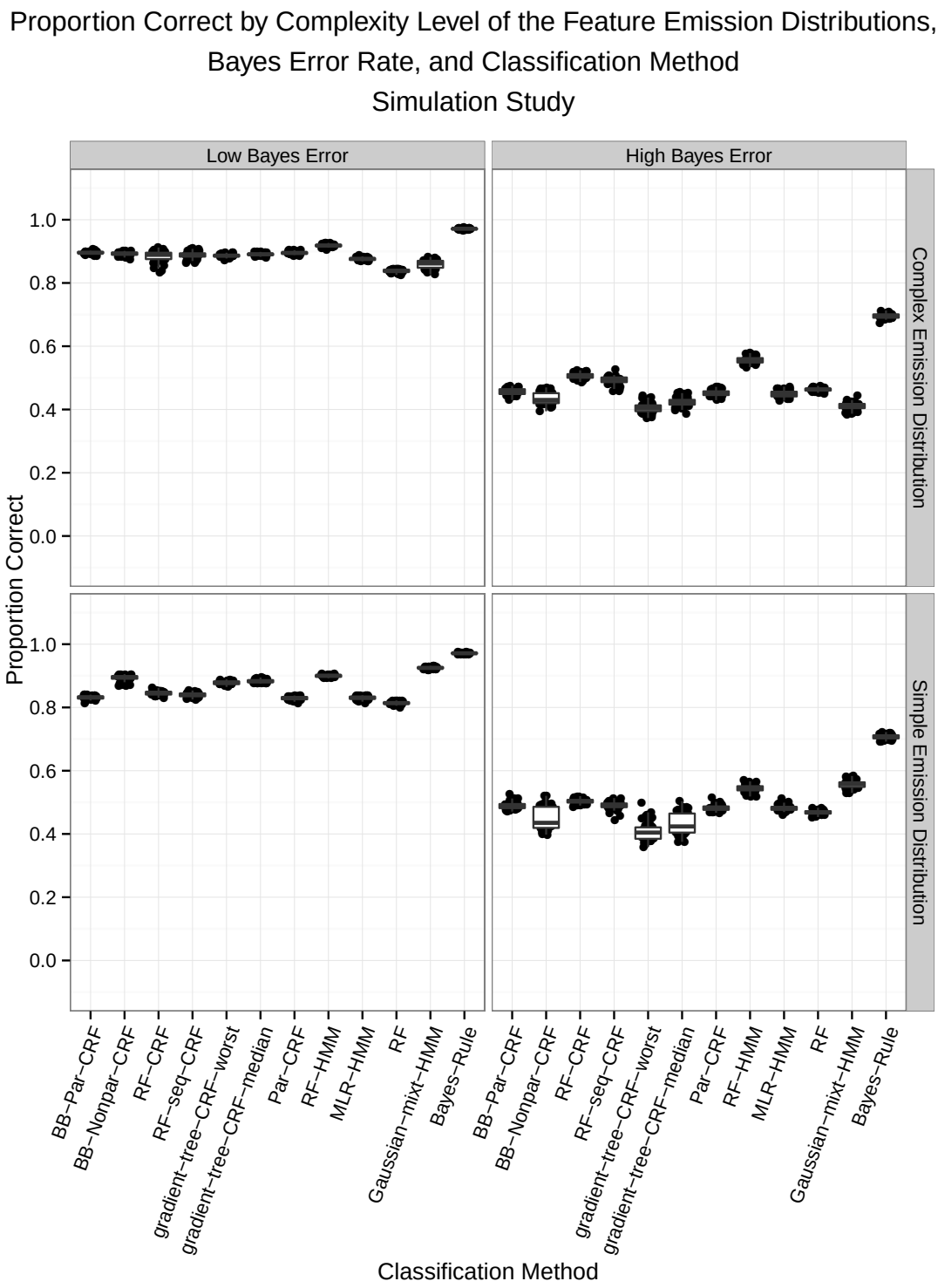


Figure 31. Box plots showing the proportion of time points classified correctly in the simulation study. A separate box plot is displayed for each combination of the complexity level of the feature emission distributions, the Bayes error rate, and the classification method. Each point corresponds to a combination of distribution complexity, Bayes error rate, classification method, and simulation index.

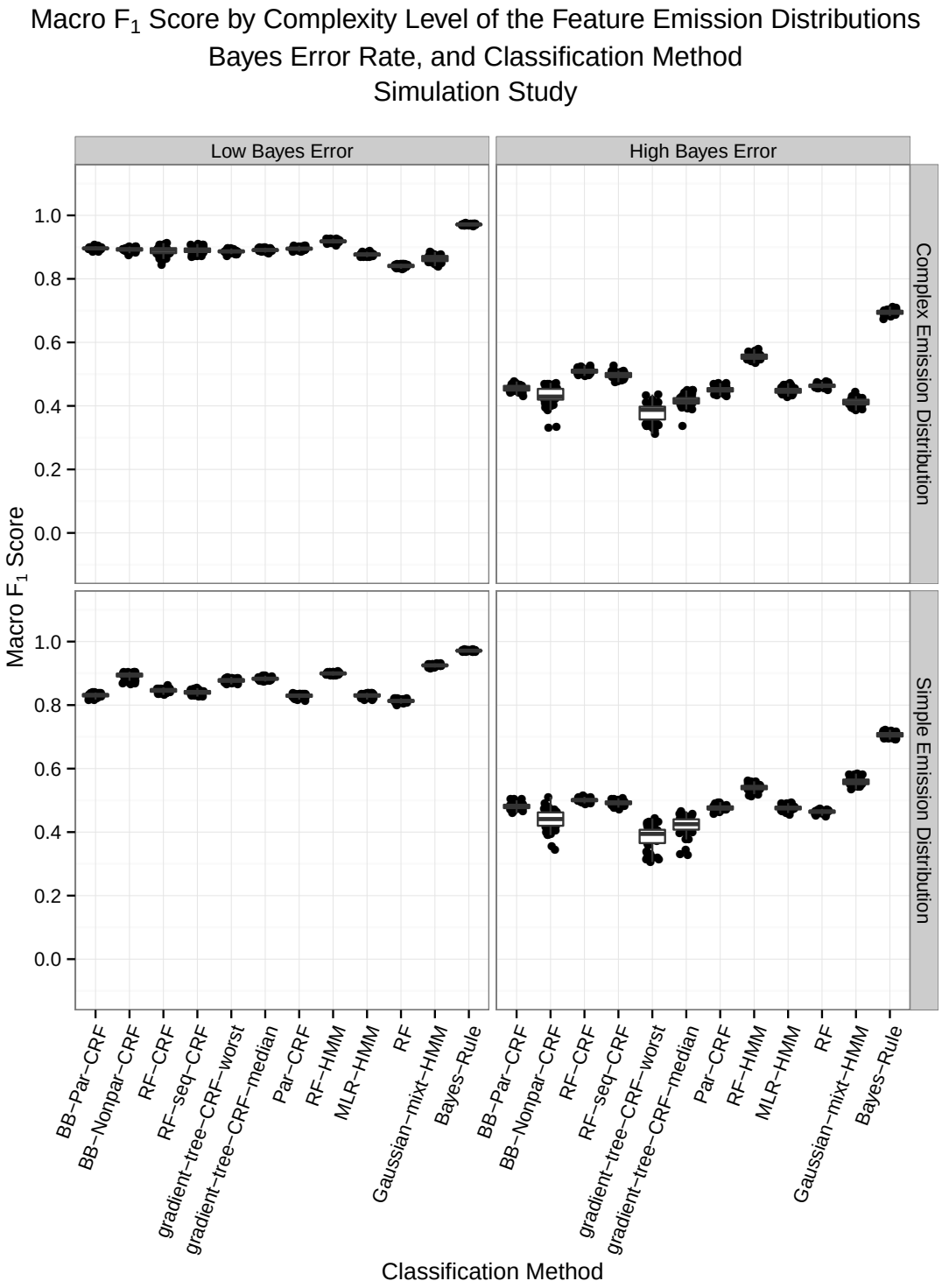


Figure 32. Box plots showing the macro F_1 score combining precision and recall across all three classes in the simulation study. A separate box plot is displayed for each combination of the complexity level of the feature emission distributions, the Bayes error rate, and the classification method. Each point corresponds to a combination of distribution complexity, Bayes error rate, classification method, and simulation index.

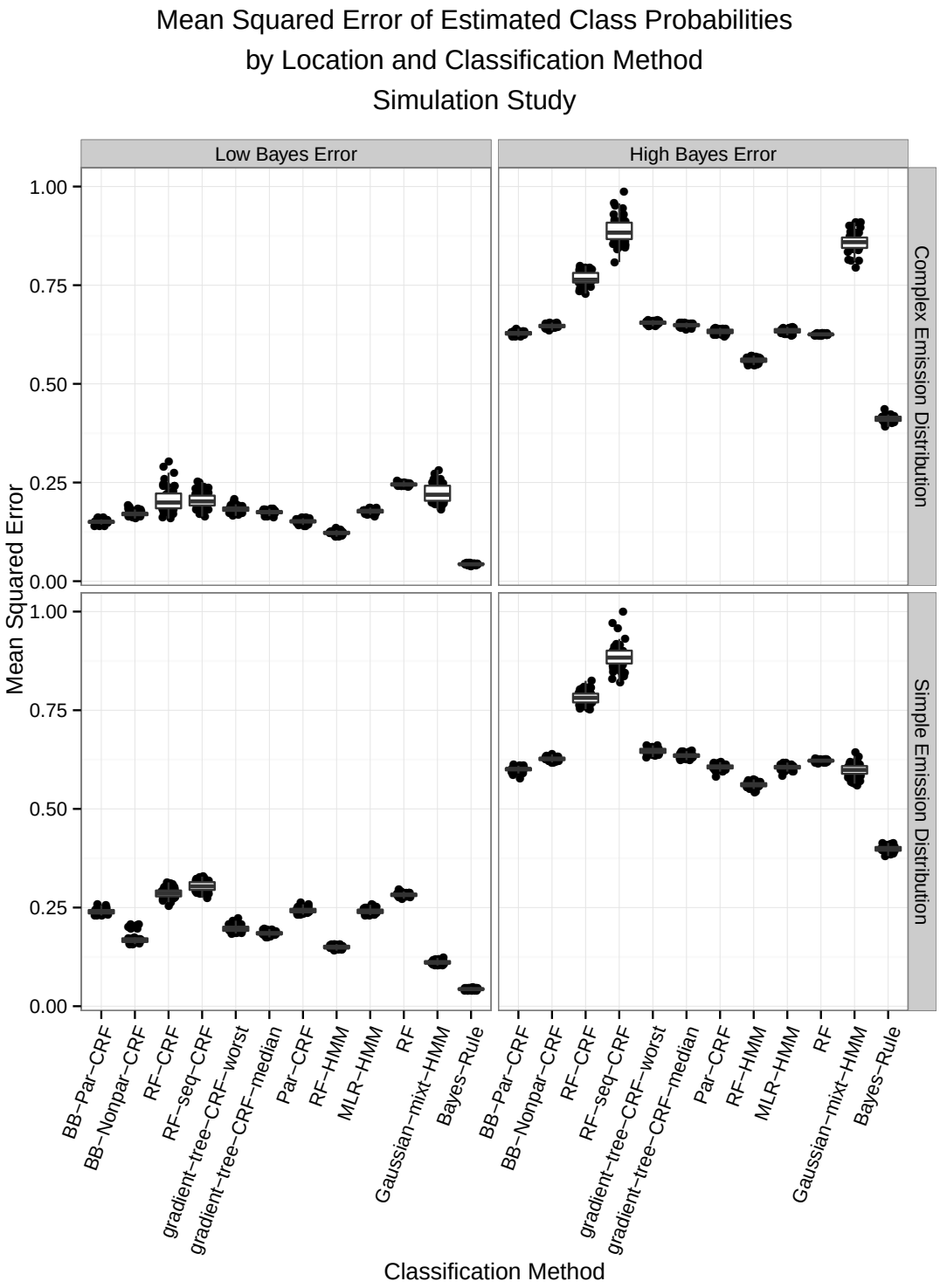


Figure 33. Box plots showing the mean squared error of the estimated class membership probabilities at each time point relative to the true class memberships in the simulation study. A separate box plot is displayed for each combination of the complexity level of the feature emission distributions, the Bayes error rate, and the classification method. Each point corresponds to a combination of distribution complexity, Bayes error rate, classification method, and simulation index.

CHAPTER 7

APPLICATIONS TO PHYSICAL ACTIVITY TYPE CLASSIFICATION

7.1 Introduction

In this Chapter, we apply the classification methods we developed in Chapter 5 and several competitor methods to classification of physical activity type in the three physical activity data sets we described in Chapter 3. We discuss our methods in Section 7.2, present the classification results for each of our three data sets in Sections 7.3 through 7.5, and conclude with a summarizing discussion in Section 7.6.

7.2 Methods

We gave a detailed description of the procedures we used to pre-process the data in Chapter 3. As a reminder, for each of our three data sets, the following factors describe how the data were gathered and preprocessed:

- **accelerometer location**: In the study by Mannini et al. [2013], accelerometers were placed on the subjects' dominant ankle and wrist. In the study by Sasaki [2013], accelerometers were placed on the subjects' dominant ankle, hip, and wrist.
- **number of categories**: In the data from Mannini et al. [2013], activities are grouped into 4 categories not including transitions; the 12.8 seconds before and after each transition were discarded. In the data from Sasaki [2013], we have two groupings of activities: one with 4 classes, and one with 6 classes. The activities that subjects performed are grouped

into either 3 or 5 "primary" activity categories, plus an additional "transition" category assigned to windows that included more than one activity type.

Our goal is to understand how classification performance depends on these factors and the classification method used. We apply the same classification methods we used in the simulation study of Chapter 6 to classify the subjects' physical activity type in each time window. These methods are as follows:

1. **BB-Par-CRF**
2. **BB-Nonpar-CRF**
3. **RF-CRF**
4. **RF-seq-CRF**
5. **RF**
6. **RF-HMM**
7. **MLR-HMM**
8. **Gaussian-mixt-HMM**
9. **gradient-tree-CRF**
10. **Par-CRF**

In addition, we applied a support vector machine (**SVM**) to the data from Mannini et al. [2013] since this is the classifier they used.

The study by Mannini et al. [2013] included $N = 33$ subjects. The laboratory component of the study conducted by Sasaki [2013] included $N = 35$ subjects, although we have ankle data for only 34 subjects. The free living component of the study included $N = 15$ subjects.

For each combination of classifier and the factors describing the data gathering process, we used a leave-one-subject-out design to gauge the success of the classifier: for each subject, we train the classification method with the observed data for all of the other subjects and use the resulting trained model to predict the activity type in each window for the held-out subject. We summarize these predictions with the same three statistics that we used in the simulation study: the proportion correct, the macro F_1 score, and the mean squared error of the estimated

class probabilities relative to the labeled class. We do not calculate the MSE statistic for the **SVM** method since that method does not directly yield estimates of the class membership probablities.

The **gradient-tree-CRF** method uses a partition of the observation sequences into a training set and a validation set used to select the number of boosting iterations. As in the simulation study, we partitioned the training data into 10 disjoint subsets and fit 10 models, holding out one subset at a time for use as the validation data in the boosting step. This partitioning in the model estimation process occurs after the first stage of partitioning for the leave one subject out cross validation used to evaluate the model performance. We report the median score and the worst score among the ten resulting model fits under the headings of **gradient-tree-CRF-median** and **gradient-tree-CRF-worst** respectively. This gives an indication of the central tendency and worst case performance of the method with respect to how the data are partitioned into training and validation sets.

7.3 Mannini *et al.* Data

We begin with a discussion of plots summarizing the results of the application to the data from Mannini et al. [2013] before turning to a description using linear mixed effects models. Figures 34 through 36 display box plots summarizing the results. We can draw several conclusions from these plots. First, we see that all of the classification methods perform better when the accelerometer is placed on the ankle than when it is on the wrist. Second, we can see that when the accelerometer is placed on the ankle, the static **RF** and **SVM** models do not perform as well as the models that account for temporal dependence. It is difficult to distinguish much of a difference between the performance of the remaining models based on the results from the ankle data. When the accelereometer is on the wrist, the **RF**, **SVM**, and **MLR-HMM** models do not perform as well as the other models. It also seems that the dynamic methods based on random forests, **RF-CRF**, **RF-seq-CRF**, and **RF-HMM**, offer slightly better performance than the other dynamic models. The performance of the **gradient-tree-CRF** method varies with the choice of which observation sequences are included in the training and validation sets; this effect is particularly strong when using the wrist data. The median performance level is similar to the other dynamic methods, but the worst case performance is worse. All of these findings are consistent whether we look at the proportion correct, the macro F_1 score, or the MSE.

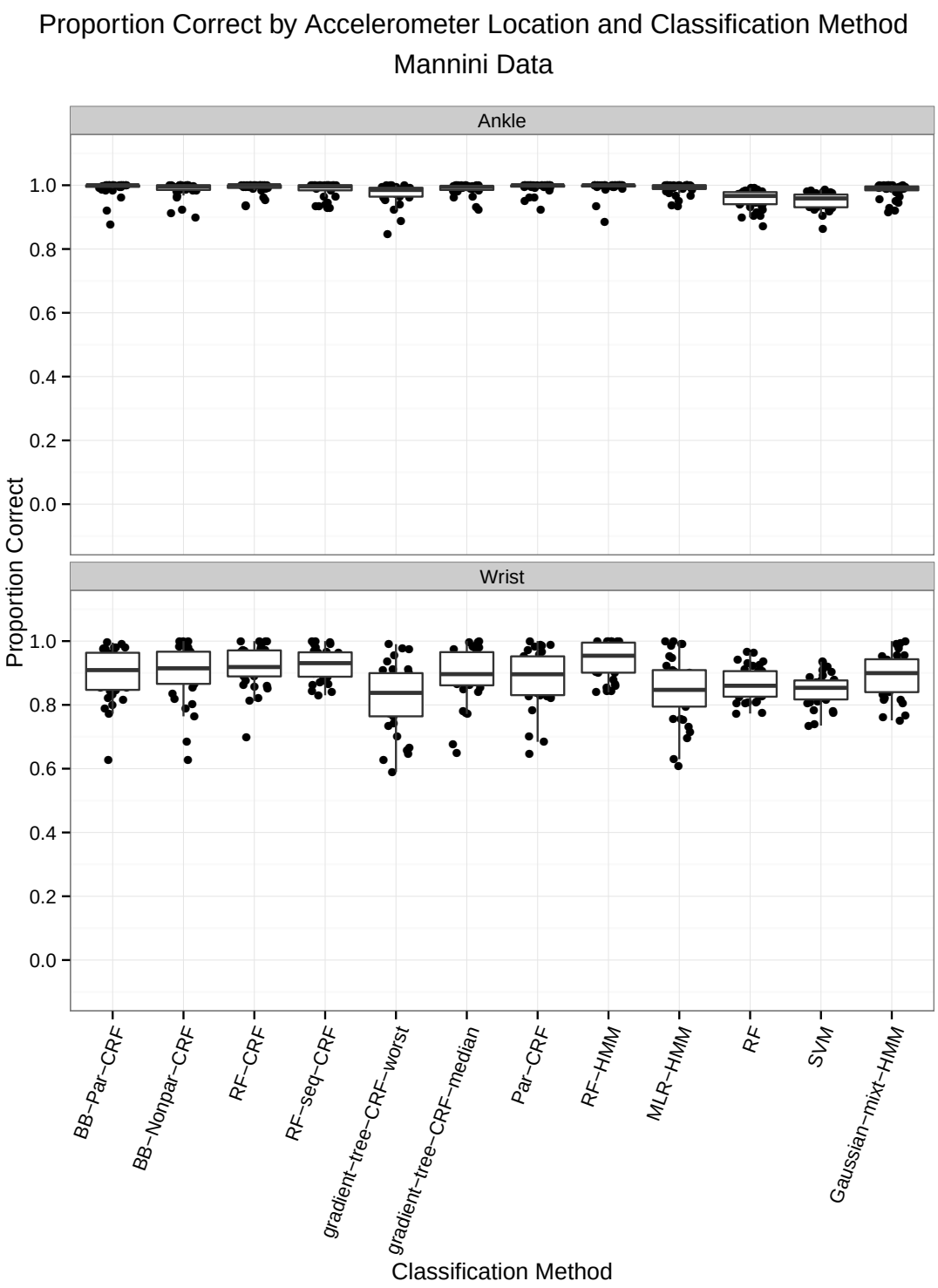


Figure 34. Box plots showing the proportion of windows classified correctly in the data from Mannini et al. [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location and classification method. Each point corresponds to a combination of accelerometer location, classification method, and subject.

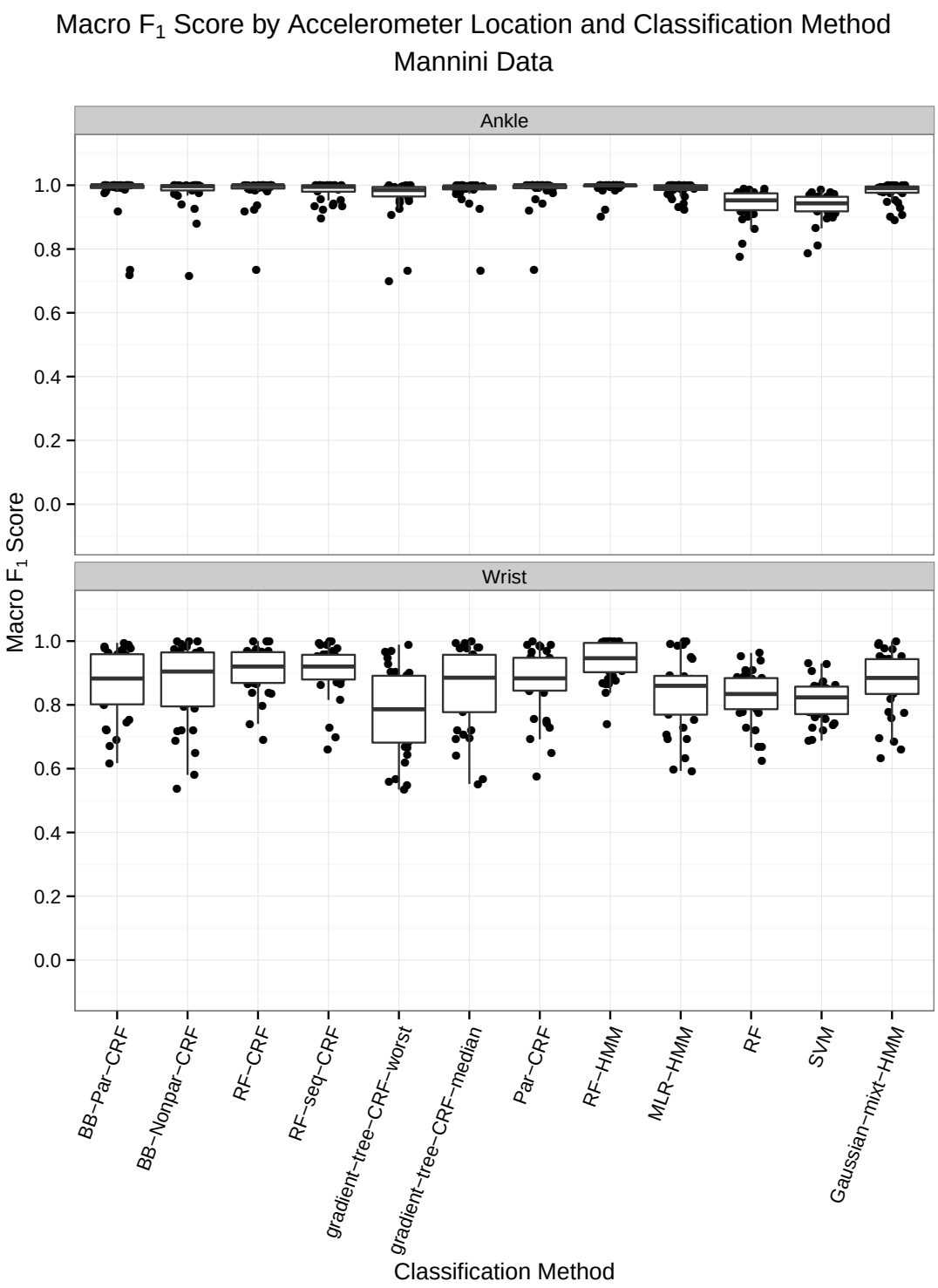


Figure 35. Box plots showing the macro F_1 score combining precision and recall across all four physical activity type categories in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location and classification method. Each point corresponds to a combination of accelerometer location, classification method, and subject.

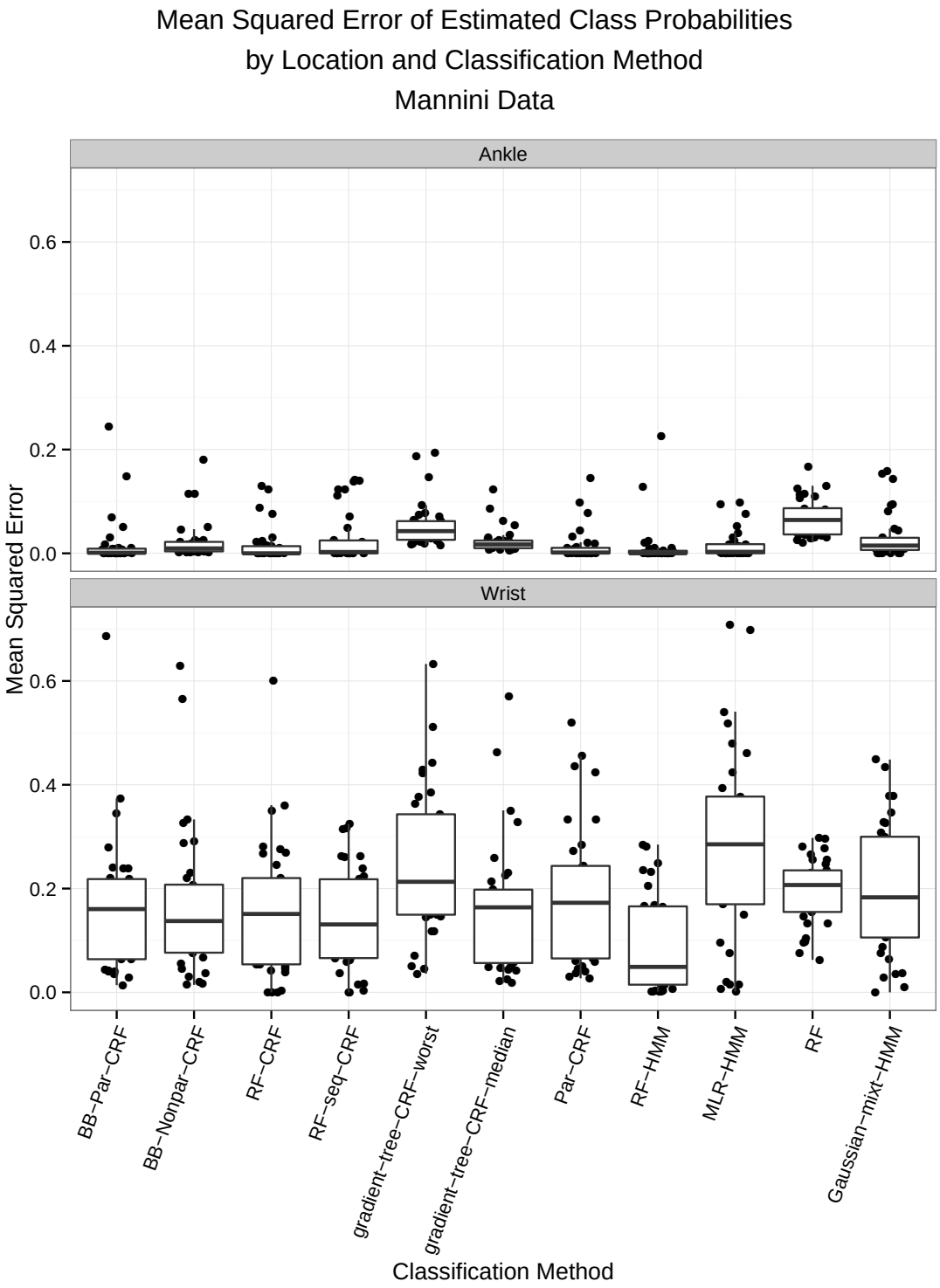


Figure 36. Box plots showing the mean squared error of the estimated class membership probabilities relative to the labeled class memberships in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location and classification method. Each point corresponds to a combination of accelerometer location, classification method, and subject.

	Predicted Class			
Labeled Class	Sedentary	Ambulation	Cycling	Other
Sedentary	2766	38	180	0
Ambulation	68	2173	154	38
Cycling	153	79	800	1
Other	3	43	16	891

Table 8. Confusion matrix for the BB-Par-CRF classification method applied to the wrist data from Mannini et al. [2013], all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Ambulation	Cycling	Other
Sedentary	2779	24	146	35
Ambulation	28	2202	145	58
Cycling	26	27	979	1
Other	9	69	22	853

Table 9. Confusion matrix for the RF-CRF classification method applied to the wrist data from Mannini et al. [2013], all subjects combined.

Figure 37 displays the proportion correct for just the **BB-Par-CRF**, **RF-CRF**, **RF**, and **Gaussian-mixt-HMM** methods with the wrist data. We selected these four methods for the display to give a representative view of how model performance varies with the subject without overcrowding the plot. We can see that there is considerable variability in classifier success between subjects. In general, the methods all do relatively well on the same subjects and relatively poorly on the same subjects. Overall, the static **RF** model tends to do worse than the other methods, but no method achieves a higher proportion of time points classified correctly than the other methods for every subject.

Tables 8, 9, 10, and 11 show confusion matrices aggregated across all subjects for the **BB-Par-CRF**, **RF-CRF**, **RF**, and **Gaussian-mixt-HMM** methods with the wrist data. We show only a few confusion matrices for the sake of brevity. We can see that for all of the classification methods, time spent in the Sedentary and Ambulation categories is most likely to be misclassified as Cycling, time in the Cycling category is most likely to be misclassified as Sedentary or Ambulation,

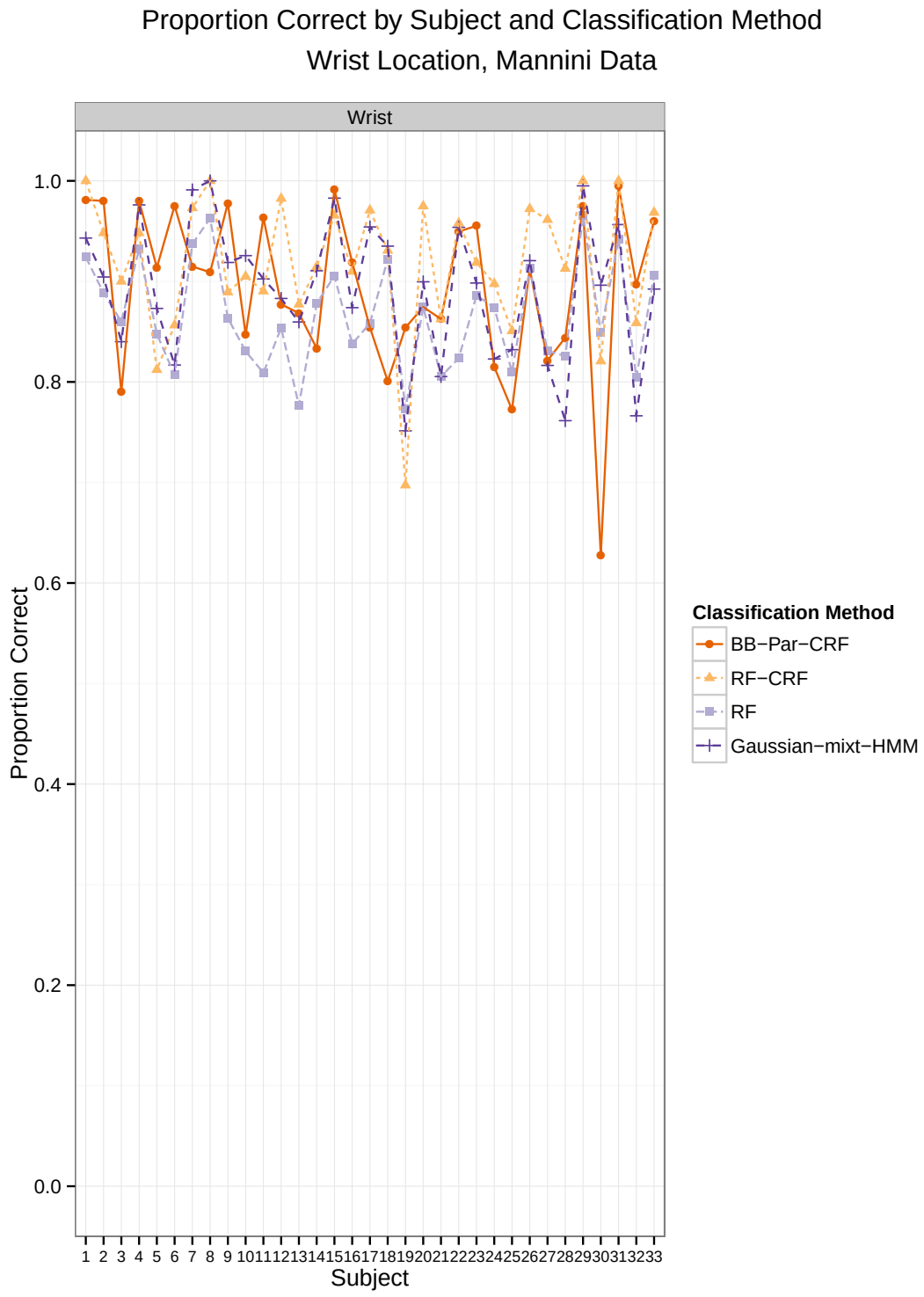


Figure 37. Proportion of time windows classified correctly by subject in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. We show these values for just four classification methods applied to data collected from accelerometers placed at the wrist location.

	Predicted Class			
Labeled Class	Sedentary	Ambulation	Cycling	Other
Sedentary	2755	25	133	71
Ambulation	63	2193	109	68
Cycling	259	64	683	27
Other	72	83	13	785

Table 10. Confusion matrix for the RF classification method applied to the wrist data from Mannini et al. [2013], all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Ambulation	Cycling	Other
Sedentary	2852	26	66	40
Ambulation	69	2088	165	111
Cycling	178	54	790	11
Other	20	30	4	899

Table 11. Confusion matrix for the Gaussian-mixt-HMM classification method applied to the wrist data from Mannini et al. [2013], all subjects combined.

and time in the Other category is most likely to be misclassified as Ambulation.

We fit the following linear mixed effects model to the classification results:

$$p_{c,l,i} = \beta_{c,l} + \gamma_i + \varepsilon_{c,l,i}, \text{ where} \tag{7.3.1}$$

$$\gamma_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{subj}) \tag{7.3.2}$$

$$\varepsilon_{c,l,i} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_{l,i})$$

Here, $p_{c,l,i}$ is the proportion of windows classified correctly using classifier c and the data gathered with the accelerometer placed at location l on subject i. The parameters $\beta_{c,l}$ are fixed effects representing the mean proportion correct for classifier method c and accelerometer location l, γ_i are random effects representing subject-specific offsets to the mean proportion correct, and $\varepsilon_{c,l,i}$ are error terms with variance specific to the combination of accelerometer location and subject.

The use of a separate mean for each combination of accelerometer location and classifier method can be justified by the box plots in Figure 34. For example, we can see that an additive structure for the mean proportion classified correctly is insufficient to capture the fact that the **MLR-HMM** method does as well as the other dynamic classification methods with the data collected at the ankle, but worse than those methods with the data collected at the wrist. The use

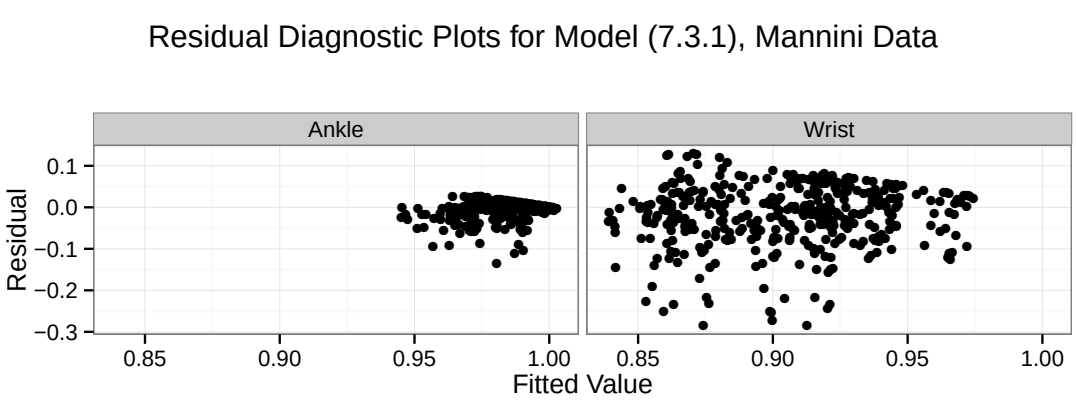
of a subject-specific random effect can be justified by Figure 37, where we saw that the different classification methods tended to do relatively well or relatively poorly on the same subjects. The heteroskedastic structure of the error term is clear from the diagnostic plots in Figure 38.

Although we focus our discussion on this model, we also considered several other variations on the model such as using transformations of the proportion correct as the outcome variable, an additive structure without interactions for the fixed effects, random effects for the combination of subject and location, and a heteroskedasticity structure that depended on only the accelerometer location. We selected this model for discussion because it strikes a balance between parsimony, faithfulness to the data, and interpretability. However, the qualitative conclusions are similar using all of these variations on the model, and they match the conclusions we drew from the plots above. We also considered models for the macro F_1 score, which give very similar qualitative conclusions; we focus on a model for the proportion correct here since it is more interpretable than the macro F_1 score.

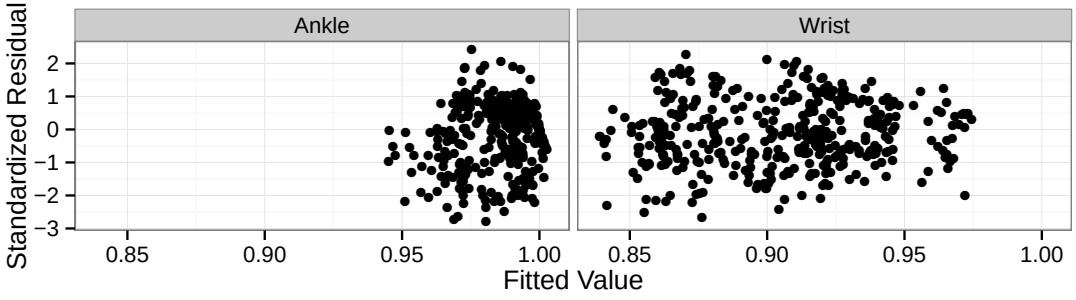
Figures 39, 40, 41, and 42 display point and interval estimates for the $\beta_{c,l}$ parameters and sets of contrasts between them. We calculated the confidence intervals using the `multcomp` package [Hothorn et al., 2008] in `R`. We caution against a strict interpretation of the confidence intervals due to the minor departures from normality of the residuals indicated in Figure 38 and the fact that these are approximate intervals based on asymptotic normality of the parameter estimates. However, the intervals serve to give a general indication of the level of uncertainty in our estimates of effect size.

The estimates displayed in Figure 40 give us another view of our earlier observation that all of the classification methods achieved a higher classification rate when using data from accelerometers placed on the ankle than they did when using data from accelerometers placed on the wrist. The size of the change in the proportion of windows classified correctly depends on the classifier being used, but the point estimates for these differences range from about 0.03 to 0.13. This is equivalent to an improvement of between about 30 minutes and 2 hours if we extrapolate to a 16 hour day while maintaining the same composition of the relative amount of time spent in each physical activity category.

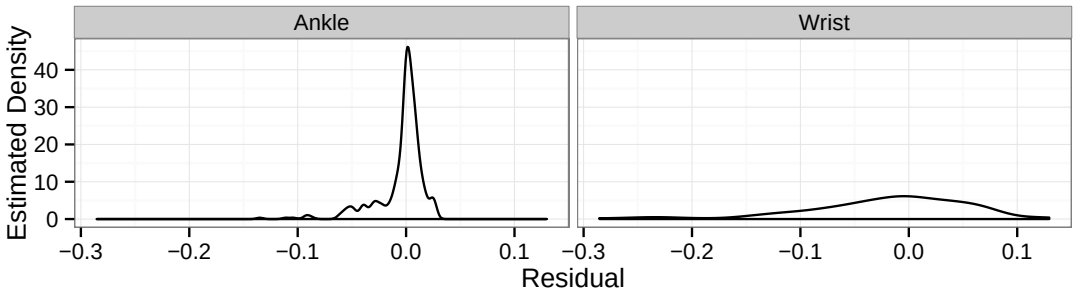
The estimates in Figure 41 show that when applied to the ankle data, all of the dynamic methods performed better than the static **RF** and **SVM** classifiers, but there is not much of a difference between the dynamic methods. The estimated gain from using a dynamic classifier instead of a



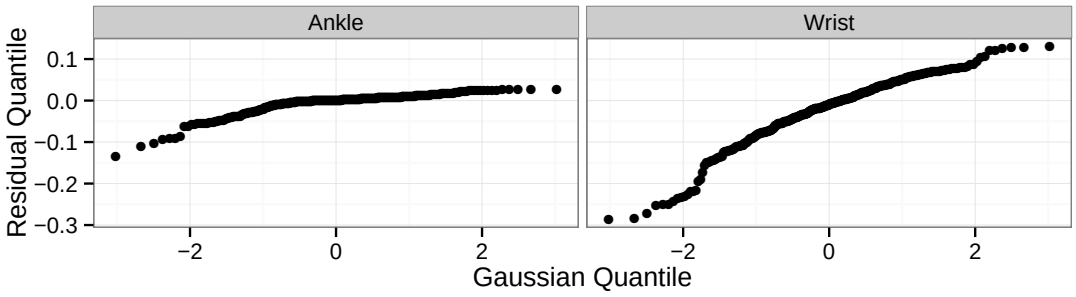
(a) Residuals vs. fitted values.



(b) Standardized residuals vs. fitted values.



(c) Marginal density estimates for residuals.



(d) Quantile-quantile plots for the residuals.

Figure 38. Diagnostic plots for model (7.3.1)**.**

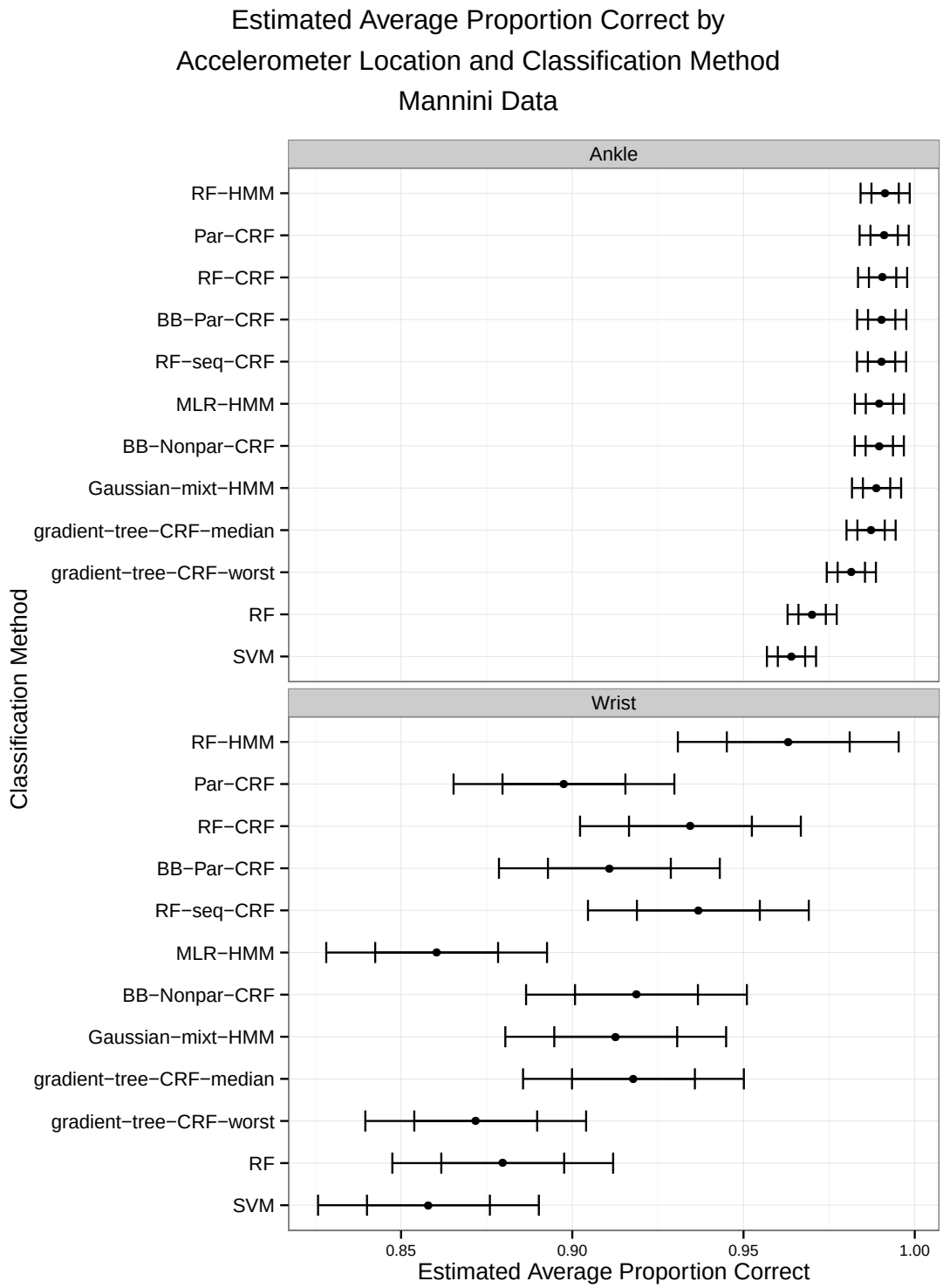


Figure 39. Point and interval estimates for the fixed effects parameters in model (7.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 39, 40, 41, and 42.**

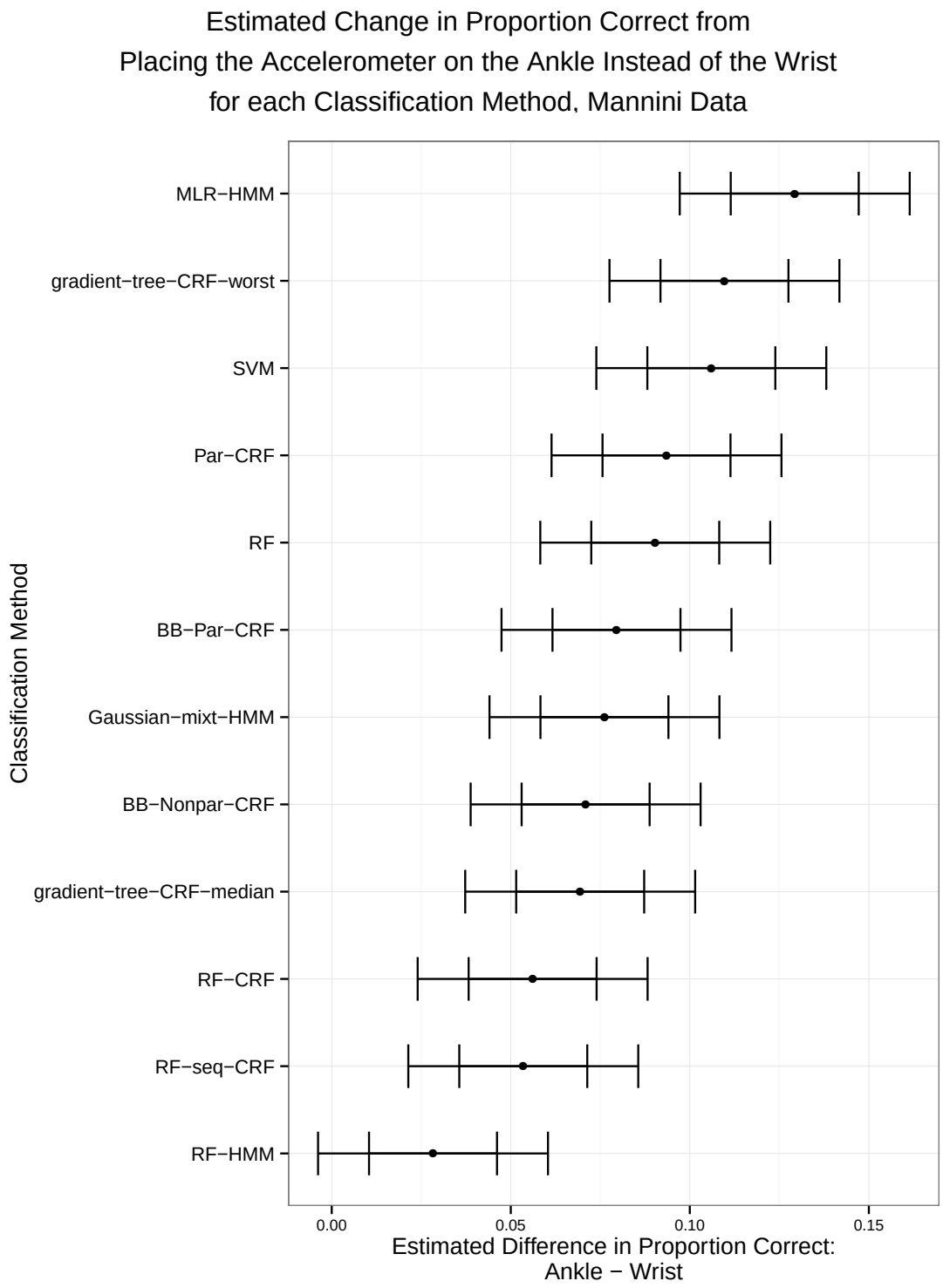


Figure 40. Point and interval estimates for the difference in performance between the ankle and wrist accelerometer locations for each classification method, based on model (7.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 39, 40, 41, and 42.**

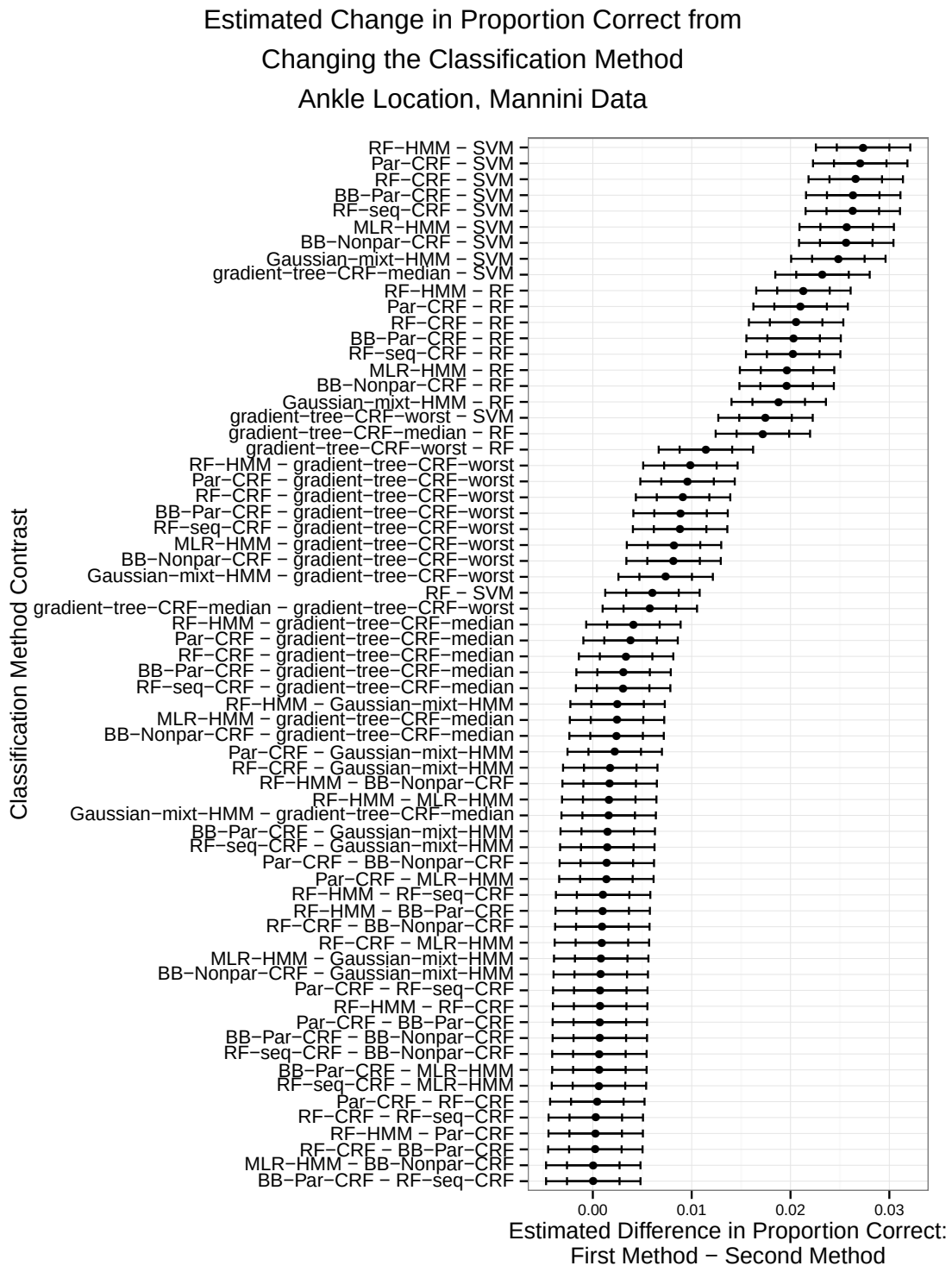


Figure 41. Point and interval estimates for the difference in performance between each pair of classification methods using data from an accelerometer placed at the ankle, based on model (7.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for family-wise 95% confidence intervals for all intervals in Figures 39, 40, 41, and 42.**

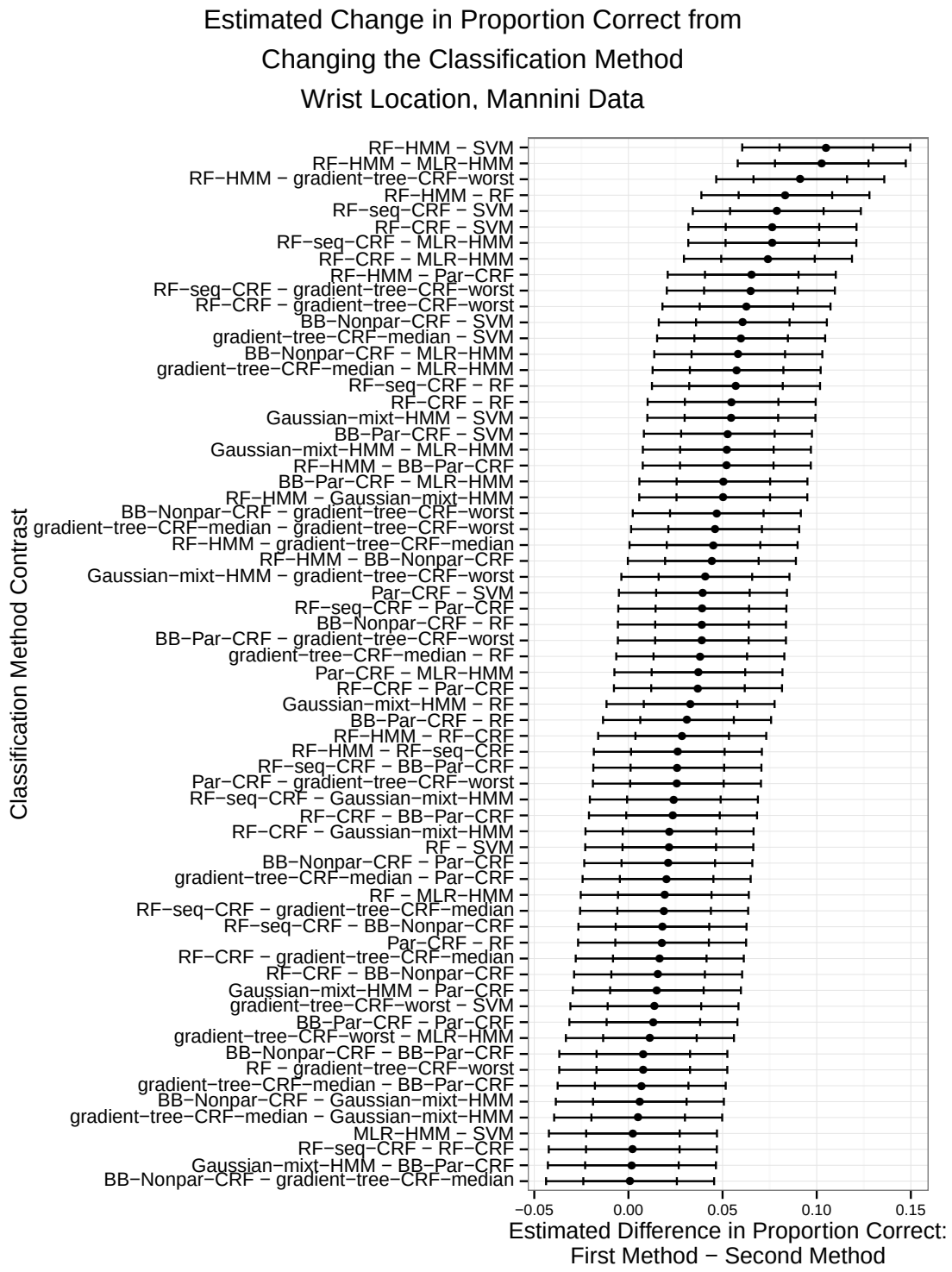


Figure 42. Point and interval estimates for the difference in performance between each pair of classification methods using data from an accelerometer placed at the wrist, based on model (7.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for family-wise 95% confidence intervals for all intervals in Figures 39, 40, 41, and 42.**

static classifier is an increase of about 0.02 to 0.03 in the proportion correct, or roughly 20 to 30 minutes in a 16 hour day. When the **gradient-tree-CRF** method is used with data from the ankle, the performance of the method is fairly stable with respect to how the data are partitioned into training and validation sequences. The estimated average proportion correct achieved in the worst case is only smaller than the median proportion correct by about 0.006, or about 6 minutes in a 16 hour day.

The estimates in Figure 42 indicate that on the wrist, where classification is more difficult, changing the classification method used can lead to larger gains in classification performance. Here again the dynamic classification methods tend to outperform the static methods, with the exception of the dynamic **MLR-HMM** method. The estimated gain in the proportion of windows classified correctly from using a dynamic method based on random forests such as the **RF-HMM** or **RF-CRF** methods instead of a static classifier is in the range of about 0.04 to 0.11, which is on the order of the increases that can be achieved by placing the accelerometer on the ankle instead of the wrist. Using the wrist data, the performance of the **gradient-tree-CRF** method is more sensitive to how the data are partitioned into training and validation sets. Here, the estimated difference between the median proportion correct and the worst case proportion correct is about 0.05, or about 50 minutes in a 16 hour day.

One possible criticism of the results we have presented in this Section is that the dynamic models may allow for more complex temporal dependencies than are warranted by the data. With data collected in the laboratory, the duration of each activity and the transitions between different activity types were determined by researchers, not the subjects. Although the sequences of activities performed were not exactly the same for all subjects, they were similar. Thus, methods that model the transition probabilities between different activity types may have artificially high classification rates since they share information across subjects about the order of activities.

In order to address this criticism, we considered a reduced parameterization of the dynamic models with only a single parameter governing transitions in activity types. For the **Gaussian-mixt-HMM** model, we used a transition matrix Q with the following form:

$$Q = [q_{r,s}], \text{ where}$$

$$q_{r,s} = \begin{cases} q \text{ if } r = s \\ (1-q)/(S-1) \text{ if } r \neq s \end{cases}$$

In words, the probability that the subject's activity type is the same in window $t+1$ as it is in window t is q, and this probability is the same for all activity types. Also, the probability of transitioning to a different activity type is the same for all pairs of activity types, and is equal to $(1-q)/(S-1)$. We also used the corresponding one-parameter parameterization of Ψ_2 in the CRF specifications of Equations (5.2.1), (5.3.1), (5.4.1), and (6.2.1).

We display the results obtained using these reduced model parameterizations in Figures 43 through 48. We can see that with the reduced parameterizations, the **RF** and **SVM** methods are no longer the worst-performing methods; instead, they now outperform a few of the dynamic models. However, the **RF-CRF**, **RF-seq-CRF**, and **RF-HMM** methods, which can be interpreted as methods that add sequential dependence structures to the **RF**, still outperform the **RF** and **SVM** by quite a bit. Thus, it appears that adding even a minimal model for sequential dependence to the static **RF** model improves its classification performance. In practice, we would expect the performance of the dynamic models to be closer to the performance under the full parameterization than to the performance with the reduced parameterization we have considered here. This is because some activity types do occur with higher frequency than others, but this is not captured by the reduced parameterization.

7.4 Sasaki Laboratory Data

Again, we begin with a discussion of plots summarizing the results and present a linear mixed effects model for these results later in the Section. Figures 49, 50, and 51 display box plots summarizing the results of the application to the laboratory data from Sasaki [2013]. We can draw several conclusions from these plots. First, we see that classification was more difficult with 6 classes than with 4. We can also see that the **BB-Par-CRF** and **BB-Nonpar-CRF** methods offer the most consistently high performance among the classification methods we have compared. The **ParCRF**, **RFHMM**, and **MLRHMM** methods do well overall, but suffer from lower performance with 6 classes and the hip data. The **RFCRF** and **RFseqCRF** methods also tend to do well, but suffer from large drops in performance with 6 classes and the accelerometer placed on the ankle or hip. The median performance of the **gradient-tree-CRF** method with respect to the partitioning of the data into training and validation sets is quite good, but the worst case performance is much worse than the other methods with 6 classes and the ankle data. The **RF** and

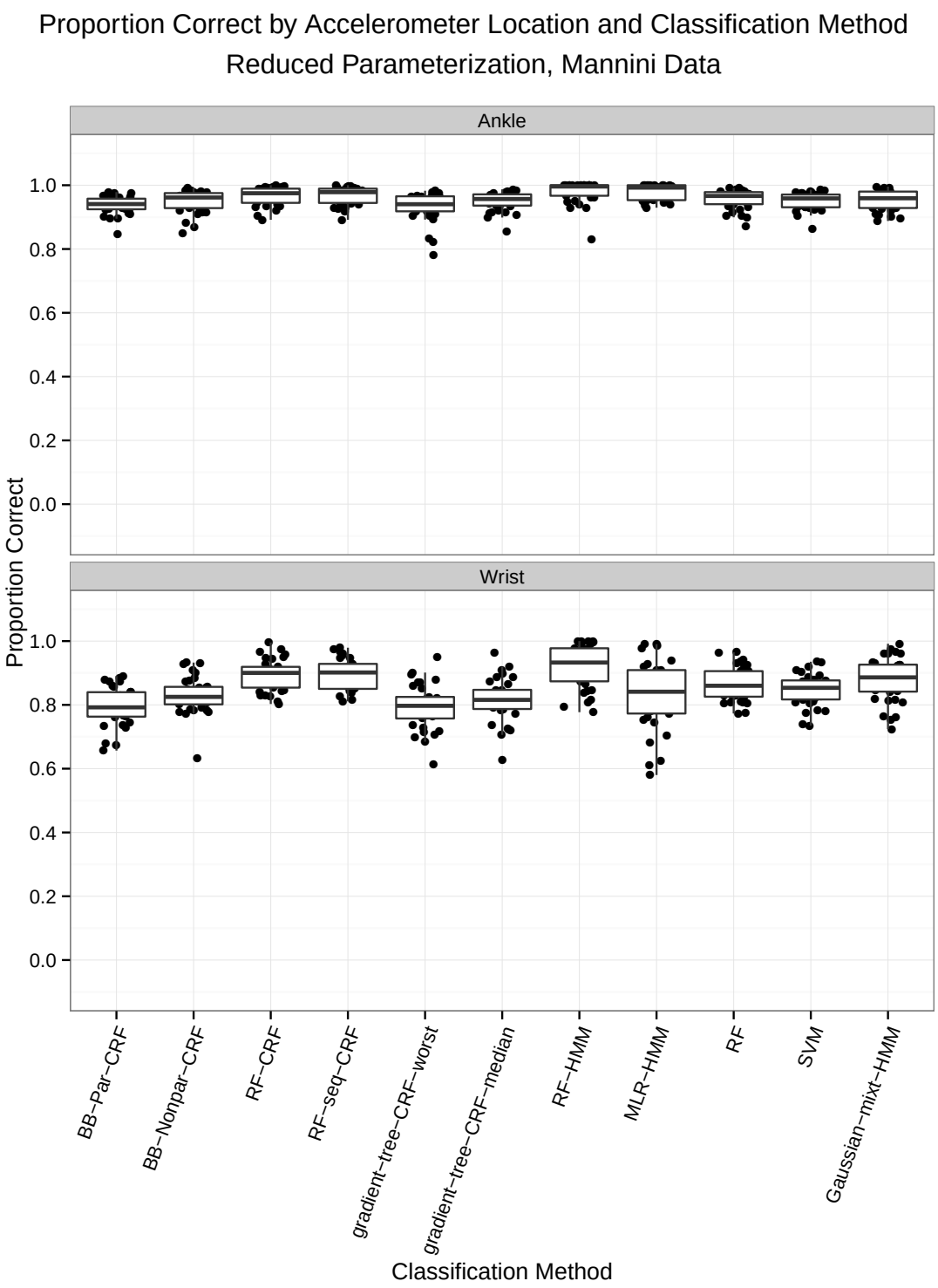


Figure 43. Box plots showing the proportion of windows classified correctly with a reduced parameterization for the dynamic methods in the data from Mannini et al. [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location and classification method. Each point corresponds to a combination of accelerometer location, classification method, and subject.

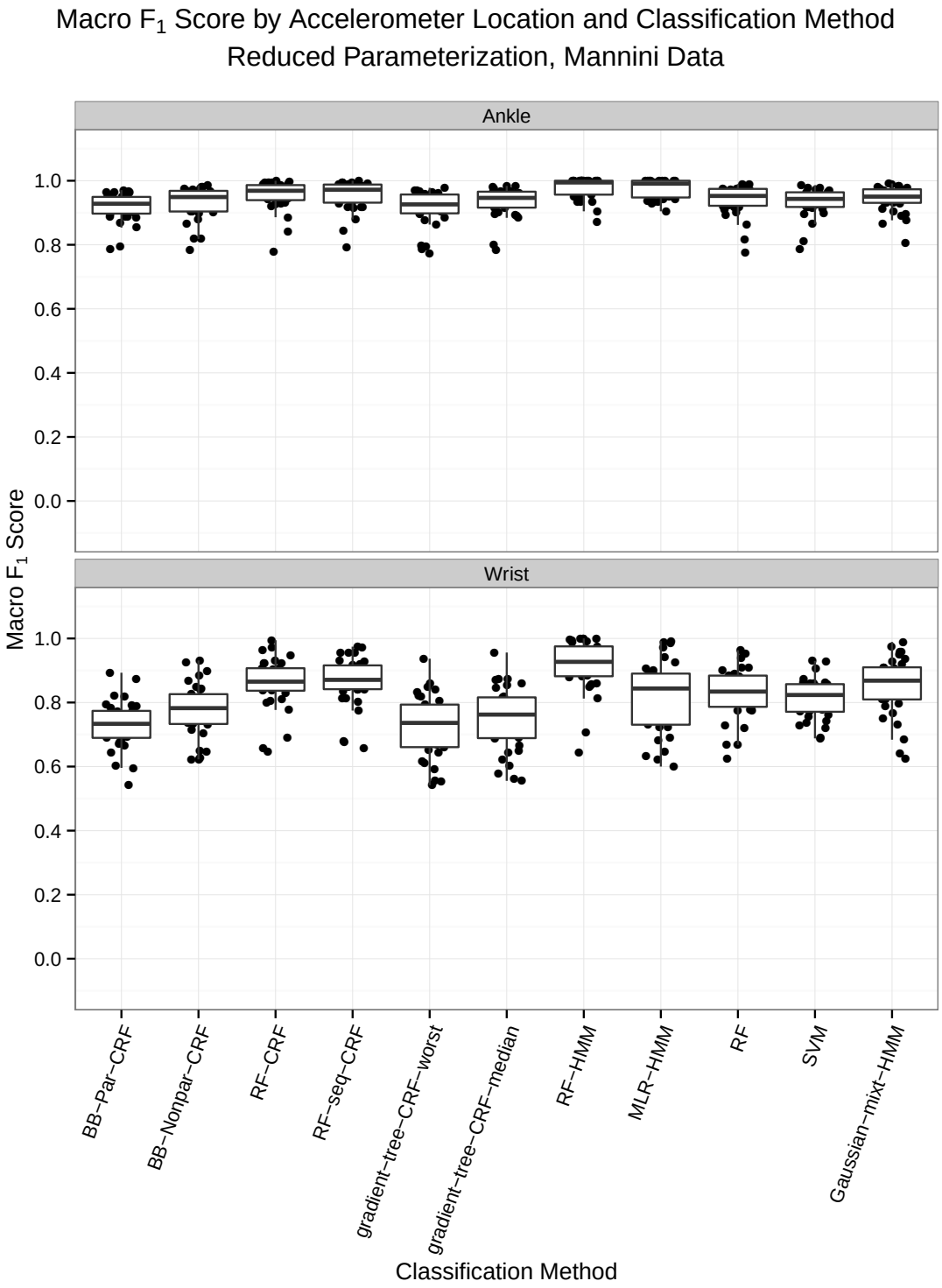


Figure 44. Box plots showing the macro F_1 score combining precision and recall across all four physical activity type categories with a reduced parameterization for the dynamic methods in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location and classification method. Each point corresponds to a combination of accelerometer location, classification method, and subject.

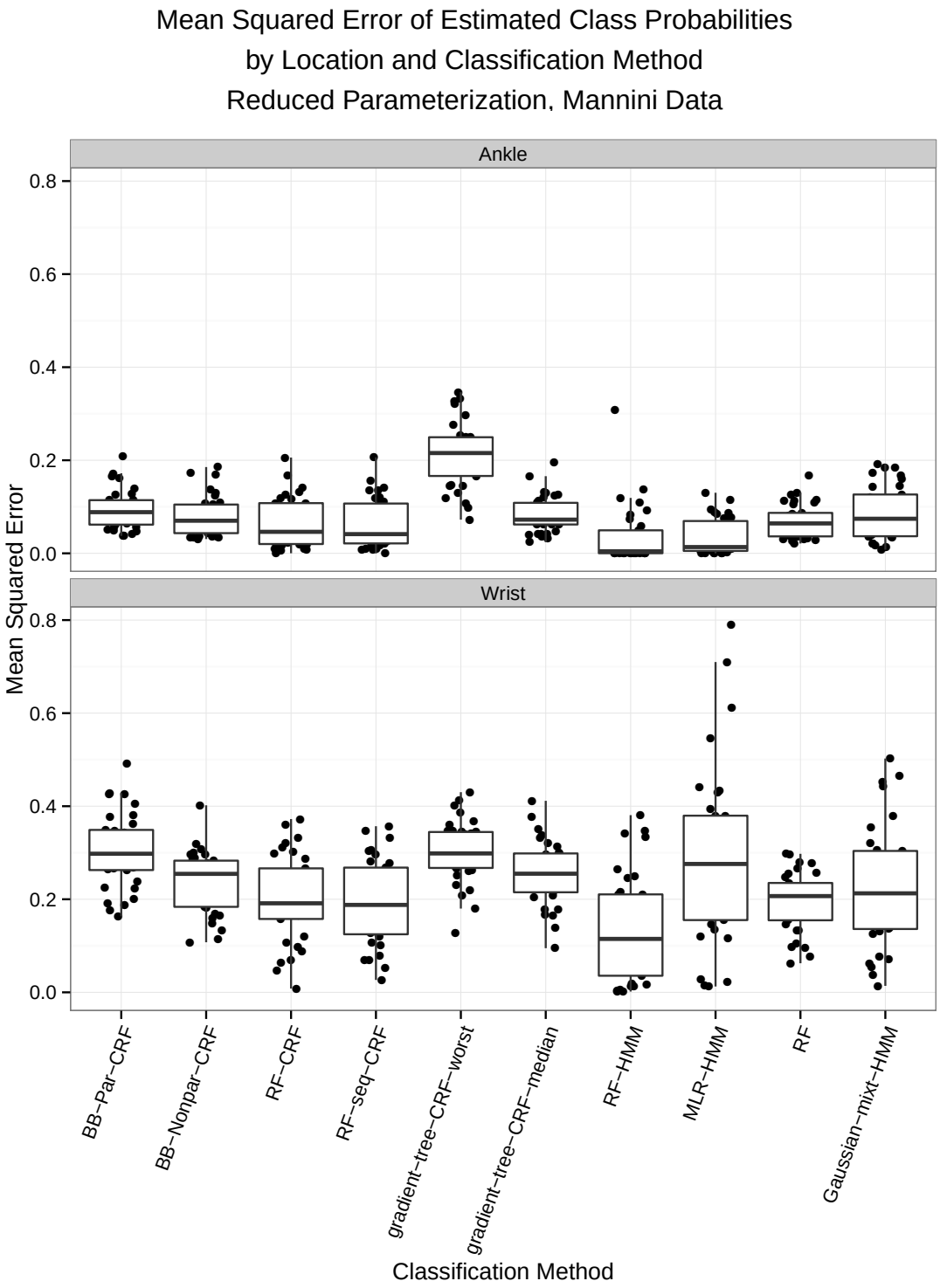


Figure 45. Box plots showing the mean squared error of the estimated class membership probabilities relative to the labeled class memberships with a reduced parameterization for the dynamic methods in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location and classification method. Each point corresponds to a combination of accelerometer location, classification method, and subject.

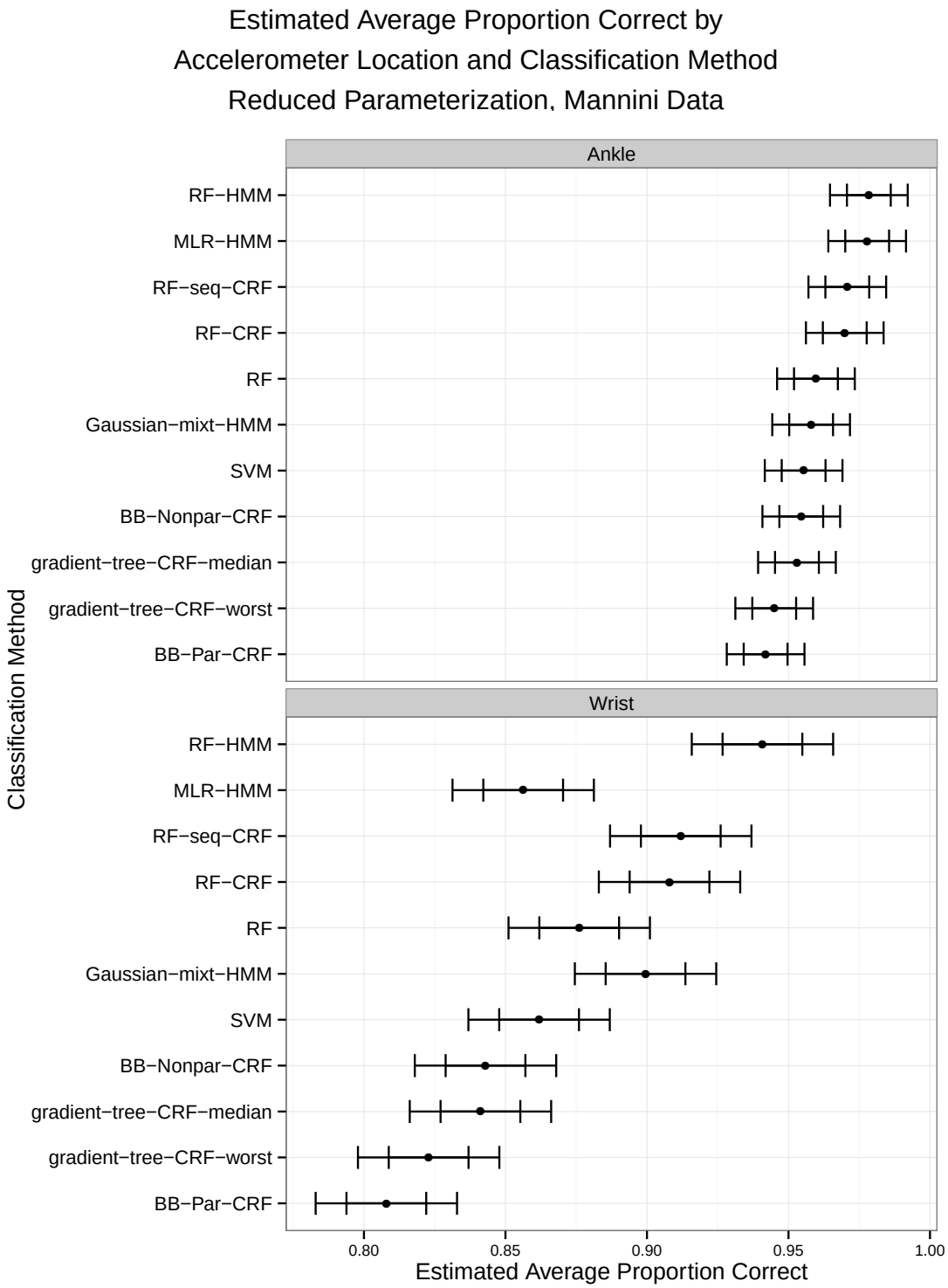


Figure 46. Point and interval estimates for the fixed effects parameters in model (7.3.1) **for the classification results with a reduced parameterization for the dynamic methods. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 46, 47, and 48.**

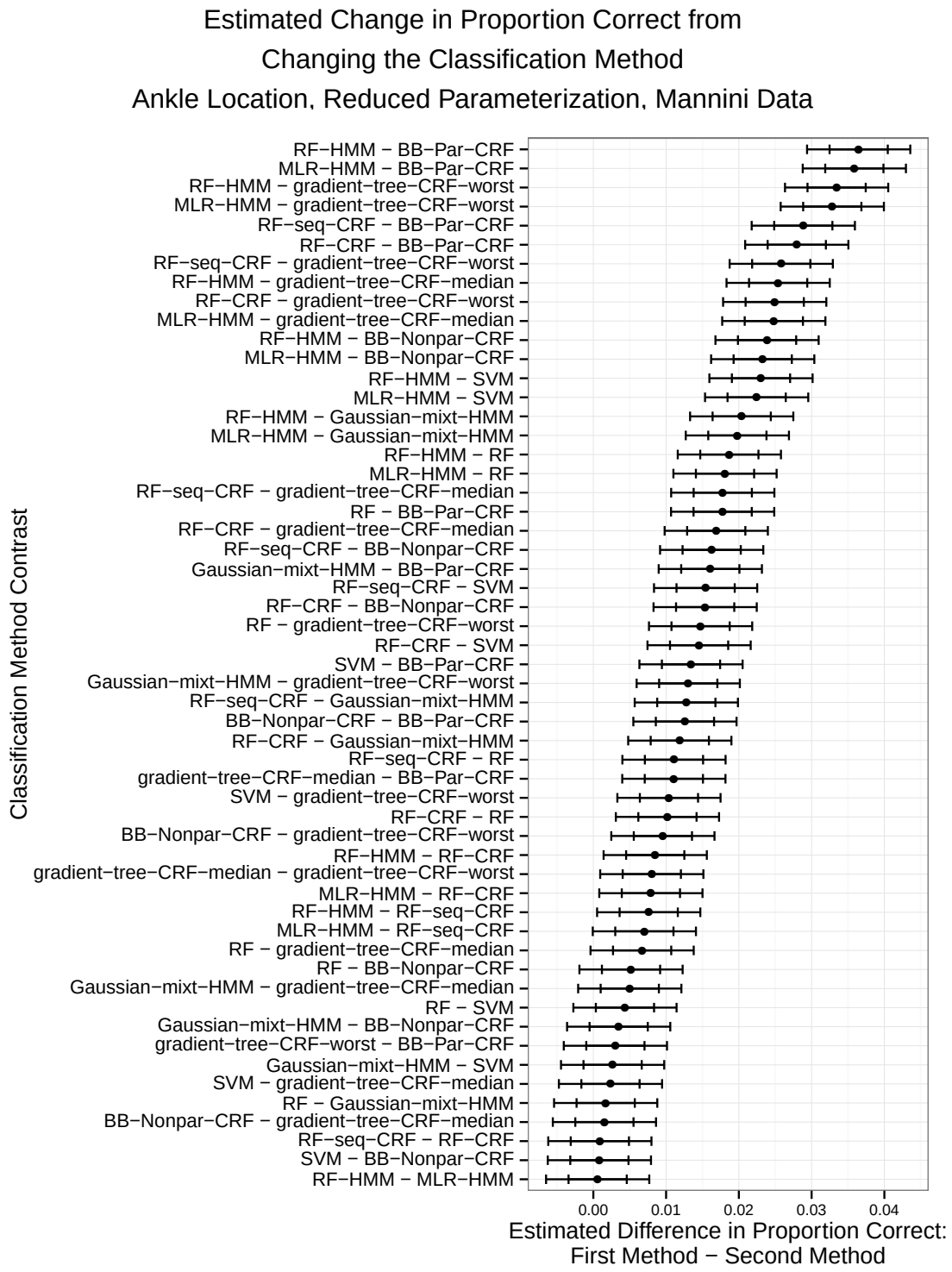


Figure 47. Point and interval estimates for the difference in performance between each pair of classification methods using data from an accelerometer placed at the ankle, based on model (7.3.1) **for the classification results with a reduced parameterization for the dynamic methods. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 46, 47, and 48.**

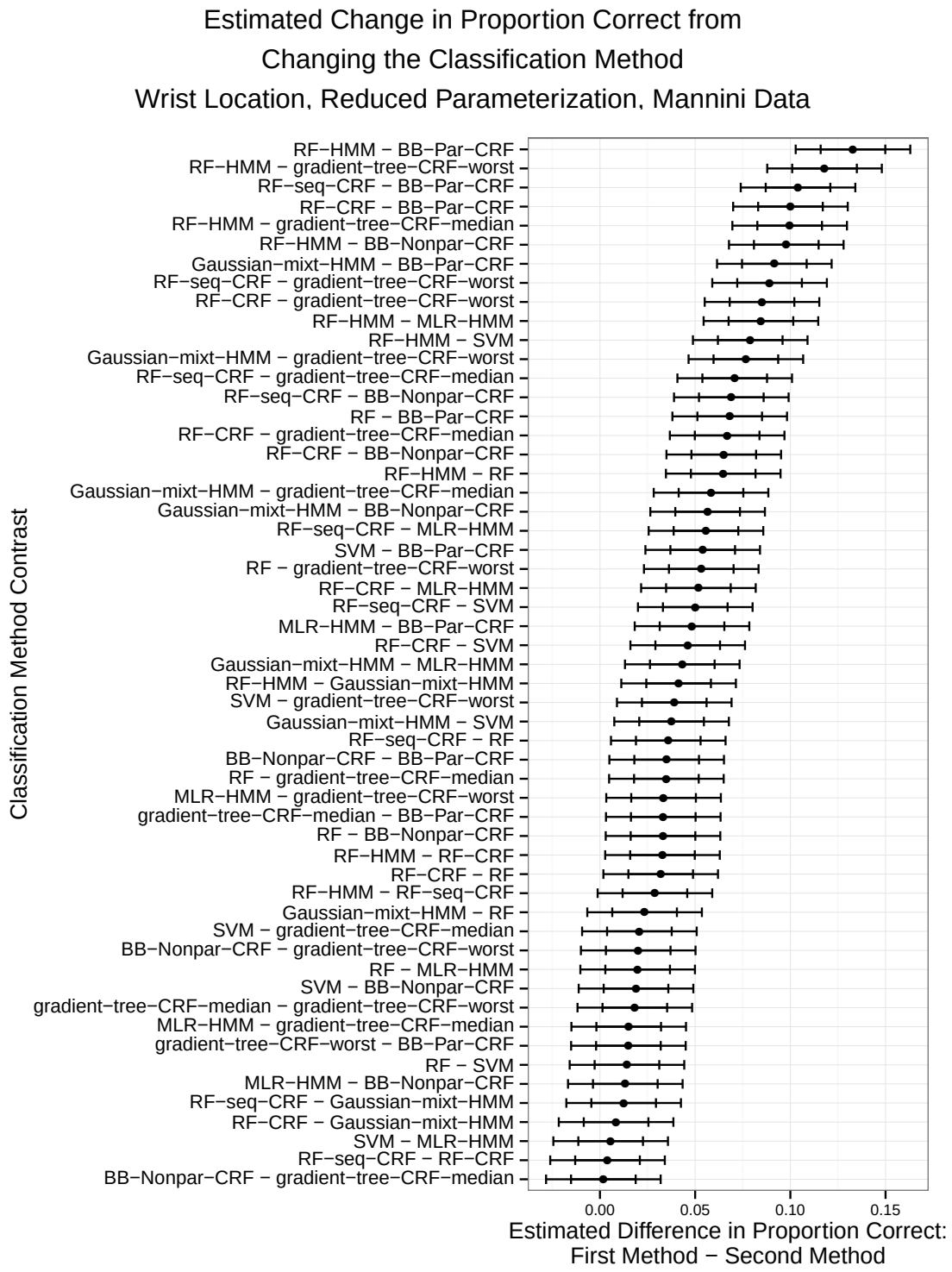


Figure 48. Point and interval estimates for the difference in performance between each pair of classification methods using data from an accelerometer placed at the wrist, based on model (7.3.1) for the classification results with a reduced parameterization for the dynamic methods. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 46, 47, and 48.

Gaussian-mixt-HMM methods consistently underperform relative to the other methods with all accelerometer locations and classification schemes. All of these results are consistent whether we look at the proportion of windows classified correctly, the macro F_1 score, or the MSE of the estimated class probabilities.

Figure 52 displays the proportion correct for just the **BB-Par-CRF**, **RF-CRF**, **RF**, and **Gaussian-mixt-HMM** methods with the wrist data. The story is similar to what we saw with the data from Mannini et al. [2013]: there is considerable variation in classifier performance across subjects, and no classifier achieves the highest proportion of windows classified correctly for all subjects. However, the **BB-Par-CRF** method is consistently among the best of these four methods.

Tables 12, 13, 14, and 15 show confusion matrices for these results aggregated across all subjects. We can see that the **BB-Par-CRF** method performs very well, and the majority of the errors involve the Transition class. The **RF-CRF** method also does fairly well for most classes, but it mislabels a large number of windows where the true class is Moving Intermittently as Recreational. In the laboratory data, the Moving Intermittently class was formed by merging the Dusting, Gardening, Vacuuming, Self-Care, Laundry, and Organizing the Room sub-categories. The windows that were labeled as Moving Intermittently and misclassified as Recreational activity were distributed fairly evenly across these sub-categories. The confusion matrices for the **RF** and **Gaussian-mixt-HMM** methods indicate that those methods had trouble distinguishing among the Moving Intermittently, Ambulation, and Recreational categories.

We fit the following linear mixed effects model to the classification results:

$$p_{c,l,k,i} = \beta_{c,l,k} + \gamma_i + \varepsilon_{c,l,k,i}, \text{ where} \tag{7.4.1}$$

$$\gamma_i \sim N(0, \sigma^2_{subj}) \tag{7.4.2}$$

$$\varepsilon_{c,l,k,i} \sim N(0, \sigma^2_{l,i})$$

Here, $p_{c,l,k,i}$ is the proportion of windows classified correctly using classifier c and the data gathered with the accelerometer placed at location l on subject i when there were k classes. The parameters $\beta_{c,l,k}$ are fixed effects representing the mean proportion correct for accelerometer location l and classifier method c with k classes, the γ_i are random effects representing subject-specific offsets to the mean proportion correct, and $\varepsilon_{c,l,k,i}$ is an error term with variance specific to the combination of accelerometer location and subject. This model is similar to the one we used for the results from Mannini et al. [2013]; the only difference is that we now have the index

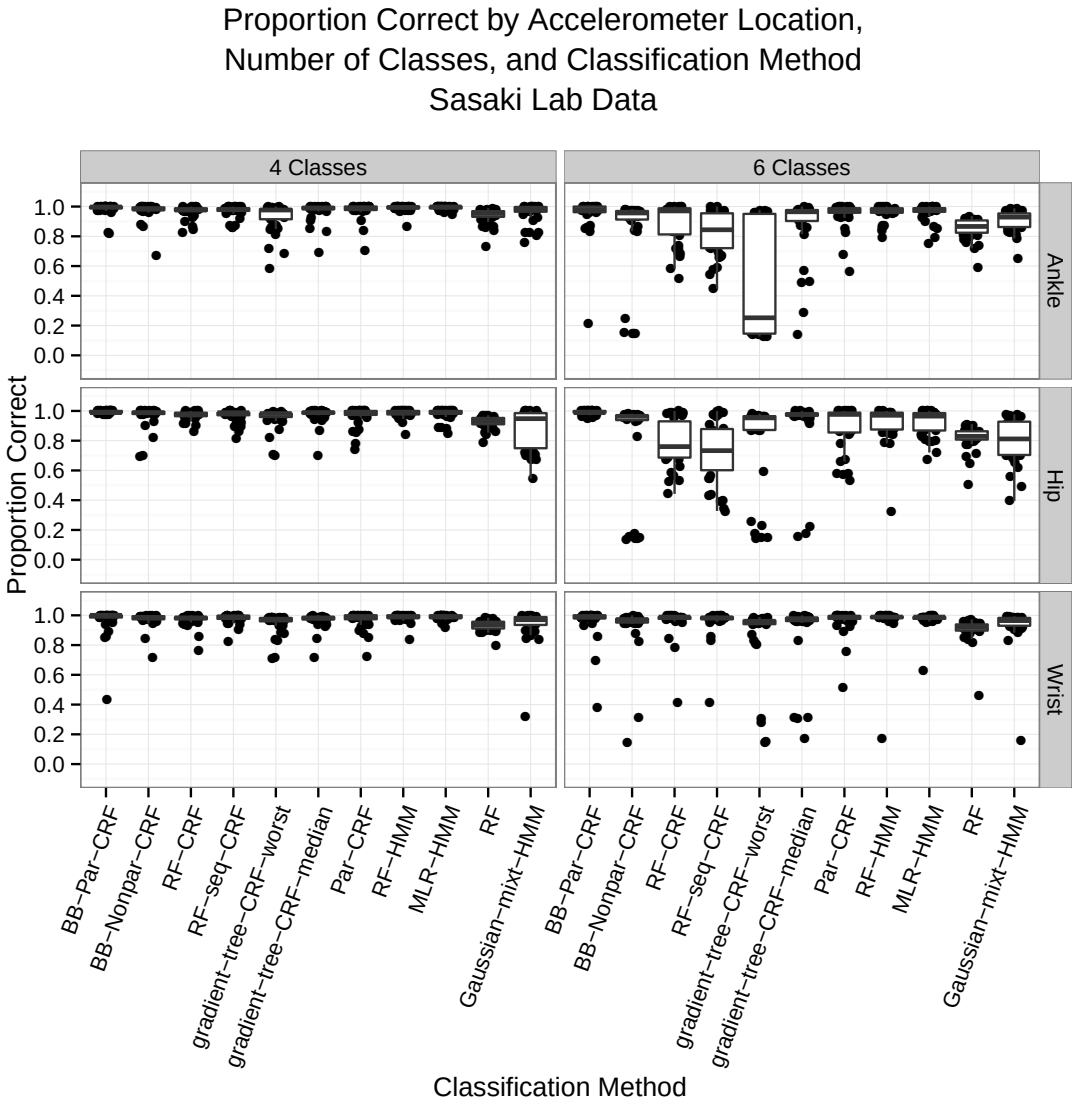


Figure 49. Box plots showing the proportion of windows classified correctly in the laboratory data from Sasaki [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and number of classes. Each point corresponds to a combination of accelerometer location, classification method, number of classes, and subject.

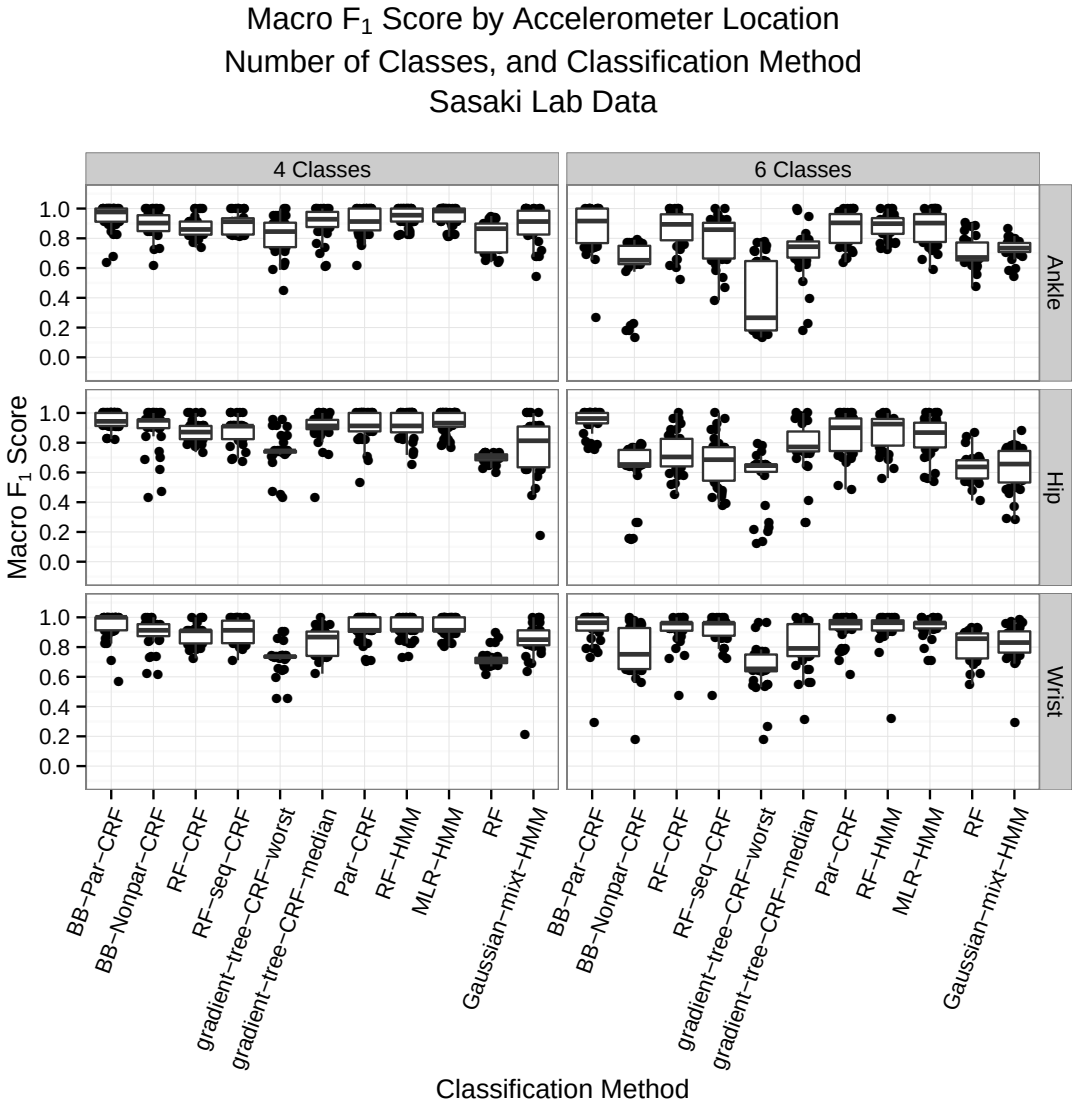


Figure 50. Box plots showing the macro F_1 score combining precision and recall across all physical activity type categories in the laboratory data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and number of classes. Each point corresponds to a combination of accelerometer location, classification method, number of classes, and subject.

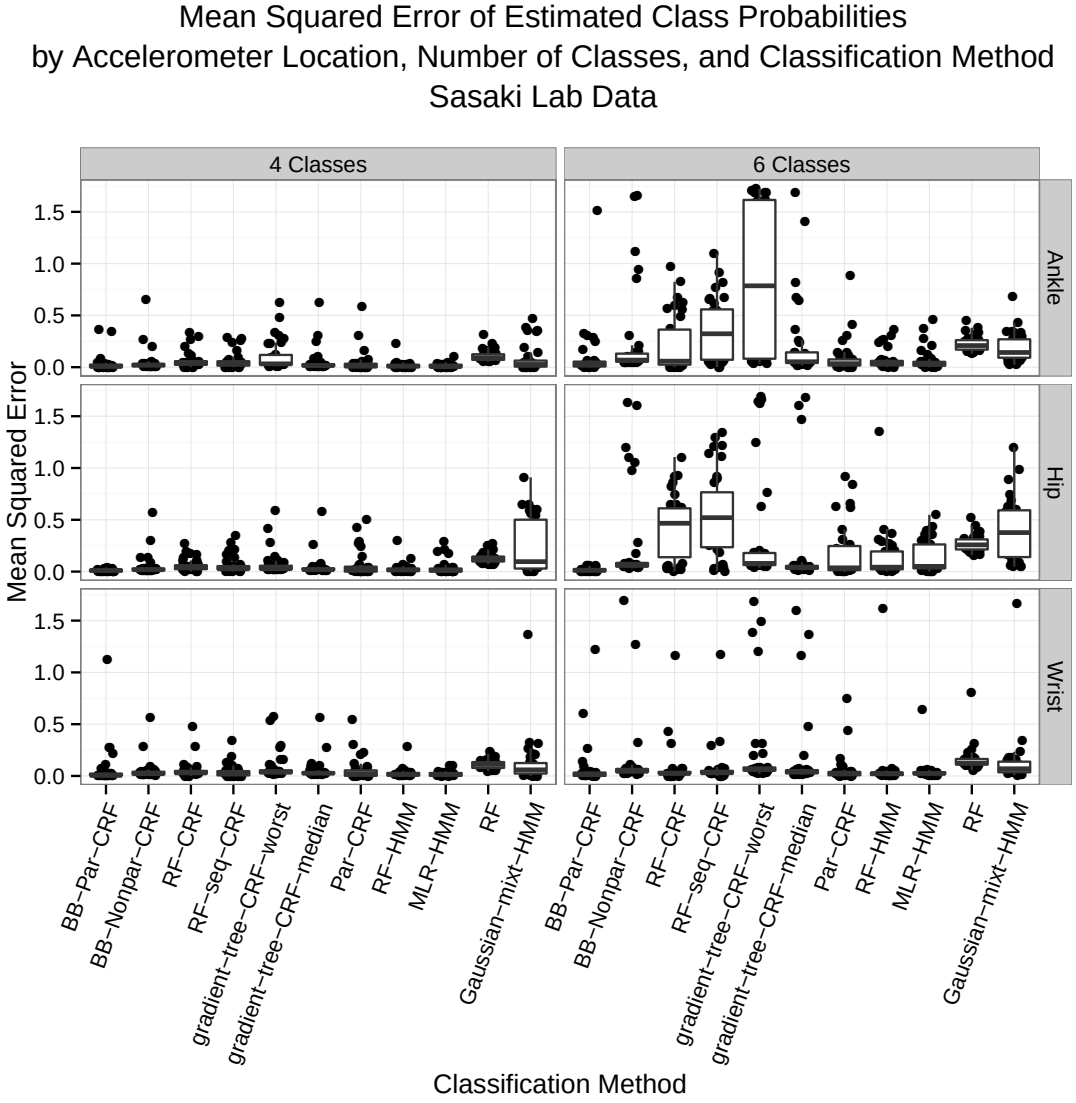


Figure 51. Box plots showing the mean squared error of the estimated classification probabilities relativ to the labeled class memberships in the laboratory data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and number of classes. Each point corresponds to a combination of accelerometer location, classification method, number of classes, and subject.

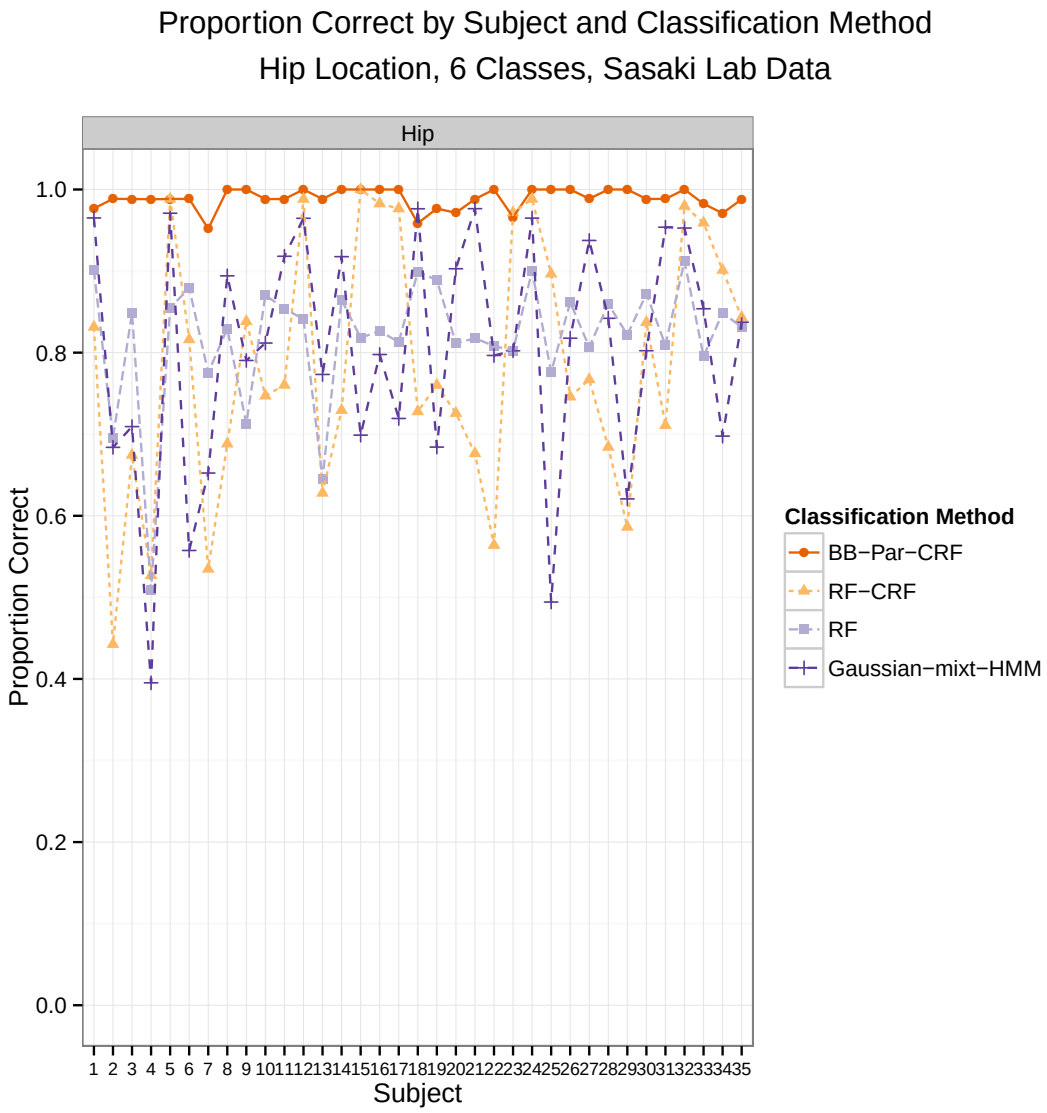


Figure 52. Proportion of time windows classified correctly by subject in the data from Sasaki [2013] with 6 classes, using a leave-one-subject-out procedure. We show these values for just four classification methods applied to data collected from accelerometers placed at the hip.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	932	1	0	0	0	5
Standing	6	28	0	0	0	0
Moving Intermittently	9	0	2420	0	0	4
Ambulation	0	0	0	1624	0	4
Recreational	0	1	0	3	794	9
Transition	8	0	8	11	0	144

Table 12. Confusion matrix for the BB-Par-CRF classification method applied to the lab hip data from Sasaki [2013] with 6 classes, all subjects combined.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	858	1	0	2	66	11
Standing	21	12	0	0	1	0
Moving Intermittently	15	1	1398	0	1008	11
Ambulation	0	0	1	1612	2	13
Recreational	0	0	69	0	729	9
Transition	30	0	13	9	16	103

Table 13. Confusion matrix for the RF-CRF classification method applied to the lab hip data from Sasaki [2013] with 6 classes, all subjects combined.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	857	6	41	0	34	0
Standing	28	4	0	0	2	0
Moving Intermittently	61	1	2140	33	197	1
Ambulation	0	0	81	1540	7	0
Recreational	46	1	409	5	338	8
Transition	4	0	91	9	25	42

Table 14. Confusion matrix for the RF classification method applied to the lab hip data from Sasaki [2013] with 6 classes, all subjects combined.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	855	0	43	0	0	40
Standing	32	0	2	0	0	0
Moving Intermittently	79	0	2179	15	0	160
Ambulation	0	0	198	1128	166	136
Recreational	42	0	191	0	529	45
Transition	4	0	23	3	7	134

Table 15. Confusion matrix for the Gaussian-mixt-HMM classification method applied to the lab hip data from Sasaki [2013] with 6 classes, all subjects combined.

k representing the number of classes used.

The considerations in formulating this model are similar to those we discussed in Section 7.3. We omit the diagnostic residual plots; they show similar issues to what we saw in the earlier model, with the distribution of residuals exhibiting a heavy left tail with several low outliers. As before, we view this model as a descriptive tool only; we do not advocate for a strict interpretation of the confidence intervals obtained from it. The qualitative results were the same with a range of similar models for both the proportion correct and the macro F_1 score.

We display point and interval estimates for the average proportion correct for each combination of classifier, accelerometer location, and number of classes, along with sets of contrasts between these quantities in Figures 53, 54, 55, 56, 57, and 58. These estimates confirm the observations we made from the box plots above.

The estimates in Figure 54 show that with four classes, there is essentially no difference in the classification rates achieved with data from the ankle, hip, or wrist. With six classes, the performance of some of the methods starts to degrade when the accelerometer is on the hip. This is most prominent with the **RF-CRF** and **RF-seq-CRF** methods, but also affects the **RF** and **Gaussian-mixt-HMM** methods.

Figures 55, 56, and 57 show that with four classes, the **RF** and to a lesser extent the **Gaussian-mixt-HMM** model tend to underperform relative to the other classification methods. This is also true for the case when there are six classes, except for the **RF-CRF** and **RF-seq-CRF** methods, which do not do well in that case. Point estimates for the gain in the proportion correct from using a dynamic classifier instead of a static RF vary from about 0.02 to 0.13 depending on the classifier used, the accelerometer location, and the number of classes. This translates to roughly 20 minutes to 2 hours if we extrapolate to a 16 hour day while maintaining the relative amounts of time spent in each activity class. The estimated gain from using a dynamic model that conditions on the accelerometer features instead of the **Gaussian-mixt-HMM** is smaller, but ranges from about 0.01 to 0.1, or about 10 to 100 minutes in a 16 hour day.

7.5 Sasaki Free Living Data

Again, we begin with a discussion of plots summarizing the results before discussing a linear mixed effects model describing the results. Figures 59, 60, and 61 display box plots summarizing

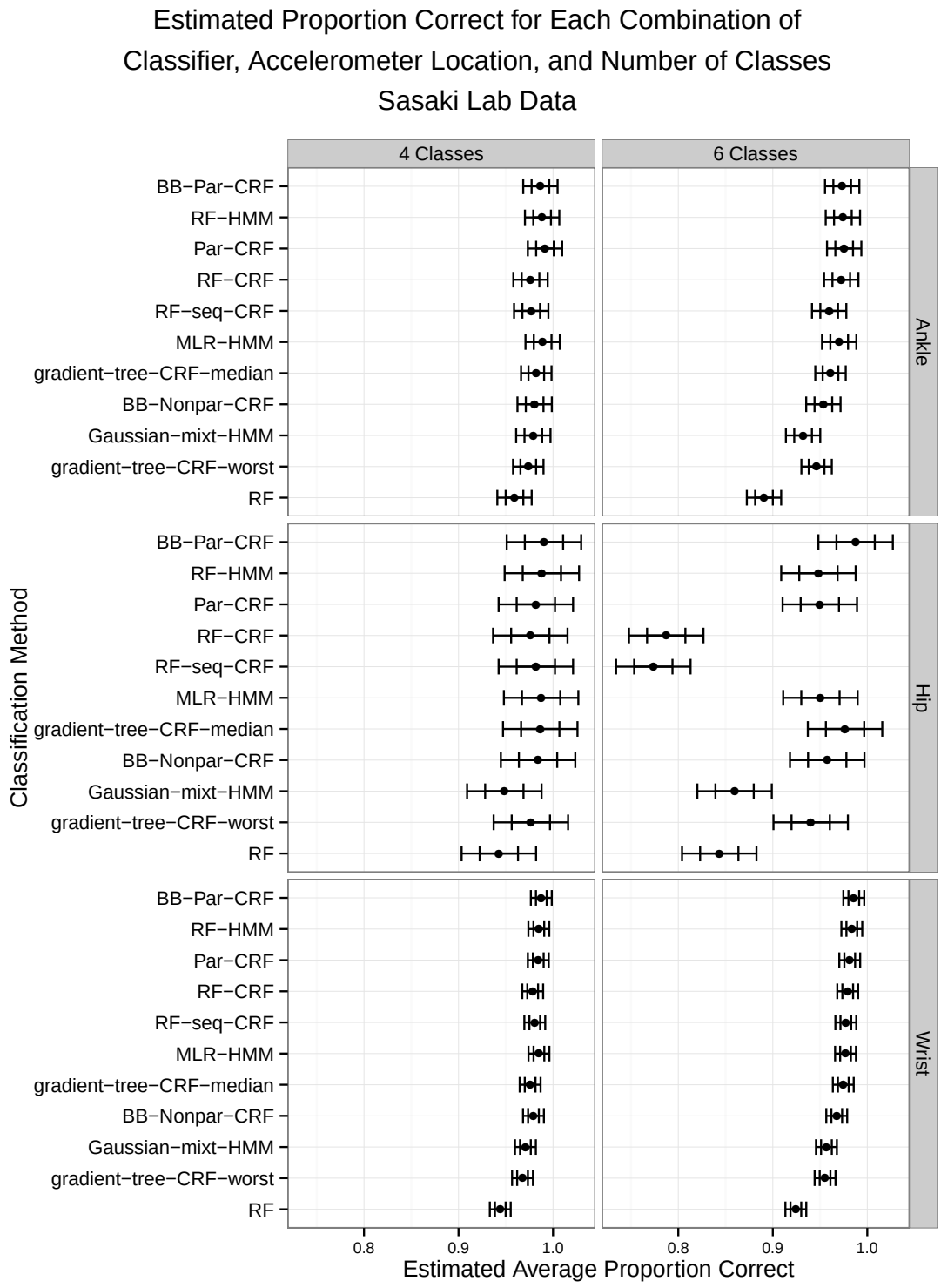


Figure 53. Point and interval estimates for the fixed effects parameters in model (7.4.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 53, 54, 55, 56, 57, and 58.**

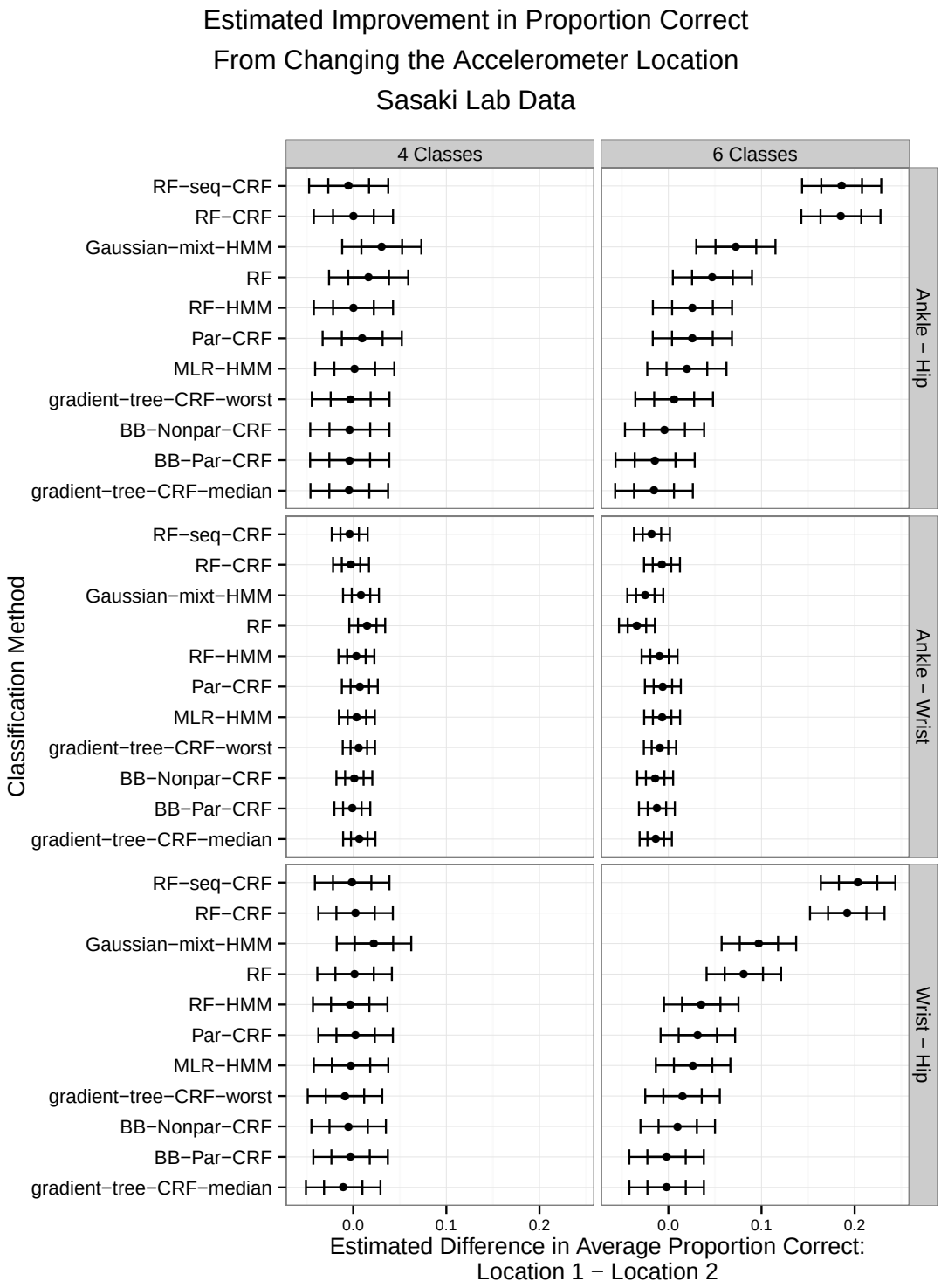


Figure 54. Point and interval estimates for the difference in performance between each pair of accelerometer locations, holding fixed the classification method and the number of classes. The confidence intervals are based on model (7.4.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 53, 54, 55, 56, 57, and 58.**

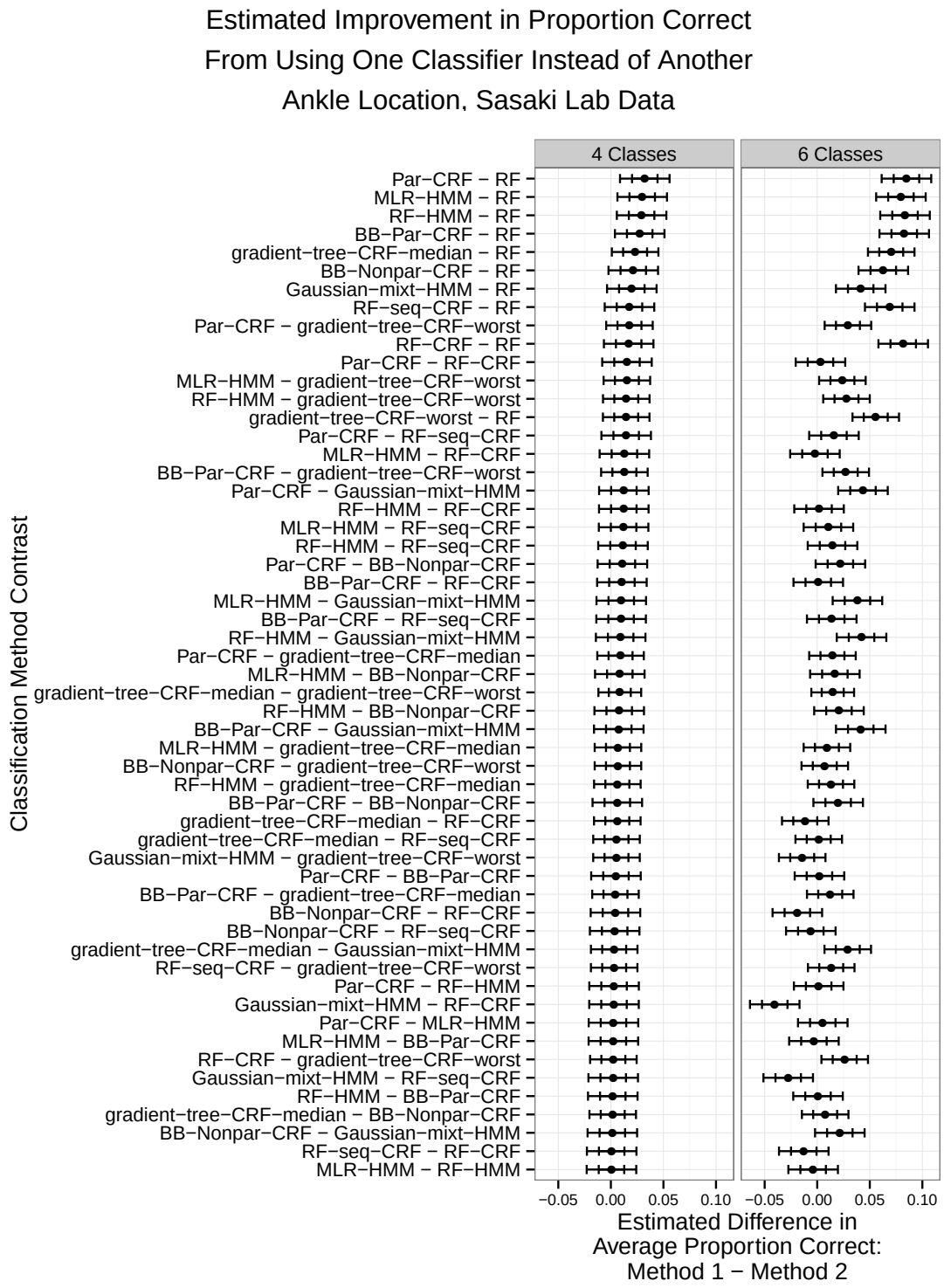


Figure 55. Point and interval estimates for the difference in average proportion of windows classified correctly for each pair of methods using data from the ankle. The estimates are based on model (7.4.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 53, 54, 55, 56, 57, and 58.**

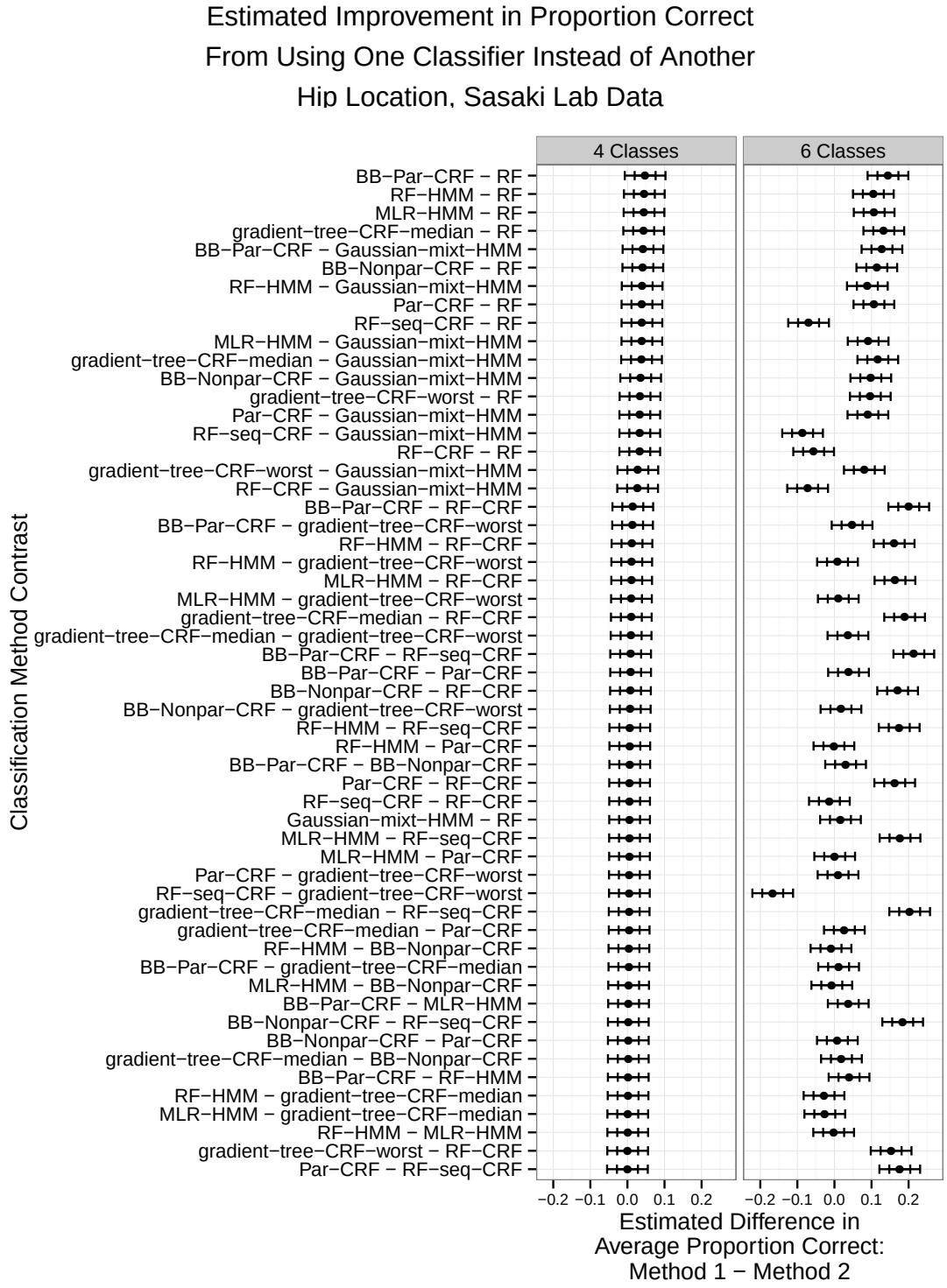


Figure 56. Point and interval estimates for the difference in average proportion of windows classified correctly for each pair of methods using data from the hip. The estimates are based on model (7.4.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 53, 54, 55, 56, 57, and 58.**

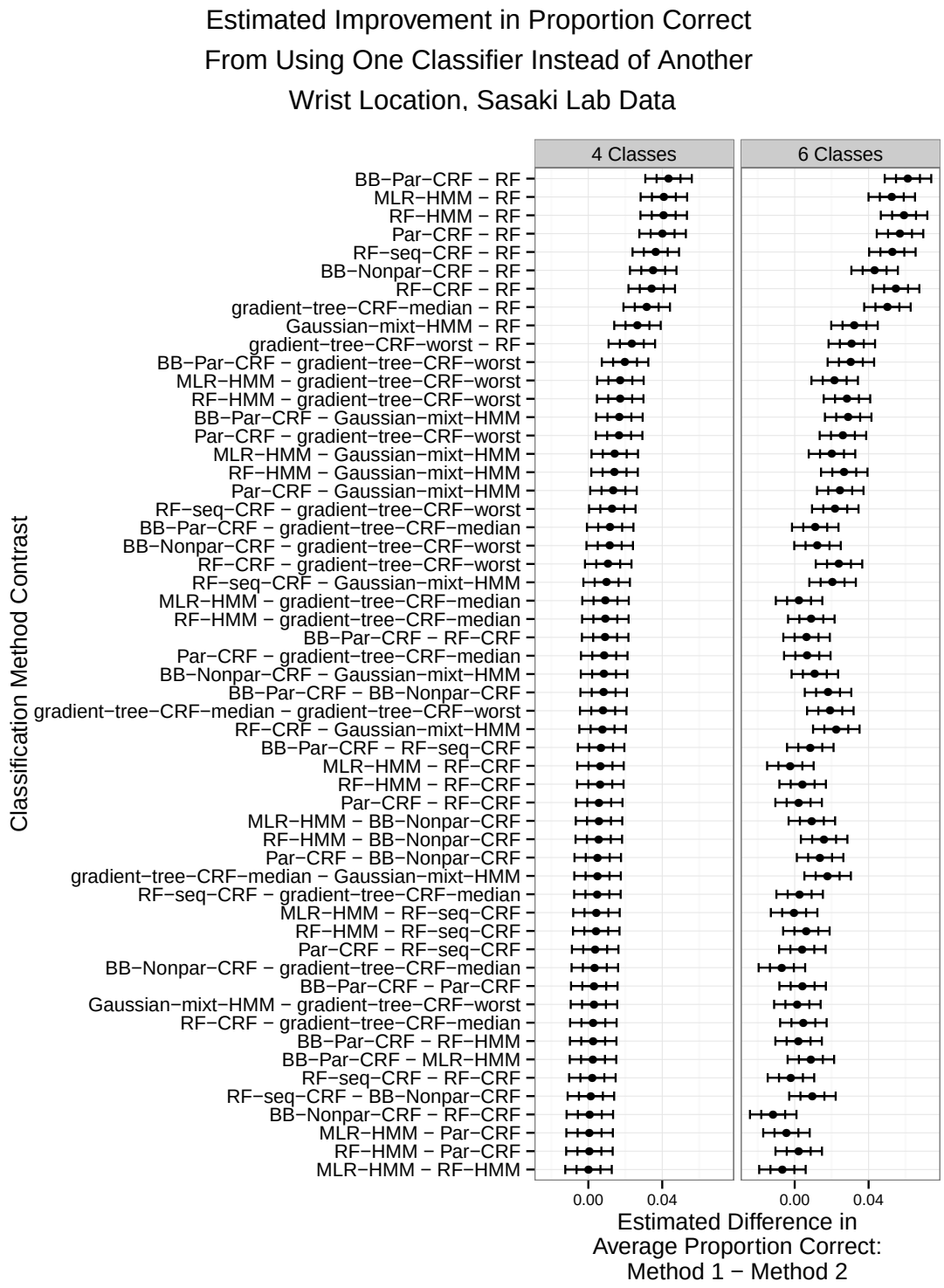


Figure 57. Point and interval estimates for the difference in average proportion of windows classified correctly for each pair of methods using data from the wrist. The estimates are based on model (7.4.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 53, 54, 55, 56, 57, and 58.**

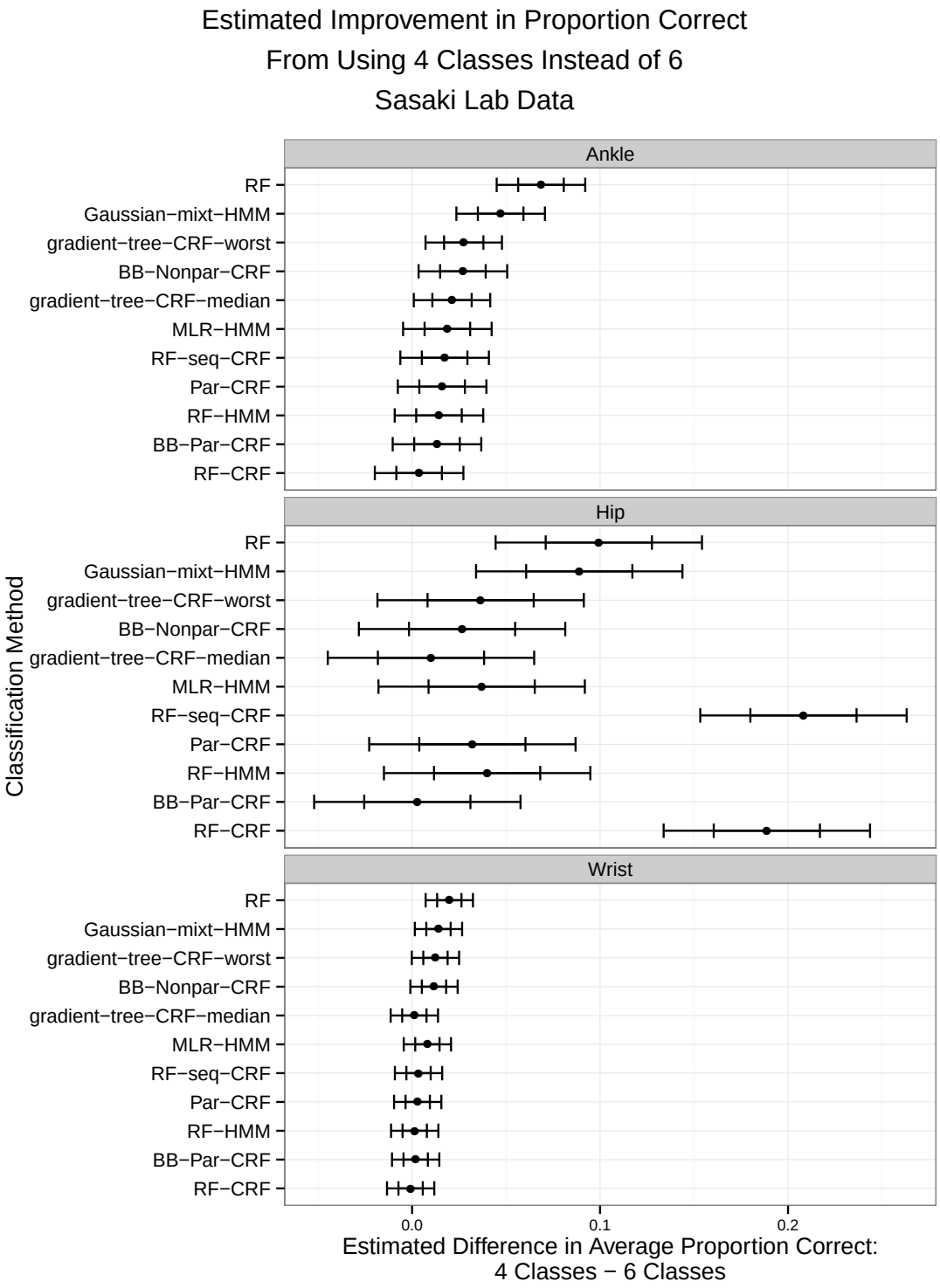


Figure 58. Point and interval estimates for the difference in the average proportion of windows classified correctly between the cases where 4 and 6 classes are used. The estimates are based on model (7.4.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 53, 54, 55, 56, 57, and 58.**

the results of the application to the free living data from Sasaki [2013]. Our first observation is that performance is much worse than it was with the laboratory data. This is consistent with previous work that has found classification to be more difficult in the free living setting than in the laboratory setting. We can also see that classification is more difficult with the data from the wrist than with data from the ankle or hip. This effect is more pronounced with 6 classes than with 4.

As with the previous data sets, there is a great deal of overlap in the box plots within each combination of the accelerometer location and the number of classes. The **Gaussian-mixt-HMM** method consistently achieves a slightly lower proportion correct than the other methods, as does the worst-case performance of the gradient-tree-CRF method. The **Par-CRF** method also suffers from poor performance with the wrist data and 6 classes. From the plots of the proportion of time points classified correctly and the macro F_1 score, it also appears that the **RF** method may offer slightly lower performance than the best of the other methods when there are 6 classes. However, the plot of MSE values offers an interesting counterpoint to this observation: the **RF** method gives a smaller range of MSE values that are roughly in line with the best of the other methods. Thus, if our objective is to minimize the worst-case difference between the estimated class probabilities and the true class labels, it appears that the **RF** method would be a reasonable choice.

Figure 63 displays the proportion correct for just the **BB-Par-CRF**, **RF-CRF**, **RF**, and **Gaussian-mixt-HMM** methods with the free living wrist data. We observe similar patterns to what we saw with the laboratory data and the data from Mannini et al. [2013], with a great deal of variability in the proportion correct achieved across different subjects.

Tables 16, 17, 18, and 19 show confusion matrices for these data aggregated across all subjects. Recreational activity was consistently difficult to distinguish from the other activity types. One explanation for this is simply that the free living data include a limited number of observations in this category. Of the 15 subjects who participated in the free living component of the study, only three engaged in activity that was categorized as Recreational, with a combined total of about 2 hours and 15 minutes spent in this activity category. Since we used a leave-one-subject-out procedure to train the models and some of the estimation procedures further subdivide the data into training and validation sets, the number of observations in the Recreational category used for estimating the model parameters could be quite small. Further, the Recreational cate-

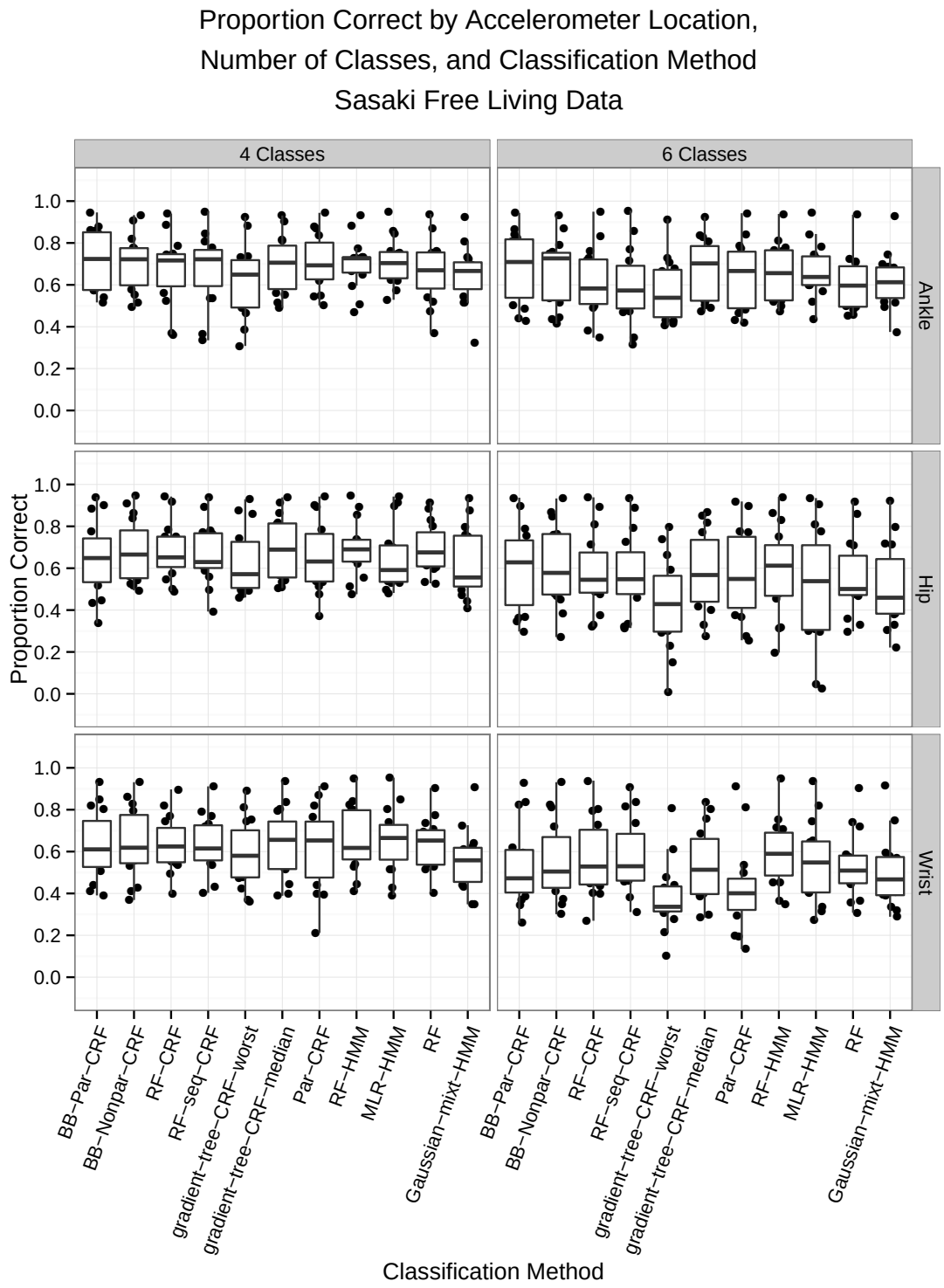


Figure 59. Box plots showing the proportion of windows classified correctly in the free living data from Sasaki [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and number of classes. Each point corresponds to a combination of accelerometer location, classification method, number of classes, and subject.

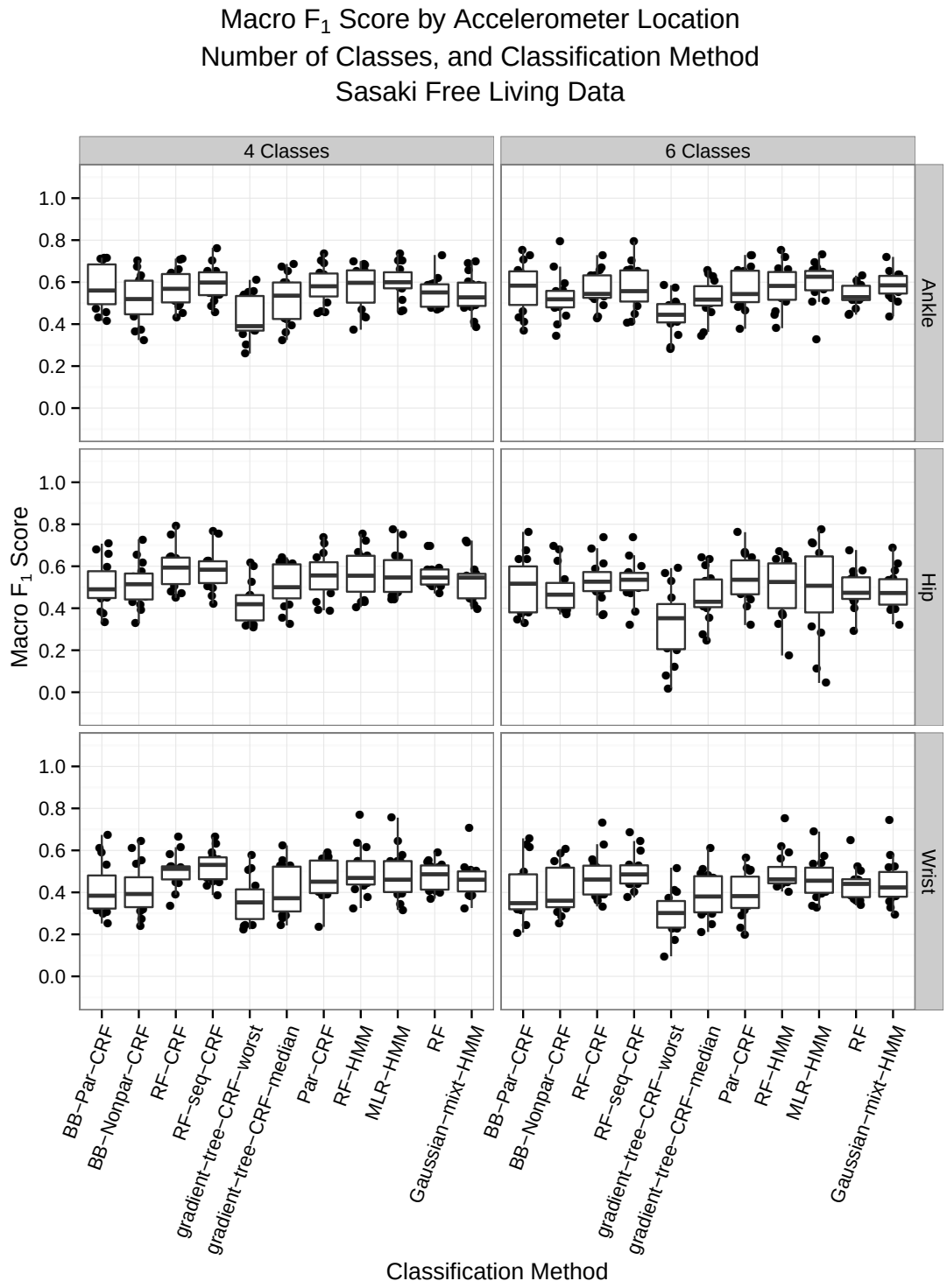


Figure 60. Box plots showing the macro F_1 score combining precision and recall across all four physical activity type categories in the free living data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and number of classes. Each point corresponds to a combination of accelerometer location, classification method, number of classes, and subject.

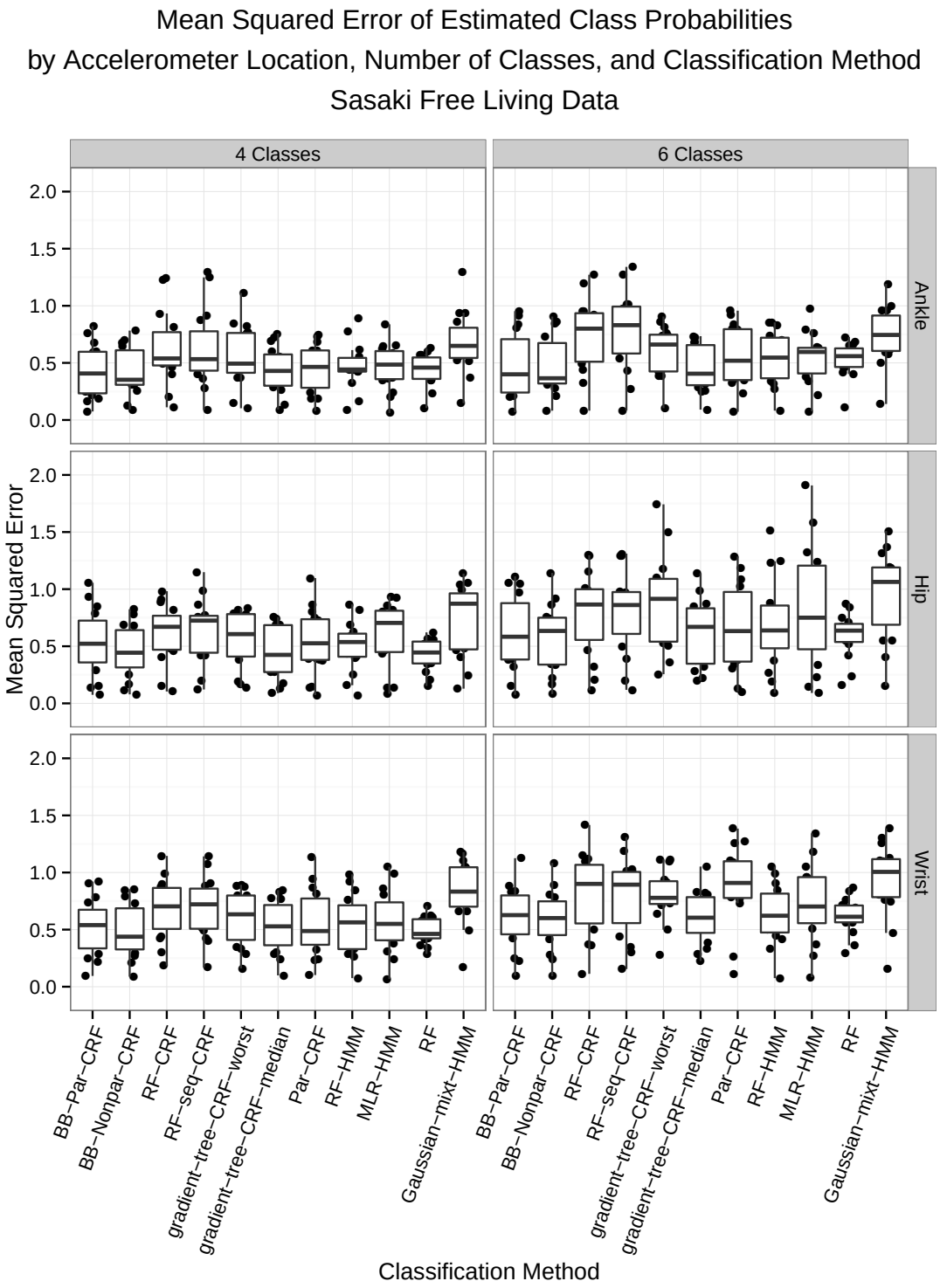


Figure 61. Box plots showing the mean squared error of the estimated classification probabilities relative to the labeled class memberships in the free living data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and number of classes. Each point corresponds to a combination of accelerometer location, classification method, number of classes, and subject.

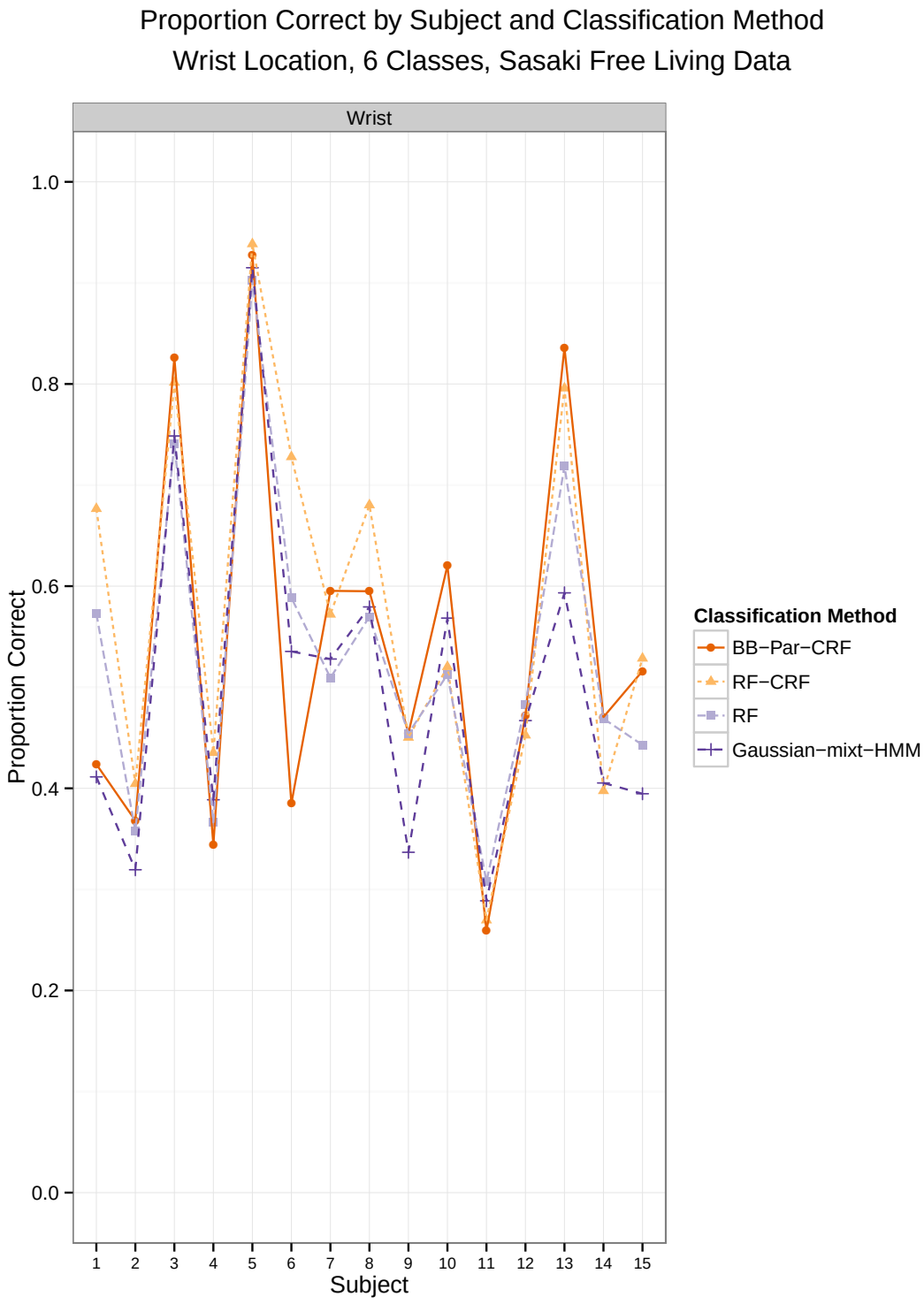


Figure 62. Proportion of time windows classified correctly by subject in the free living data from Sasaki [2013] with 6 classes, using a leave-one-subject-out procedure. We show these values for just four classification methods applied to data collected from accelerometers placed at the wrist location.

gory was formed by combining the Aerobic Exercise, Resistance Exercise, and Balance Exercise sub-categories. These sub-classes entail fairly different types of movement resulting in different patterns in the accelerometer signal. We believe that classification rates for this category could be improved with more training data.

For all of the classification methods, the Sedentary, Standing, and Moving Intermittently categories are often confused for each other. The Sedentary and Standing categories have proven to be difficult to distinguish using accelerometer data in previous studies as well, simply because these activity types can both entail limited movement and similar accelerometer orientations. However, we also believe that some of these disagreements between the labeled class and the estimated class may be due to mislabeled windows. Figure 63 presents some evidence for this. The figure shows the labeled and estimated classes along with the vector magnitude from the accelerometer placed at the ankle for subject 8 in the free living component of the study. In this figure, many windows that are labeled as Moving Intermittently are classified as Sedentary, Standing, Locomotion, or Transition. We cannot be sure of what the subject was doing during these time spans, but the predicted classes seem plausible in many cases. For example, there are several periods of length ranging from about 30 seconds to several minutes where the accelerometer recorded minimal movement at the ankle and the labeled class was Moving Intermittently. It seems possible that the subject was sedentary or standing still during these time periods. Similarly, Moving Intermittently could entail short periods of walking. Thus, it could be that for some of the windows when the labeled class if Moving Intermittently but the predicted class is Locomotion, the predicted class is an accurate description of the subject's behavior.

We fit the following linear mixed effects model to the classification results:

$$p_{c,l,k,i} = \beta_{c,l,k} + \gamma_i + \varepsilon_{c,l,k,i}, \text{ where} \tag{7.5.1}$$

$$\gamma_i \sim N(0, \sigma^2_{subj}) \tag{7.5.2}$$

$$\varepsilon_{c,l,k,i} \sim N(0, \sigma^2_{l,i})$$

This is the same model that we employed for the laboratory data. As with the models we discussed in the earlier Sections, diagnostic residual plots show that the distribution of residuals has a heavy left tail with several low outliers. Again, we primarily view this model as a descriptive tool to complement the box plots above, and the qualitative results were the same with a range of similar models for both the proportion correct and the macro F_1 score.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	1696	408	189	0	0	4
Standing	262	685	493	3	0	11
Moving Intermittently	79	224	1115	52	0	54
Ambulation	0	2	256	1298	0	14
Recreational	154	233	57	168	0	32
Transition	64	154	410	94	0	98

Table 16. Confusion matrix for the BB-Par-CRF classification method applied to the free living hip data from Sasaki [2013] with 6 classes, all subjects combined.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	1597	528	108	0	18	46
Standing	290	637	379	0	52	96
Moving Intermittently	119	148	1025	26	20	186
Ambulation	2	3	230	1202	16	117
Recreational	84	165	103	142	95	55
Transition	77	63	338	63	12	267

Table 17. Confusion matrix for the RF-CRF classification method applied to the free living hip data from Sasaki [2013] with 6 classes, all subjects combined.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	1614	577	55	0	33	18
Standing	332	731	249	0	108	34
Moving Intermittently	185	278	755	69	65	172
Ambulation	6	3	144	1247	30	140
Recreational	91	170	92	144	119	28
Transition	94	110	358	93	33	132

Table 18. Confusion matrix for the RF classification method applied to the free living hip data from Sasaki [2013] with 6 classes, all subjects combined.

Labeled Class	Predicted Class					
	Sedentary	Standing	Moving Intermittently	Ambulation	Recreational	Transition
Sedentary	1608	444	44	0	42	159
Standing	281	619	189	1	118	246
Moving Intermittently	97	254	410	95	49	619
Ambulation	2	0	143	999	1	425
Recreational	115	135	31	108	113	142
Transition	49	89	124	95	5	458

Table 19. Confusion matrix for the Gaussian-mixt-HMM classification method applied to the free living hip data from Sasaki [2013] with 6 classes, all subjects combined.

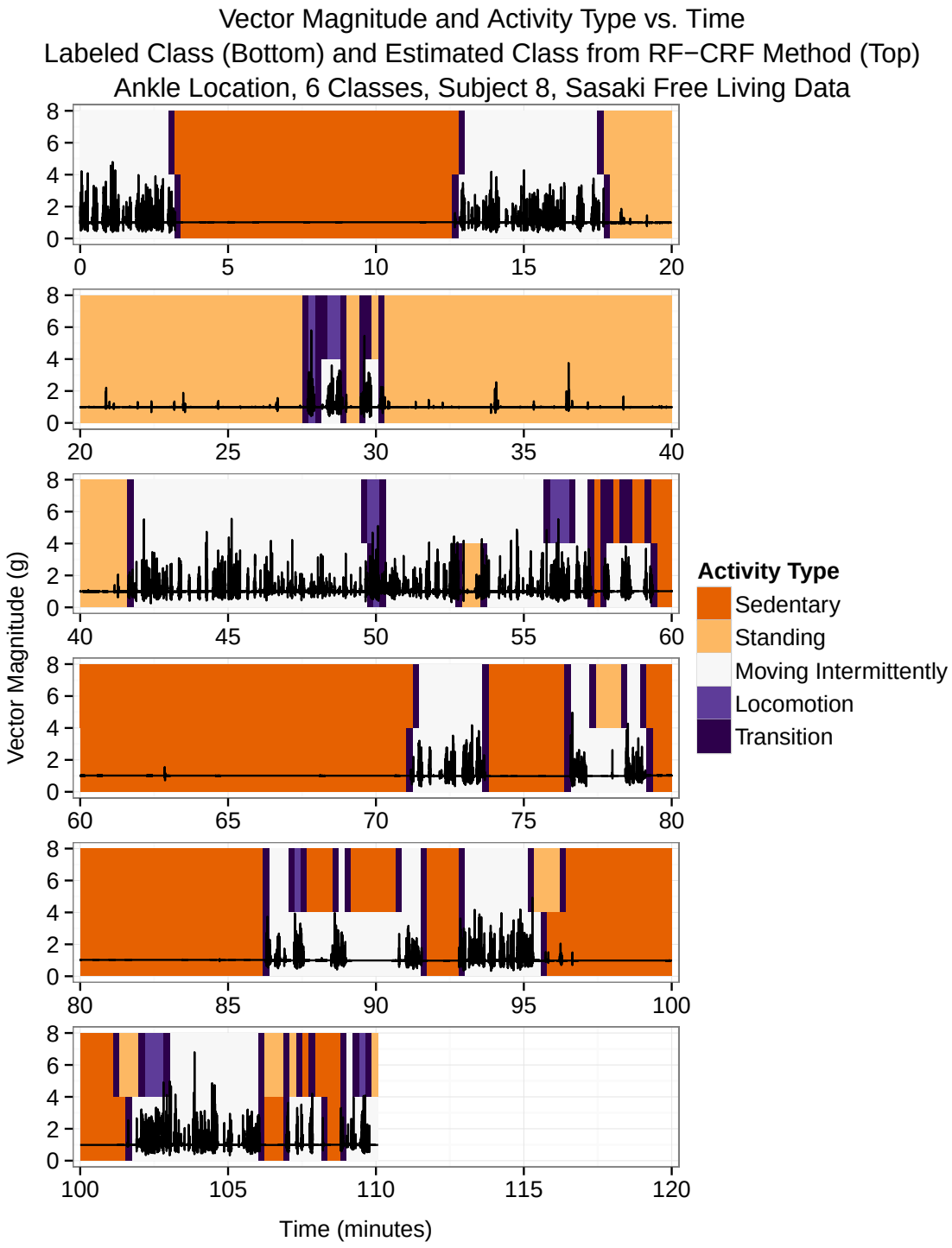


Figure 63. A plot showing the labeled class and the predicted class from the RF-CRF method for Subject 8 in the free living data from Sasaki [2013], using the ankle data with 6 classes. The background color indicates the labeled class in the bottom half of each panel and the predicted class in the top half of each panel. The black line indicates the vector magnitude of the accelerometer signal. No windows were classified as Recreational activity by either the direct observation labels or the classifier's predictions for this subject.

We display point and interval estimates for the average proportion correct for each combination of classifier, accelerometer location, and number of classes, along with sets of contrasts between these quantities in Figures 64, 65, 66, 67, 68, and 69. These estimates confirm the observations we made from the box plots above.

Among all of the factors considered here, the largest boost in the proportion correct is achieved by using data from an accelerometer placed on the ankle or hip instead of an accelerometer placed on the wrist. With 6 classes, point estimates for the increase in the proportion correct from changing the accelerometer location range from about 0.04 to 0.21, or between about 40 minutes and 3 and a half hours in a 16 hour day, depending on the classification method used. The gains are smaller if we consider only 4 classes.

As we saw from the box plots, the **Par-CRF** method had poor classification performance when using the wrist data and 6 classes. This is reflected in the estimates displayed in Figures 66, 67, and 68 comparing the classification methods to each other. However, the other dynamic discriminative methods consistently yield small improvements in performance relative to the **Gaussian-mixt-HMM** method for all combinations of the accelerometer location and number of classes, and relative to the **RF** method in the cases with 6 classes. The point estimates for these improvements range from about 0.01 to about 0.08; these gains correspond to as much as about 75 minutes if we extrapolate to a 16 hour day.

7.6 Discussion

In this Section we tie together the results of the applications to classification of physical activity type in our three data sets. There are many factors related to how the data are collected, preprocessed, and analyzed that can each affect the performance of the resulting classifier. In this work, we have focused on the effects that the data collection setting (i.e., in the laboratory vs. free living), the accelerometer location, the number of classes, and the model and estimation strategy have on classification performance.

As in previous studies, we found that classification performance decreased noticeably when we used free living data instead of data collected in the laboratory. For instance, with 6 classes and data from accelerometers placed at the ankle, the overall average proportion correct achieved by the most successful methods dropped from about 0.97 in the lab data from Sasaki [2013] to

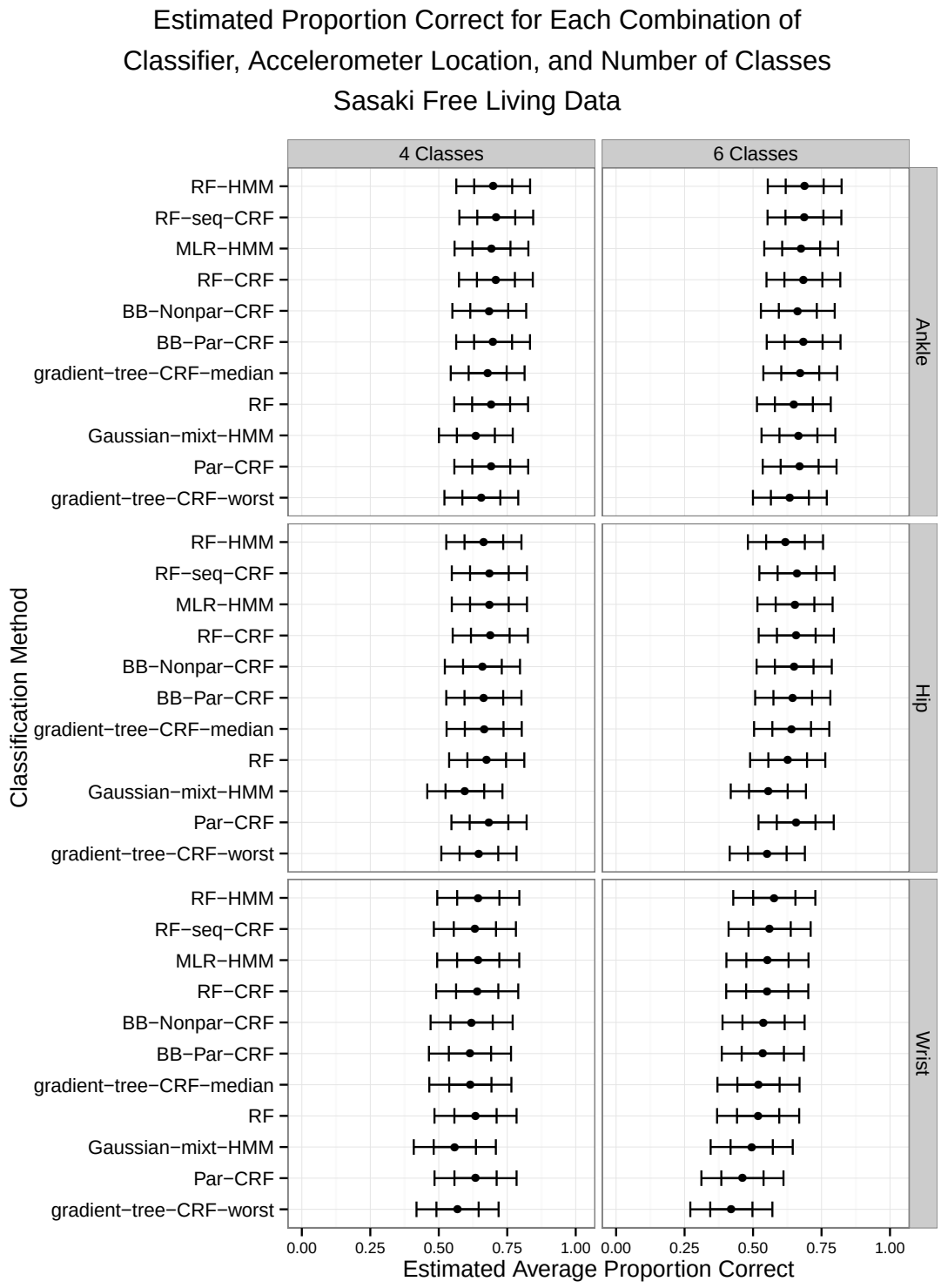


Figure 64. Point and interval estimates for the fixed effects parameters in model (7.5.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 64, 65, 66, 67, 68, and 69.**

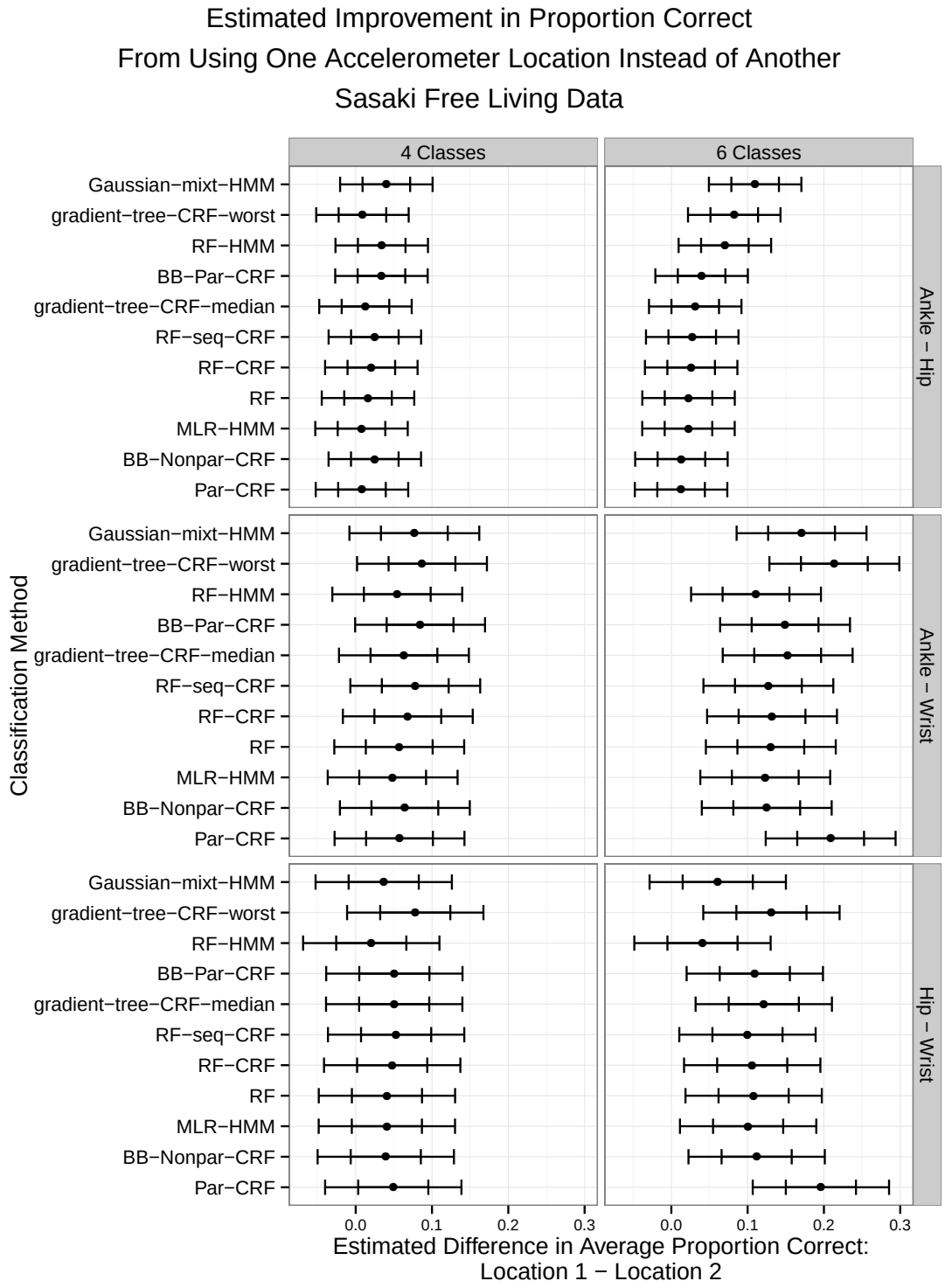


Figure 65. Point and interval estimates for the difference in performance between each pair of accelerometer locations, holding fixed the classification method and the number of classes. The confidence intervals are based on model (??). The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 64, 65, 66, 67, 68, and 69.

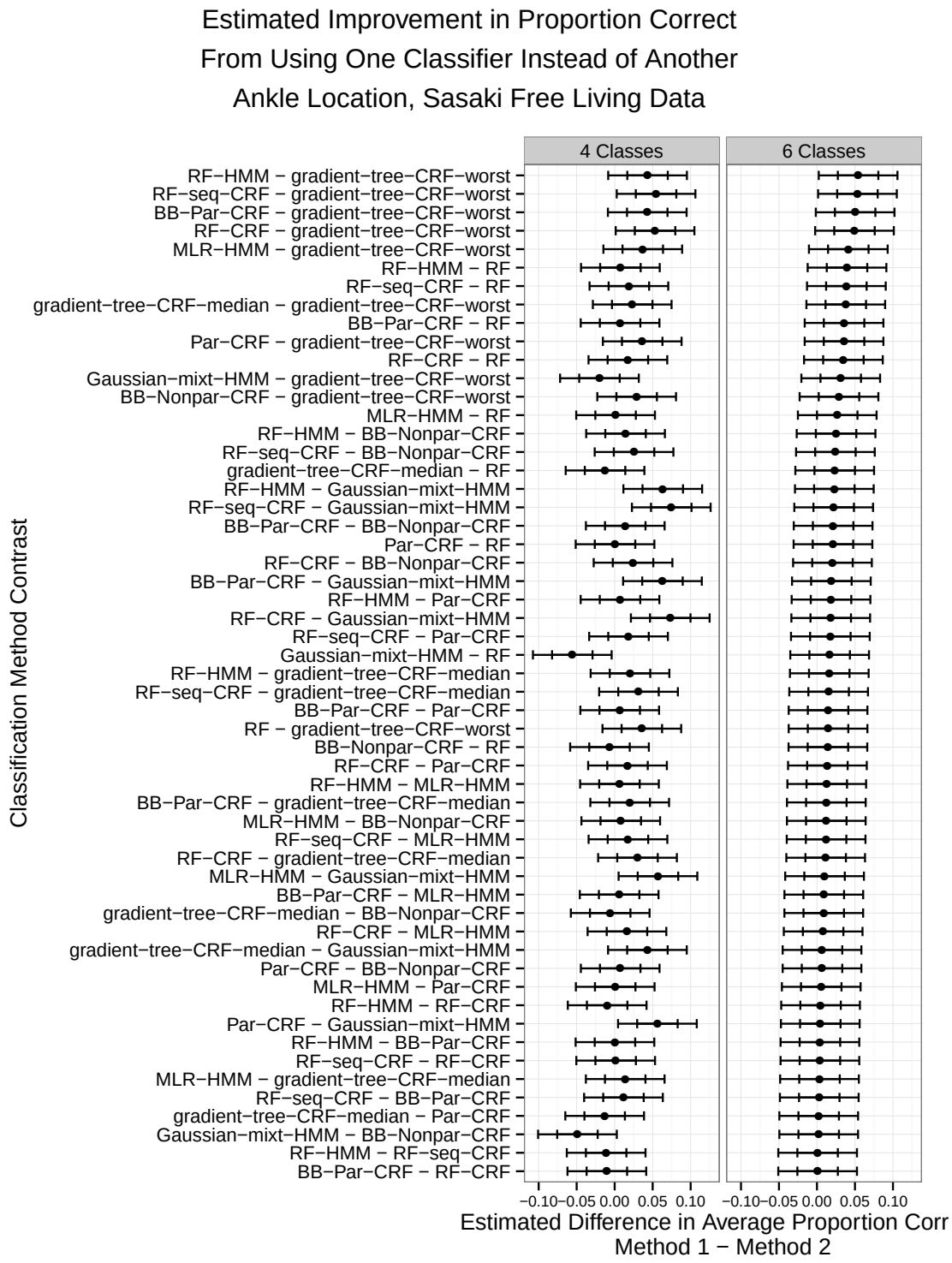


Figure 66. Point and interval estimates for the difference in average proportion of windows classified correctly for each pair of methods using data from the ankle. The estimates are based on model (7.5.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 64, 65, 66, 67, 68, and 69.**

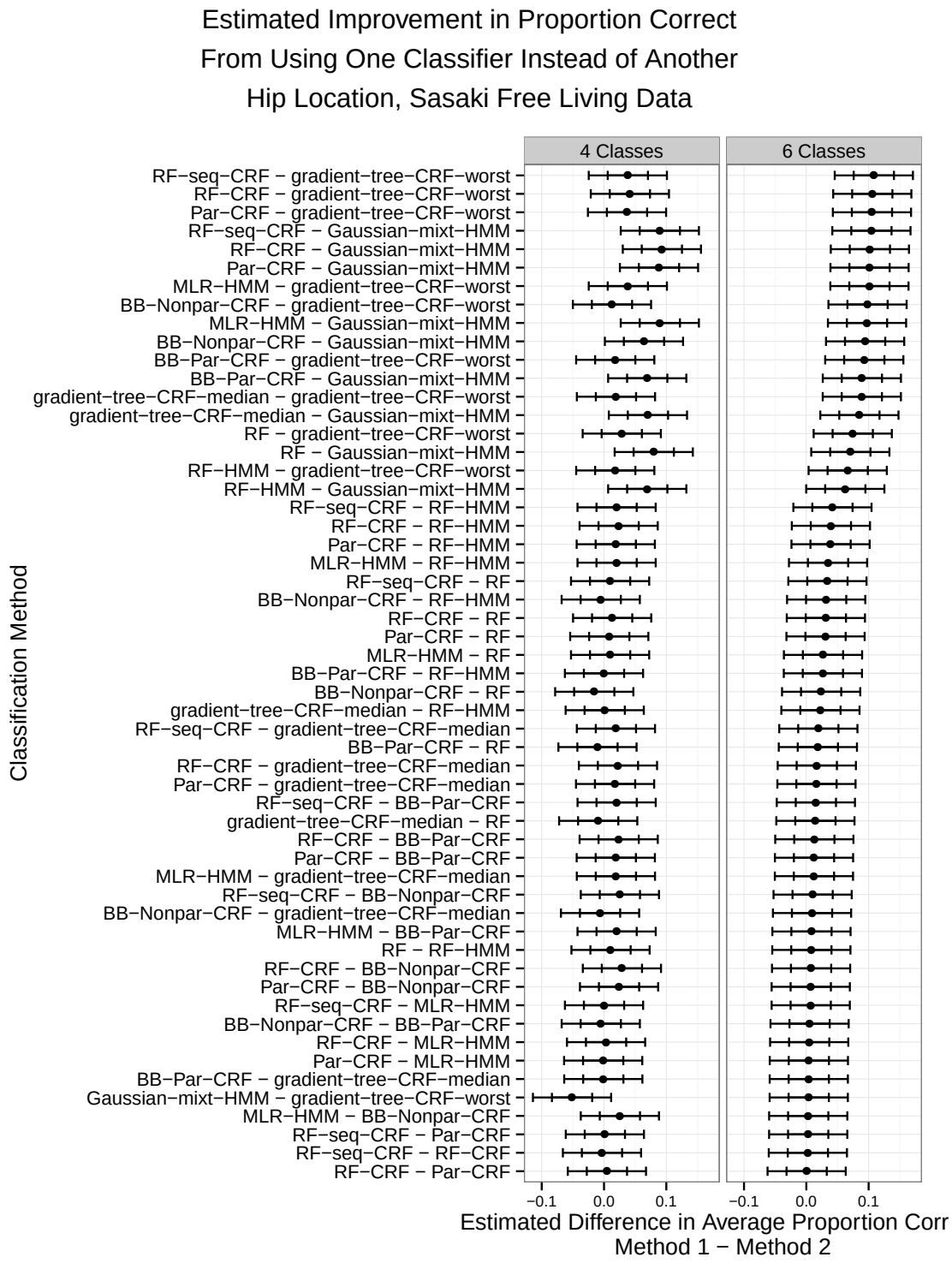


Figure 67. Point and interval estimates for the difference in average proportion of windows classified correctly for each pair of methods using data from the hip. The estimates are based on model (7.5.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 64, 65, 66, 67, 68, and 69.**

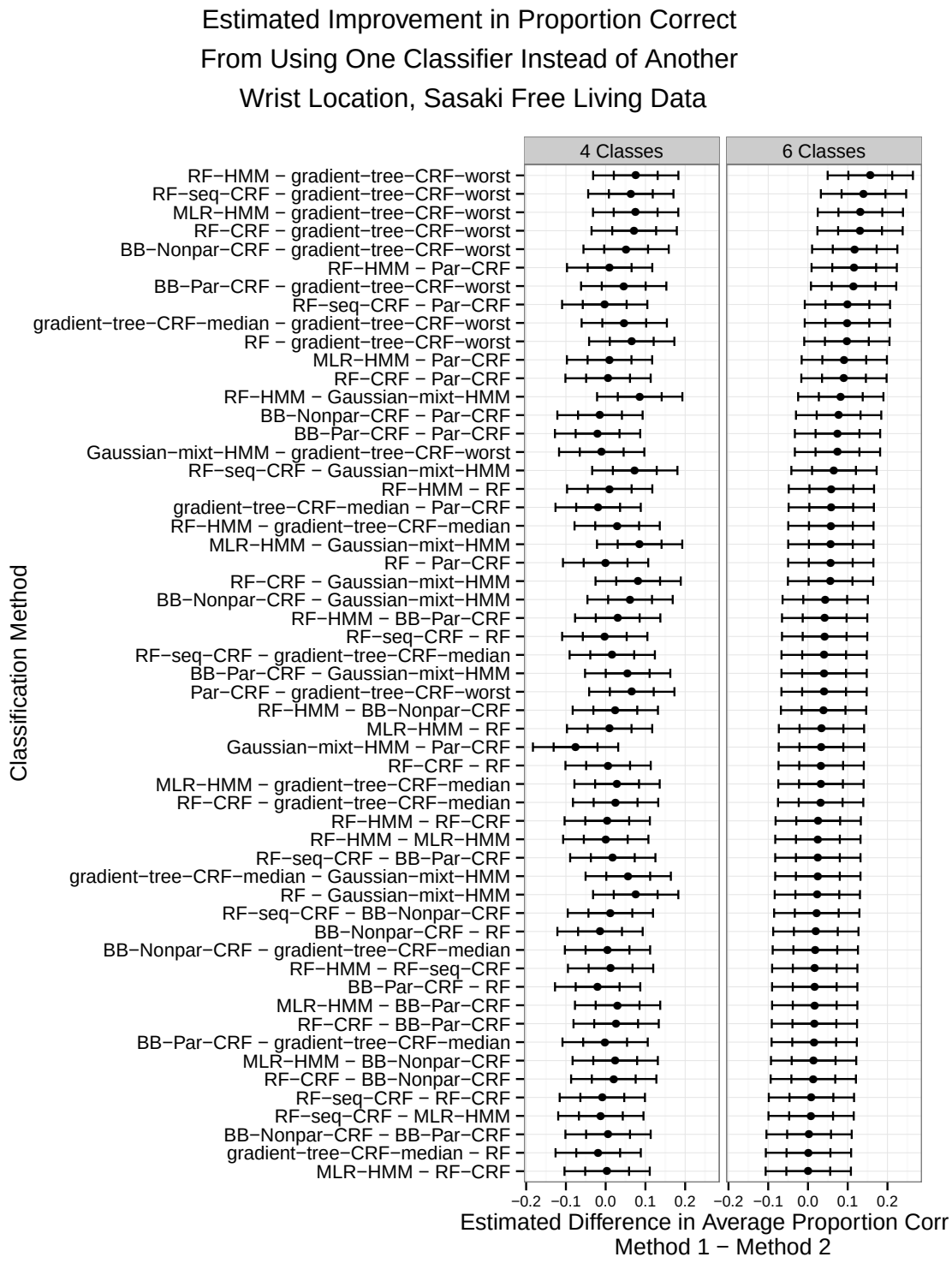


Figure 68. Point and interval estimates for the difference in average proportion of windows classified correctly for each pair of methods using data from the wrist. The estimates are based on model (7.5.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 64, 65, 66, 67, 68, and 69.**

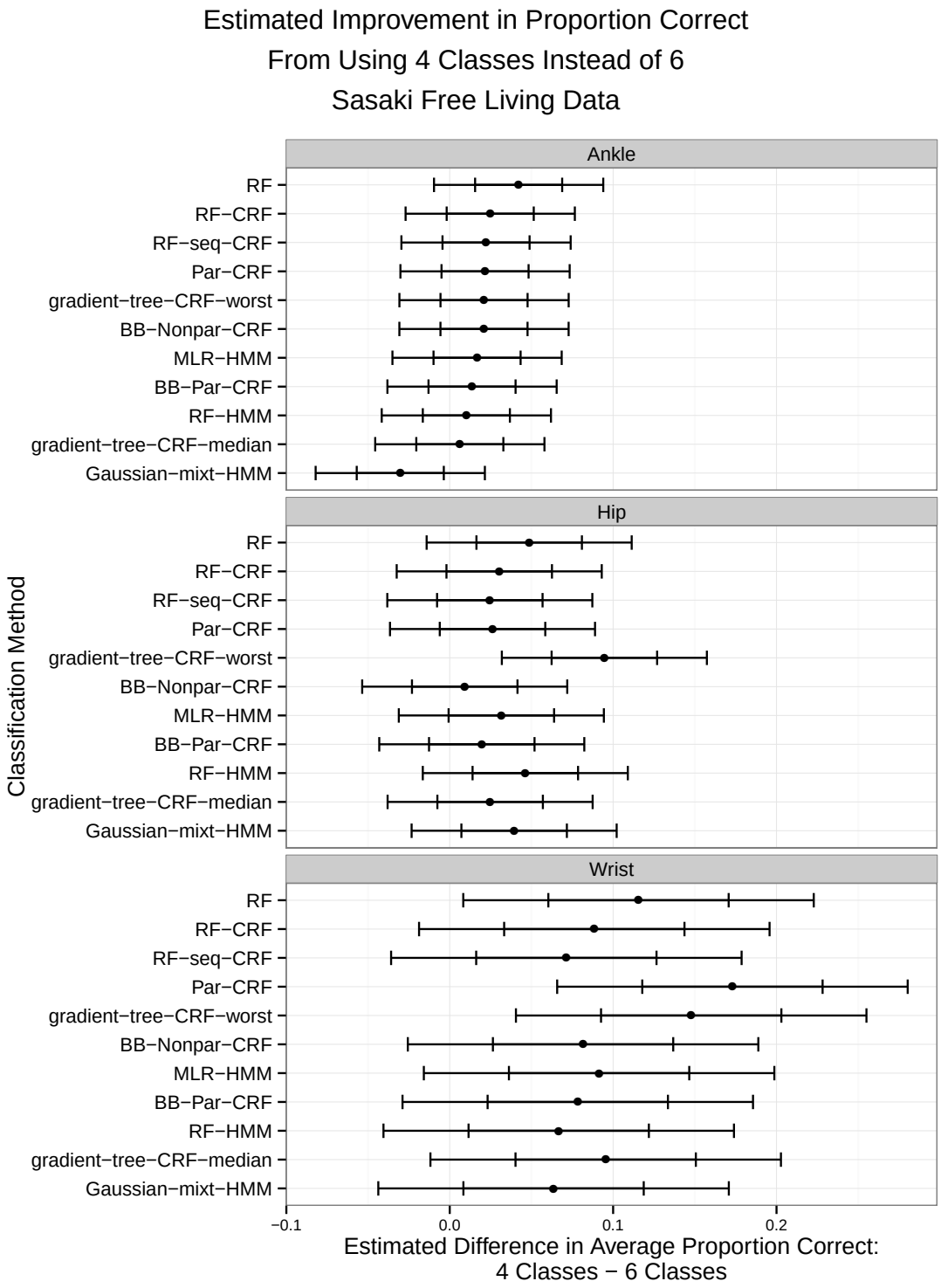


Figure 69. Point and interval estimates for the difference in performance between the cases where 4 and 6 classes are used. The estimates are based on model (7.5.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 64, 65, 66, 67, 68, and 69.**

about 0.7 in the free living data. This drop in performance may be partially due to the fact that the classification problem is fundamentally more challenging in the free living setting and partially due to limitations of the particular free living data set we are working with. We will discuss these issues in more depth in Chapter 9.

The three new estimation procedures for CRFs that we introduced in Chapter 5 had varying levels of success. The **BB-Par-CRF** method was among the best performing methods in all cases. It was matched by the **Par-CRF** method, which uses a much simpler and faster maximum likelihood estimation procedure, in every case but two: the lab hip data from Sasaki [2013] with 6 classes and the free living wrist data from Sasaki [2013] with 6 classes. In the first case, the estimated difference in the average proportion correct achieved by the two methods was about 0.04; in the second case, it was about 0.08. It appears that our bagging and boosting estimation strategy can reduce the potential for maximum likelihood estimation of the CRF model to overfit the training data in some cases. Another method that has a similar structure to our **BB-Par-CRF** is the **MLR-HMM** method of McShane et al. [2013]. The **MLR-HMM** consistently had lower classification performance than the **BB-Par-CRF** method.

Our **BB-Nonpar-CRF** method was also among the best performing methods in all cases. As we discussed in Chapter 5, this method makes two main modifications to the estimation strategy used in the **gradient-tree-CRF** method: we bag the observation sequences to create many training data sets and combine the resulting fits in a LOP, and we randomly select a subset of the features to use in each boosting iteration as a strategy for regularizing the model. It appears that of these two modifications, the first was more useful. In all cases, the performance of our **BB-Nonpar-CRF** method was about the same as the median performance achieved by the **gradient-tree-CRF** method across different partitions of the data into training and validation sets used to select the stopping point for boosting. The **BB-Nonpar-CRF** method did better than the worst case performance of the **gradient-tree-CRF** with respect to this partitioning of the data. Experiments not reported in detail here revealed that we could also stabilize the classification results of the **gradient-tree-CRF** method at about the median value by combining the results from the 10-fold cross validated fits in a LOP. In other words, in the applications to our data sets the main improvement offered by our method was not due to bagging or random selection of feature subsets, but instead the reduction in the Monte Carlo variability of the classifier performance that is achieved by combining the fits from multiple training sets in a LOP. This

benefit can be achieved equally well and with reduced computation time by using 10 fold cross validation instead of bagging with a large number of bags.

Our remaining methods, the **RF-CRF** and closely related **RF-seq-CRF**, offered inconsistent performance. In many cases, they were among the best performing methods. However, in other cases they suffered large drops in performance. We do not have a satisfying explanation for this behavior. The most similar methods in our comparison were the **RF-HMM** and **BB-Nonpar-CRF** methods; both of those approaches outperformed the **RF-CRF** method in many cases. Additionally, our scheme for bagging at the level of individual time points did not improve the performance of the method relative to bagging at the level of complete observation sequences. The performance of the **RF-CRF** and **RF-seq-CRF** methods was similar in all cases.

The **RF**, which does not account for temporal dependence, had among the lowest proportion of windows classified correctly for all accelerometer locations in the data from Mannini et al. [2013] and the lab data from Sasaki [2013]. We only applied the static **SVM** method to the data from Mannini et al. [2013], but it was the method with the worst performance on that data set. On these laboratory data sets, the estimated increase in the average proportion correct that could be realized by using one of the **RF-HMM**, **BB-Par-CRF**, or **BB-Nonpar-CRF** methods instead of a **RF** ranges from about .02 to 0.14, depending on the data set, the accelerometer location, and the number of classes used. This is equivalent to an increase of between about 20 minutes and about 2 hours and 15 minutes in a 16 hour day.

With the free living data from Sasaki [2013], the difference in performance levels between the **RF** method and the other methods is smaller. With 4 classes, the **RF** classifier did about as well as the other methods. With 6 classes the **RF** method was always among the three methods with the smallest average proportion of windows classified correctly, but the differences from the other methods were small. One possible explanation for this is the finding in the simulation study of Chapter 6 that the differences in performance between the static **RF** and dynamic models are lower when the Bayes error rate is high than when the Bayes error rate is small. Intuitively, when the Bayes error rate is high, the dynamic methods have less to gain from sharing information about activity type across nearby time points. If this explanation is correct, we would expect the performance gap between the static **RF** and the dynamic models to increase if we could improve the classification rates in the free living data. We will discuss ideas for how this could be achieved in Chapter 9. Regardless, the combined results across all three data sets offer evidence

that accounting for sequential dependence in activity types can lead to improved classification performance.

The **Gaussian-mixt-HMM** did fairly well with the data from Mannini et al. [2013]. However, it was at or near the bottom of the pack for all but one of the combinations of factors in the two data sets from Sasaki [2013]; the one case where it did relatively well here was with the ankle accelerometer location with 6 classes in the free living data. With 4 classes, the estimated improvement from using one of the **RF-HMM**, **BB-Par-CRF**, or **BB-Nonpar-CRF** methods instead of the **Gaussian-mixt-HMM** ranged from about 0.05 to 0.1, or between about 48 and 96 minutes in a 16 hour day. Outside of the case where the **Gaussian-mixt-HMM** performed about as well as the other methods, with 6 classes the point estimates for the gain in the average proportion correct that is achieved by using one of the **RF-HMM**, **BB-Par-CRF**, or **BB-Nonpar-CRF** methods ranged from about 0.01 to 0.09. While the **Gaussian-mixt-HMM** did well in some cases, the overall pattern was that it was among the methods with the worst performance.

The placement of the accelerometer can also have a large impact on classification performance. In all three data sets, the best classification performance was achieved using data from accelerometers placed at the ankle. Both Mannini et al. [2013] and Sasaki [2013] made this observation previously for static classification methods. Our contribution is an extension of this result to a wider range of classification models.

The relative success of classification using data from the other locations was less consistent. In the data from Mannini et al. [2013] and the free living data from Sasaki [2013], classification was more difficult using recordings from accelerometers placed at the wrist than it was using data from the ankle. However, in the laboratory component of the study by Sasaki [2013] classification performance was about the same using data from the wrist or the ankle. In both components of the study by Sasaki [2013], the performance of most methods was similar in the ankle and the hip. For those methods that did exhibit a difference in performance between these locations, performance was worse using the hip data than it was using the ankle data. Classification was easier with the wrist data than the hip data in the laboratory setting, but in the free living setting classification was easier with the hip data than the wrist data. The study by Mannini et al. [2013] did not include accelerometers at the hip.

The size of the gain in classification performance that could be achieved by switching to the ankle location varied quite a bit with the classification method used. As we observed above,

some classification methods were less robust to the accelerometer location than others. For instance, in the data from Mannini et al. [2013], the proportion correct from the **RF-HMM** method only improved by about 0.03 when the accelerometer was placed on the ankle instead of the wrist, while the proportion correct from the **Gaussian-mixt-HMM** improved by about 0.08 and the proportion correct from the **RF** improved by about 0.09. A similar relationship holds in the data from Sasaki [2013]. Overall, the **Gaussian-mixt-HMM** method most consistently exhibited large differences in performance from one accelerometer location to another across all three data sets. Additionally, the **MLR-HMM**, **RF-CRF**, **RF-seq-CRF**, and **Par-CRF** methods, which did well in most cases but experienced large drops in performance in some of the harder cases, had large differences in performance across different accelerometer locations for some of the data sets. However, even the most stable of methods often experienced large jumps in performance when the accelerometer was placed on the ankle. For example, in the free living data from Sasaki [2013] with 6 classes, the proportion correct improved by at least 0.11 for all methods when the accelerometer was placed on the ankle instead of the wrist.

CHAPTER 8

APPLICATIONS TO PHYSICAL ACTIVITY INTENSITY CLASSIFICATION

8.1 Introduction

In this Chapter, we apply the classification methods we developed in Chapter 5 and several competitor methods to classification of physical activity intensity in the three physical activity data sets we described in Chapter 3. Our goal is to understand how classification performance depends on the accelerometer location, the classification method used, and the treatment of information about an individual's activity type in estimating activity intensity.

As we discussed in the literature review, physical activity intensity is a continuous quantity, measured in METs. One MET corresponds to the energy used sitting at rest, and other energy expenditure levels are defined relative to that value. However, in this work we focus on categorization of intensity as either Sedentary (< 1.5 METs), Light (≥ 1.5 and < 3 METs), Moderate (≥ 3 and ≤ 6 METs), or Vigorous (> 6 METs) activity.

Our primary reason for pursuing classification rather than regression is that the direct observation data from the free living component of Sasaki [2013] include annotations of the categorical intensity level, but not records of numeric METs. We have focused on methods that are applicable to those data since our ultimate goal is to develop methods for use with free living data. These intensity categories are also of direct interest to researchers. Many hypothesized relationships between physical activity levels and health outcomes are stated in terms of time spent in these categories. Similarly, the Physical Activity Guidelines published by the U.S. Department of Health and Human Services [U.S. Department of Health and Human Services, 2008] are given in terms of time spent in these categories, so a measure of time spent in each intensity category is of interest to researchers studying the impacts of public policy.

Two general approaches to physical activity intensity estimation have been proposed in the literature, distinguished by whether or not physical activity type is estimated as a first step before estimating the intensity. In applications with linear models relating a simple univariate summary of the accelerometer signal in each second ("counts") to METs, Albinali et al. [2010] demonstrate that the two stage procedure can yield improved results relative to the one stage procedure for estimation of continuous MET values. To our knowledge, a careful comparison of the relative merits of these approaches has not been conducted for more flexible models that directly model activity intensity using a richer summary of the accelerometer recordings. We will discuss the differences between our work and the study of Albinali et al. [2010] in more detail in Section 8.6.

In order to gain some intuition into the two-stage procedure, it may be fruitful to cast estimation of physical activity intensity as a missing data problem. In addition to the accelerometer signal, some observation sequences are annotated with activity type labels and some are not. Typically, those sequences in the data sets used to estimate the model parameters do include activity type annotations, and new sequences where we wish to perform classification do not include activity type. When the activity type is available, intuition suggests that the activity type may provide useful information that we can use in estimating the intensity. One interpretation of two-stage procedures is that they proceed by imputing the activity types for observation sequences where the activity type at each time point is not available.

In this Chapter, we will investigate two questions related to the treatment of the activity type in estimating activity intensity:

1. When the activity type is available, does it provide extra information about the intensity that was not already available from the accelerometer signal?

2. When the activity type is not available, do the imputed values contain extra information about the intensity that was not available from the accelerometer signal?

We emphasize that these questions are different from the question of whether the true or imputed activity types provide any information at all about activity intensity; we are specifically interested in whether they provide information about intensity above and beyond the information that is present in the accelerometer signal. We believe that the answers to these questions likely depend on several factors, including the set of activity type categories that is used, the vector of features that are extracted from the accelerometer signal, and the classification method.

For instance, it seems likely that if the activity type classification provides a fine level of detail about what an individual is doing and the set of features extracted from the accelerometer is fairly limited, the activity type would be a useful input to a classification algorithm. On the other hand, if we use a limited set of activity classes and a richer set of features describing the accelerometer signal, the activity type may not provide any information about intensity that was not already contained in the accelerometer features. Our investigation will focus on the particular categorization schemes, features, and classification methods we have introduced previously.

The rest of this Chapter is organized as follows. We discuss our methods in Section 8.2, present the classification results for each of our three data sets in Sections 8.3 through 8.5, and conclude with a summarizing discussion in Section 8.6.

8.2 Methods

We gave a detailed description of the procedures we used to preprocess the data in Chapter 3. As a reminder, the windows were labeled as either Sedentary, Light, Moderate, or Vigorous physical activity, or as a Transition if the window contained more than one labeled intensity level. There are no windows with the Transition label in the data from Mannini et al. [2013]; the 12.8 seconds immediately before and after each transition were discarded.

In the free living data from Sasaki [2013] the intensity category at each time point was recorded in the direct observation logs. In the laboratory data from Sasaki [2013] we calculated the steady-state average energy expenditure in minutes 3 through 5 of each activity using recordings of breath-by-breath oxygen usage from an Oxycon metabolic cart, and then converted the resulting continuous MET values to the corresponding discrete intensity categories. In the data from Mannini et al., we combined the fine-grained categorizations of physical activity type recorded in the direct observation logs with the population average intensity level associated with each activity type as published in the Compendium of Physical Activities [Ainsworth et al., 2011] to obtain continuous MET values for each window, and then converted these to the corresponding discrete intensity categories.

We apply the same classification methods we used in the simulation study of Chapter 6 and the application to classification of activity type in Chapter 7 to classify the subjects' physical activity intensity in each time window. These methods are as follows:

1. **BB-Par-CRF**
2. **BB-Nonpar-CRF**
3. **RF-CRF**
4. **RF-seq-CRF**
5. **RF**
6. **RF-HMM**
7. **MLR-HMM**
8. **Gaussian-mixt-HMM**
9. **gradient-tree-CRF**
10. **Par-CRF**

As in Chapter 7, we use a leave-one-subject-out design to gauge the success of the classifiers: for each subject, we train the classification methods with the observed data for all of the other subjects and use the resulting trained models to predict the activity type in each window for the held-out subject. We summarize these predictions with the same statistics that we used in Chapter 7: the proportion correct, the macro F_1 score, and the mean squared error of the estimated class probabilities relative to the true class indicator. We also use the same treatment of the **gradient-tree-CRF** method that we employed in Chapter 7, reporting the median and worst-case results for the model fits obtained with 10 different partitions of the data into training and validation subsets.

We explored three approaches to using information about activity type in our models for activity intensity. Throughout this Chapter, we will refer to these approaches as our *estimation strategies*. In the first, we use a one stage procedure that makes no use of information about the activity type. In this case, the feature vectors used in the model are exactly the same as the features used for the models in Chapter 7.

Our second estimation strategy is also a one stage procedure, but now we use the observed activity type included in our labeled data sets as an input to estimating activity intensity. The exact way that we use the type information depends on the model parameterization. For the

parametric models, **Par-CRF** and **BB-Par-CRF**, we introduce a separate set of coefficients for each activity type label:

$$\boldsymbol{\beta}_s = (\boldsymbol{\beta}'_{s,1}, \ldots, \boldsymbol{\beta}'_{s,K})', \tag{8.2.1}$$

where $\boldsymbol{\beta}_s$ is the vector of coefficients associated with intensity class s, which is partitioned into a separate group of coefficients $\boldsymbol{\beta}_{s,k}$ for each activity type class k. We also modify the feature vectors, setting

$$\mathbf{x}_{i,t} = (\mathbf{x}'_{i,t,1}, \ldots, \mathbf{x}'_{i,t,K})' \text{ where} \tag{8.2.2}$$

$\mathbf{x}'_{i,t,k}$ is the vector of features extracted from the accelerometer (including a leading 1 corresponding to the intercept coefficient) in window t, subject i if the activity type at that time is class k, and $\mathbf{x}'_{i,t,k} = \underline{0}$ if the activity type at time t is not k.

For the remaining models, we simply augmented the feature vectors with indicator variables giving the activity type class membership:

$$\mathbf{x}_{i,t} = (\tilde{\mathbf{x}}'_{i,t}, z_{i,t,1}, \ldots, z_{i,t,K})' \text{ where} \tag{8.2.3}$$

$\tilde{\mathbf{x}}_{i,t}$ is the vector of features that were extracted from the accelerometer in window t, subject i (without a leading 1, as there isn't an intercept term in these models) and $z_{i,t,k} = 1$ if the activity type at time t is class k and 0 otherwise. All of our nonparametric CRF models use flexible classification or regression trees that allow for conjunctions between the indicator variables $z_{i,t,k}$ and the accelerometer features. We also expected that the **Gaussian-mixt-HMM** method would be flexible enough to handle this parameterization; as we will see later, this may not be the case.

Our third estimation strategy is a two-stage procedure in which we impute the activity type labels in the first stage, and then use those imputed values in our model for activity intensity. In previous work with such two-stage procedures, it has been most common to perform a "hard" classification of activity type in the first stage of the estimation process and use a separate model for each activity type to map the accelerometer features to intensity estimates in the second stage. However, we prefer to capture the uncertainty in the estimated class membership from the first stage by using the estimated class membership probabilities obtained from the activity type classifier instead of indicators of the estimated class. For the parametric models, the coefficient vector has the same form as in Equation (8.2.1). However, we now define the feature vector as

follows:

$$\mathbf{x}_{i,t} = (\widehat{p}_{i,t,1}\tilde{\mathbf{x}}'_{i,t}, \ldots, \widehat{p}_{i,t,K}\tilde{\mathbf{x}}'_{i,t})'.$$

Here, $\widehat{p}_{i,t,k}$ is the estimated probability from the first stage that the activity type at time t is class k and $\tilde{\mathbf{x}}_{i,t}$ is the vector of features extracted from the accelerometer signal in window t, augmented with a 1 corresponding to an intercept term for each activity type class. We can consider this model formulation as a CRF where the final coefficient values are an observation-specific weighted sum of the coefficients from type-specific models: $\beta_{s,d}^{(i,t)} = \sum_{k=1}^{K} \widehat{p}_{i,t,k}\beta_{s,k,d}$. The coefficient weights are the class membership probabilities for the observation at time t from the first stage of modeling. In the non-parametric models, we augment the observed feature vectors with the estimated class membership probabilities for each class. We will sometimes refer to these estimated activity type probabilities as the imputed types.

We must take special care in estimating the parameters for this two stage procedure. When we apply the model fit to a new data set, both the activity type and the activity intensity are unobserved. The second stage model relates the accelerometer features and the imputed activity type class estimates to the intensity categories. Importantly, those imputed class memberships are obtained from a first stage model that never saw the true class labels in the new data set.

In order to replicate this in training the two-stage models, we use a leave one subject out procedure during the estimation of the first-stage classifier. The overall estimation procedure is as follows:

1. For each subject i,
   (a) Use all observation sequences except for the ith to estimate a first-stage model that uses accelerometer features to classify activity type.
   (b) Use this model fit to obtain estimated activity type class membership probabilities for subject i.
2. Use the accelerometer features and the activity type class membership probabilities from step 1 to estimate the second stage model that classifies according to intensity category.

In order to evaluate the success of the methods, this entire process is embedded within leave one subject out cross validation.

The nested cross validation procedure we described above is very time consuming, so we have only implemented the two stage estimation strategy for four of the classification methods: **BB-Par-CRF**, **BB-Nonpar-CRF**, **RF-CRF**, and **RF-HMM**. These are the three methods we proposed in Chapter 5, and the competetitor method that was the most successful in the applications of Chapter 7. For the data from Sasaki [2013], we focused on first-stage classifiers that worked with the four-class system of categories. We chose this set of classes because we saw in Chapter 7 that classification rates were higher using four classes than they were using six classes.

8.3 Mannini *et al.* Data

Figures 70 through 72 display box plots summarizing the results of the application to intensity level classification with the data from Mannini et al. [2013]. As with classification according to physical activity type, most methods achieved a higher proportion correct using data from the ankle than they did using data from the wrist. If we consider the proportion of windows classified correctly, the CRF-based models all perform about as well as each other and the static **RF** algorithm. These methods all outperform the **Gaussian-mixt-HMM** method and the **MLR-HMM** method, as well as the **RF-HMM** method with the ankle data. It is interesting that the **Gaussian-mixt-HMM** method outperformed the **RF** method in classifying according to activity type with the data from Mannini et al. [2013], but the **RF** method did better when classifying according to intensity.

The plot of the macro F_1 scores shows that although the **RF-HMM** method classified a lower proportion of time points correctly and had higher MSE than the other methods when using the ankle data, it had a higher F_1 score. The confusion matrices in Tables 20 through 25 explain why. The majority of windows had a true class of either Sedentary or Moderate intensity. Methods such as **BB-Par-CRF** are therefore able to achieve a high proportion correct by classifying most time points as either Sedentary or Moderate activity. This means that many Light and Vigorous windows are misclassified, resulting in a lower F_1 score. The **RF-HMM** makes a different trade-off, misclassifying some Sedentary and Moderate activity, but also getting more of the Light and Vigorous windows right. This results in an overall lower proportion correct, but a higher F_1 score. A choice of which method to use must take into account a prioritization of how important it is to achieve correct classification for each of the different intensity categories. It is also

interesting to note that most misclassification that does occur is between Sedentary and Light activity, or between Moderate and Vigorous activity. However, when the true activity type is unknown the **BB-Par-CRF** method misclassifies about 5% of windows where the true intensity is Moderate as Sedentary activity, a more serious error.

The final observations we can make from the plots in Figures 70 through 72 relate to the strategy for handling activity type when estimating activity intensity. Using the true activity type is generally beneficial if it is available. This holds for every classification method except for the **Gaussian-mixt-HMM**, which actually had lower performance when using the true activity type. It is possible that the **Gaussian-mixt-HMM** method would have performed better if we had used the activity type information differently. For that model, we simply augmented the observed feature vectors with indicator variables giving the estimated class membership. An alternative would be to use the same parameterization that we used for the **BB-Par-CRF** and **Par-CRF** methods, described in Section 8.2 above.

When the activity type is unknown, it appears that using a two-stage procedure that imputes the activity type can lead to small improvements in performance. The performance of the two-stage procedure generally falls inbetween the performance of the one stage method that does not use any information about the activity type and the one stage method that uses the true activity type. We can see this more clearly in the plots of Figure 73, which break the results down by subject. For the data from the ankle, the two-stage estimation strategy yields performance that is almost as good as when the activity type is known. This is perhaps because classification according to type was so effective: with the ankle data, the imputed activity types obtained in the first step are often very close to the true activity types. The story is less clear when we look at the wrist data. While the **BB-Par-CRF** and **RF-CRF** methods tend to do better with an imputed estimate of the activity type than without it, the performance of the **BB-Nonpar-CRF** and **RF-HMM** methods is about the same whether or not the type is imputed in a first stage.

We fit the following linear mixed effects model to the classification results:

$$p_{c,e,l,i} = \beta_{c,e,l} + \gamma_i + \varepsilon_{c,e,l,i}, \text{ where} \tag{8.3.1}$$

$$\gamma_i \sim N(0, \sigma^2_{subj}) \tag{8.3.2}$$

$$\varepsilon_{c,e,l,i} \sim N(0, \sigma^2_{l,i})$$

Here, $p_{c,e,l,i}$ is the proportion of windows classified correctly using classifier c, estimation strat-

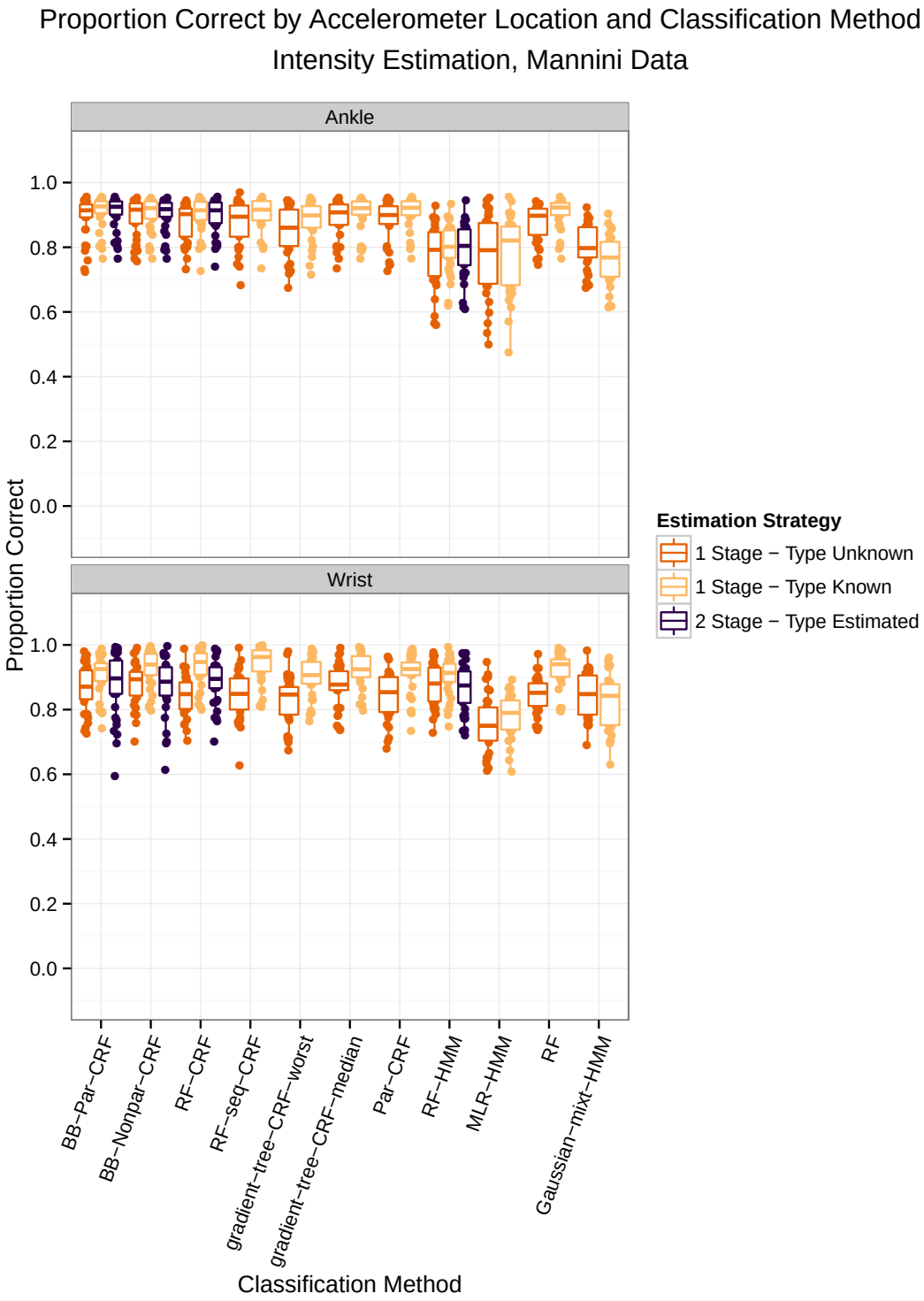


Figure 70. Box plots showing the proportion of windows classified correctly in the data from Mannini et al. [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

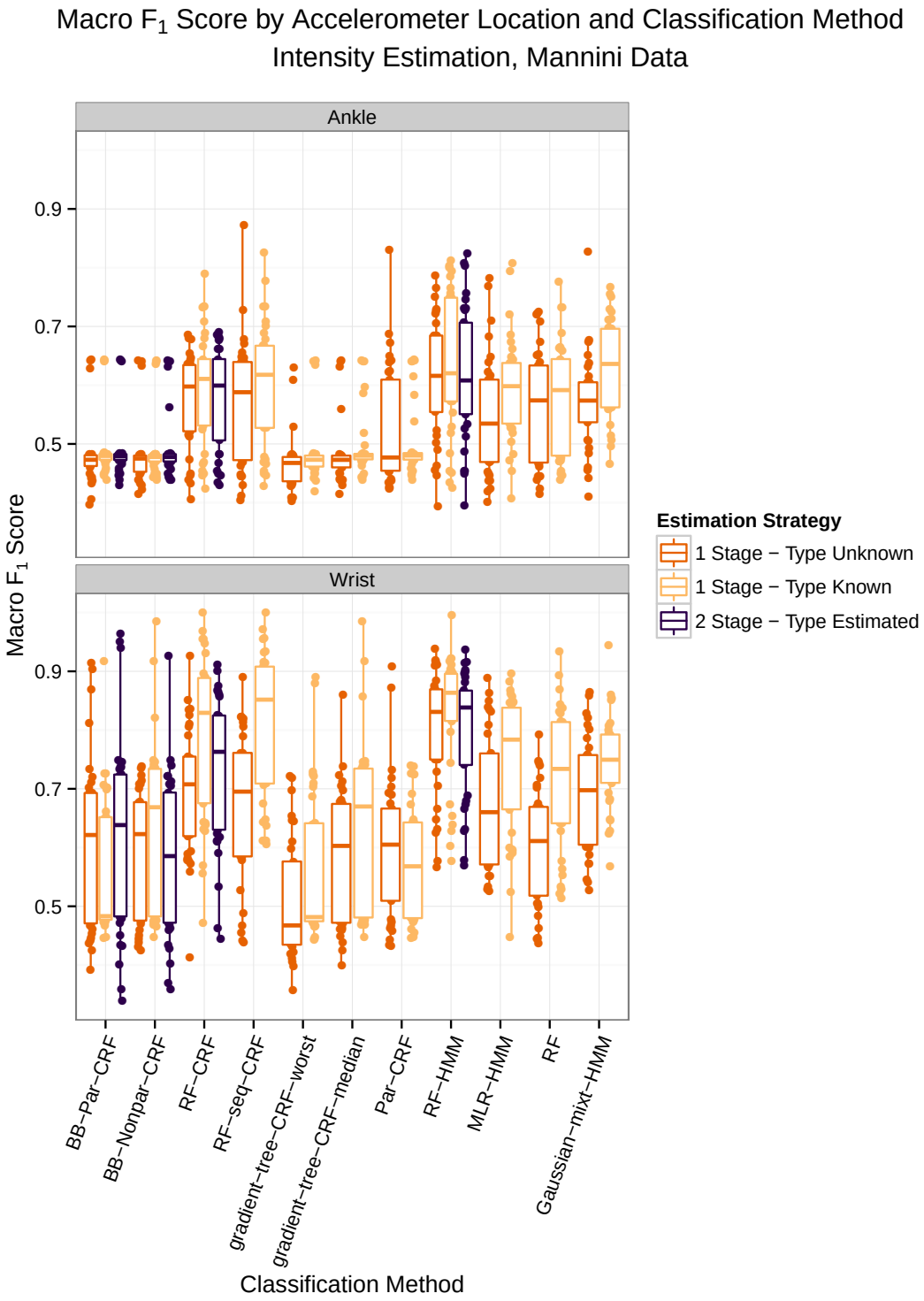


Figure 71. Box plots showing the macro F_1 score combining precision and recall across all physical activity intensity categories in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

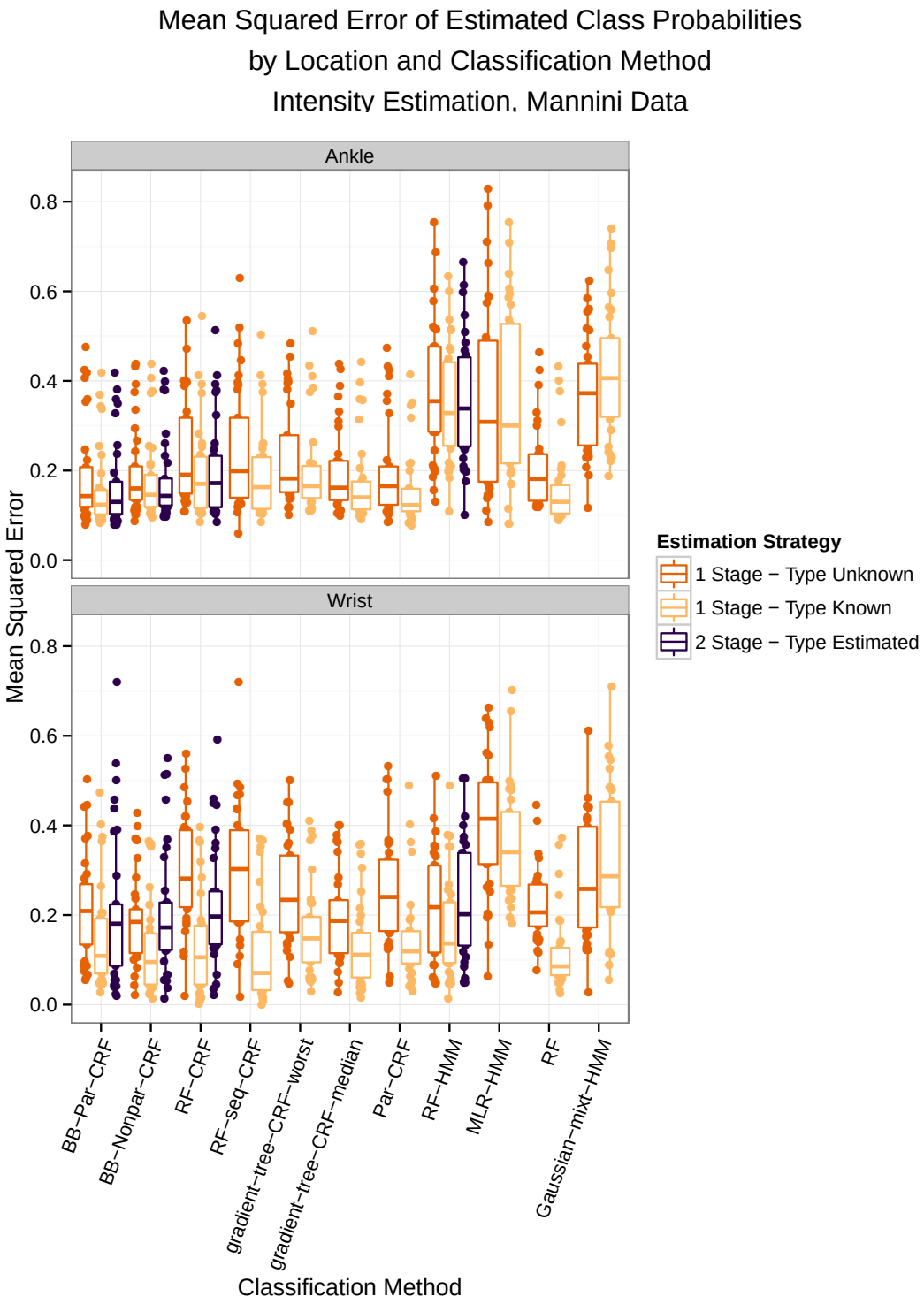


Figure 72. Box plots showing the mean squared error of the estimated class membership probabilities relative to the labeled class memberships in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

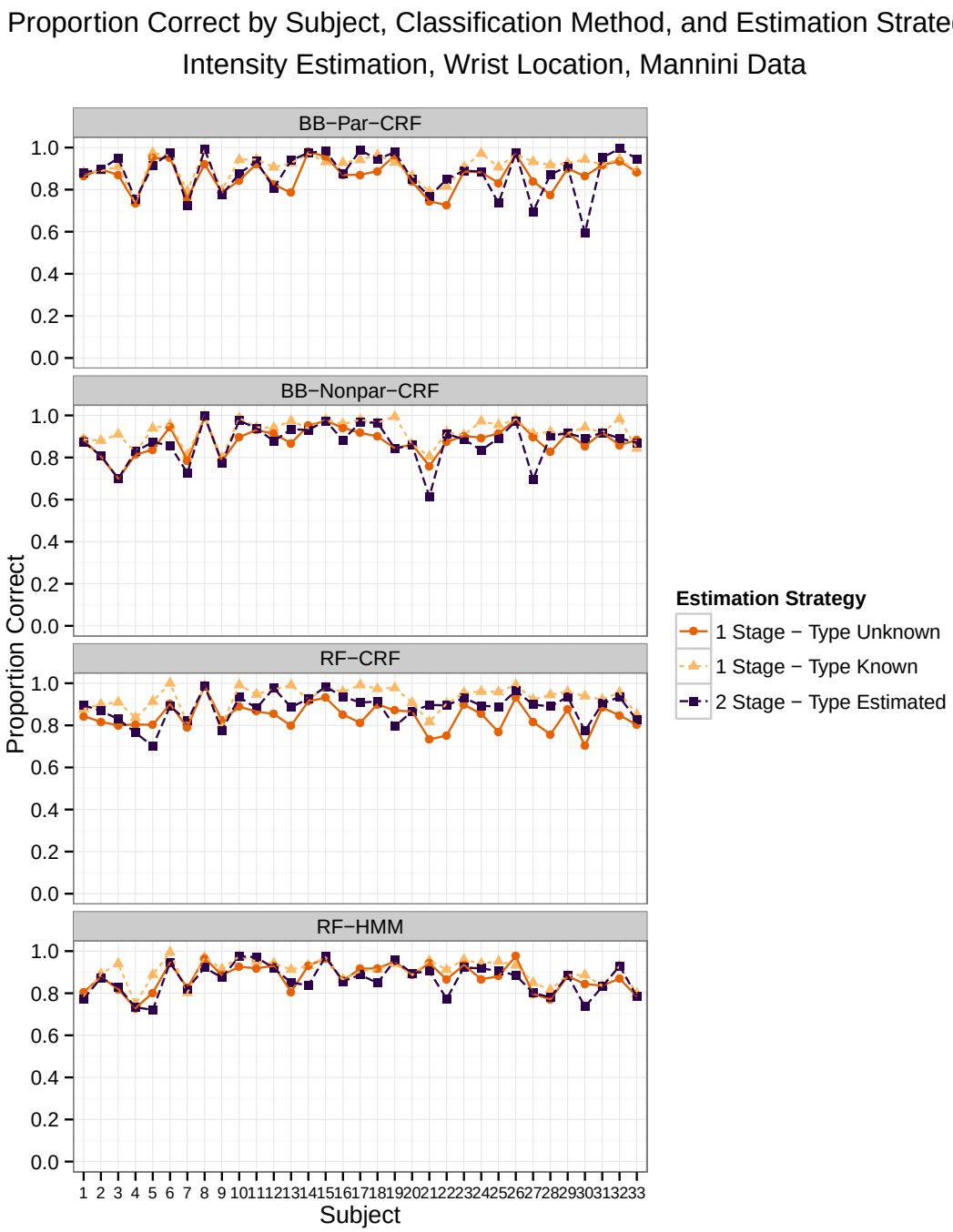


Figure 73. Proportion of time windows with intensity level classified correctly by subject in the data from Mannini et al. [2013], using a leave-one-subject-out procedure. We show these values for just the BB-Par-CRF, BB-Nonpar-CRF, RF-CRF, and RF-HMM classification methods applied to data collected from accelerometers placed at the wrist location. These are the four classification methods where we used all three estimation strategies.

	Predicted Class			
Labeled Class	Sedentary	Light	Moderate	Vigorous
Sedentary	2463	158	5	0
Light	25	333	0	0
Moderate	1	2	3677	412
Vigorous	0	0	117	210

Table 20. Confusion matrix for the RF-HMM classification method applied to the wrist data from Mannini et al. [2013] with a one stage estimation strategy where the true activity type is known for each window, all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Light	Moderate	Vigorous
Sedentary	2398	171	50	7
Light	25	333	0	0
Moderate	93	36	3559	404
Vigorous	0	0	115	212

Table 21. Confusion matrix for the RF-HMM classification method applied to the wrist data from Mannini et al. [2013] with a one stage estimation strategy where the true activity type is unknown for each window, all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Light	Moderate	Vigorous
Sedentary	2355	196	69	6
Light	25	333	0	0
Moderate	35	10	3507	540
Vigorous	0	0	113	214

Table 22. Confusion matrix for the RF-HMM classification method applied to the wrist data from Mannini et al. [2013] with a two stage estimation strategy where the unobserved activity type is imputed for each window, all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Light	Moderate	Vigorous
Sedentary	2340	123	163	0
Light	152	204	2	0
Moderate	233	3	3856	0
Vigorous	0	0	327	0

Table 23. Confusion matrix for the BB-Par-CRF classification method applied to the wrist data from Mannini et al. [2013] with a one stage estimation strategy where the true activity type is known for each window, all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Light	Moderate	Vigorous
Sedentary	2331	121	174	0
Light	142	213	3	0
Moderate	228	3	3861	0
Vigorous	0	0	327	0

Table 24. Confusion matrix for the BB-Par-CRF classification method applied to the wrist data from Mannini et al. [2013] with a one stage estimation strategy where the true activity type is unknown for each window, all subjects combined.

	Predicted Class			
Labeled Class	Sedentary	Light	Moderate	Vigorous
Sedentary	2364	86	176	0
Light	117	222	19	0
Moderate	173	2	3917	0
Vigorous	0	0	327	0

Table 25. Confusion matrix for the BB-Par-CRF classification method applied to the wrist data from Mannini et al. [2013] with a two stage estimation strategy where the unobserved activity type is imputed for each window, all subjects combined.

egy e, and the data gathered with the accelerometer placed at location l on subject i. The parameters $\beta_{c,e,l}$ are fixed effects representing the mean proportion correct for classifier method c and accelerometer location l, γ_i are random effects representing subject-specific offsets to the mean proportion correct, and $\varepsilon_{c,e,l,i}$ are error terms with variance specific to the combination of accelerometer location and subject. The considerations in formulating the model are similar to those we discussed for the models in Chapter 7.

Figures 74 through 79 display point and interval estimates for the $\beta_{c,l}$ parameters and sets of contrasts between them. As before, we caution against a strict interpretation of the confidence intervals; however, they do give a rough sense of the uncertainty in the point estimates.

The estimates displayed in Figure 75 show that unlike activity type classification, classification of intensity is not necessarily easier with data from the ankle or the wrist. Some methods, such as **MLR-HMM**, do better with the ankle data and others, such as **RF-HMM**, do better with wrist data. However, in the cases where the activity type is unknown or we use the two stage estimation strategy and impute the type, the point estimates indicate that for most classification methods a small improvement in the proportion correct can be achieved by using data from the

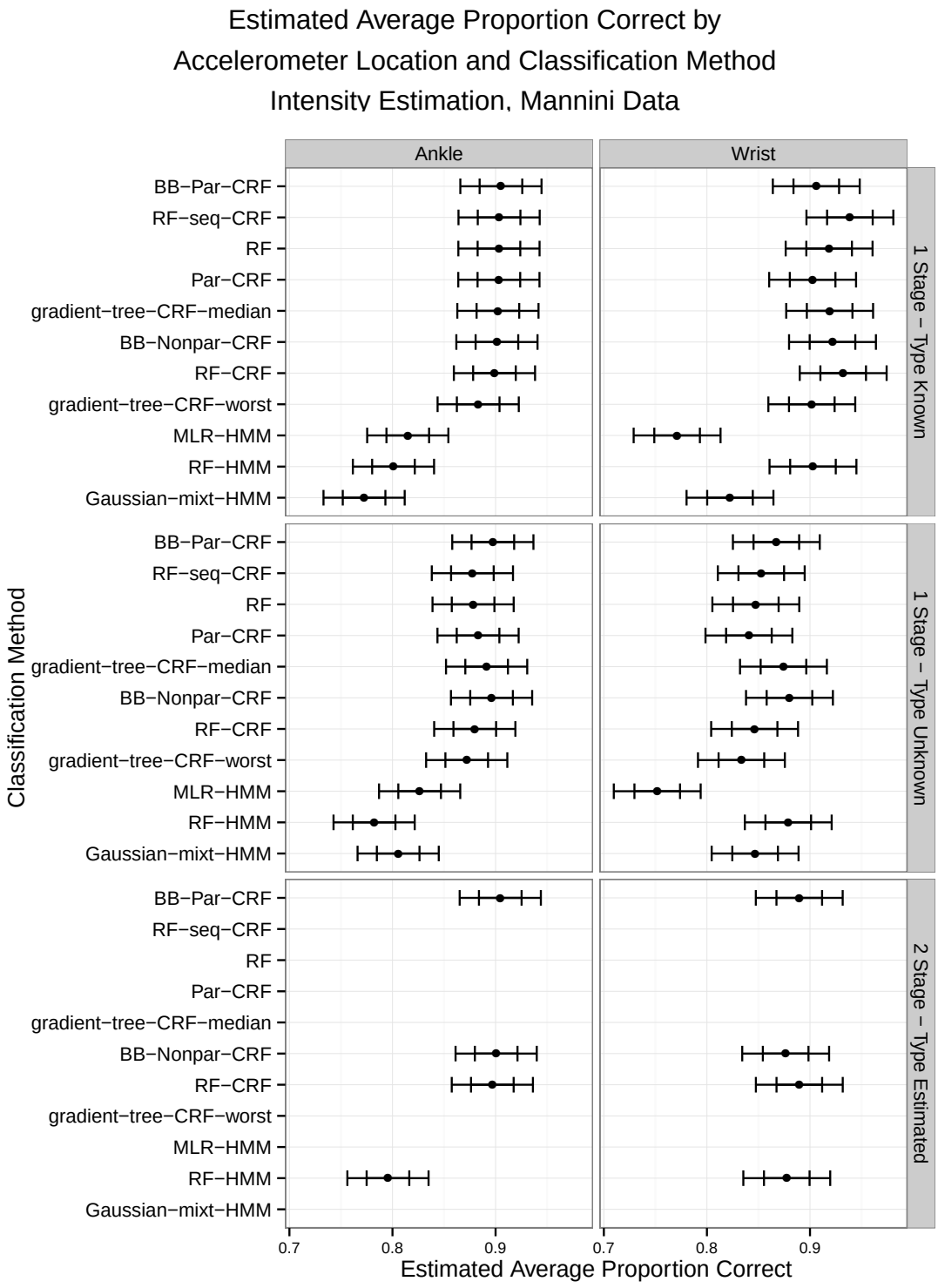


Figure 74. Point and interval estimates for the fixed effects parameters in model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 74, 75, 76, 77, 78, and 79.**

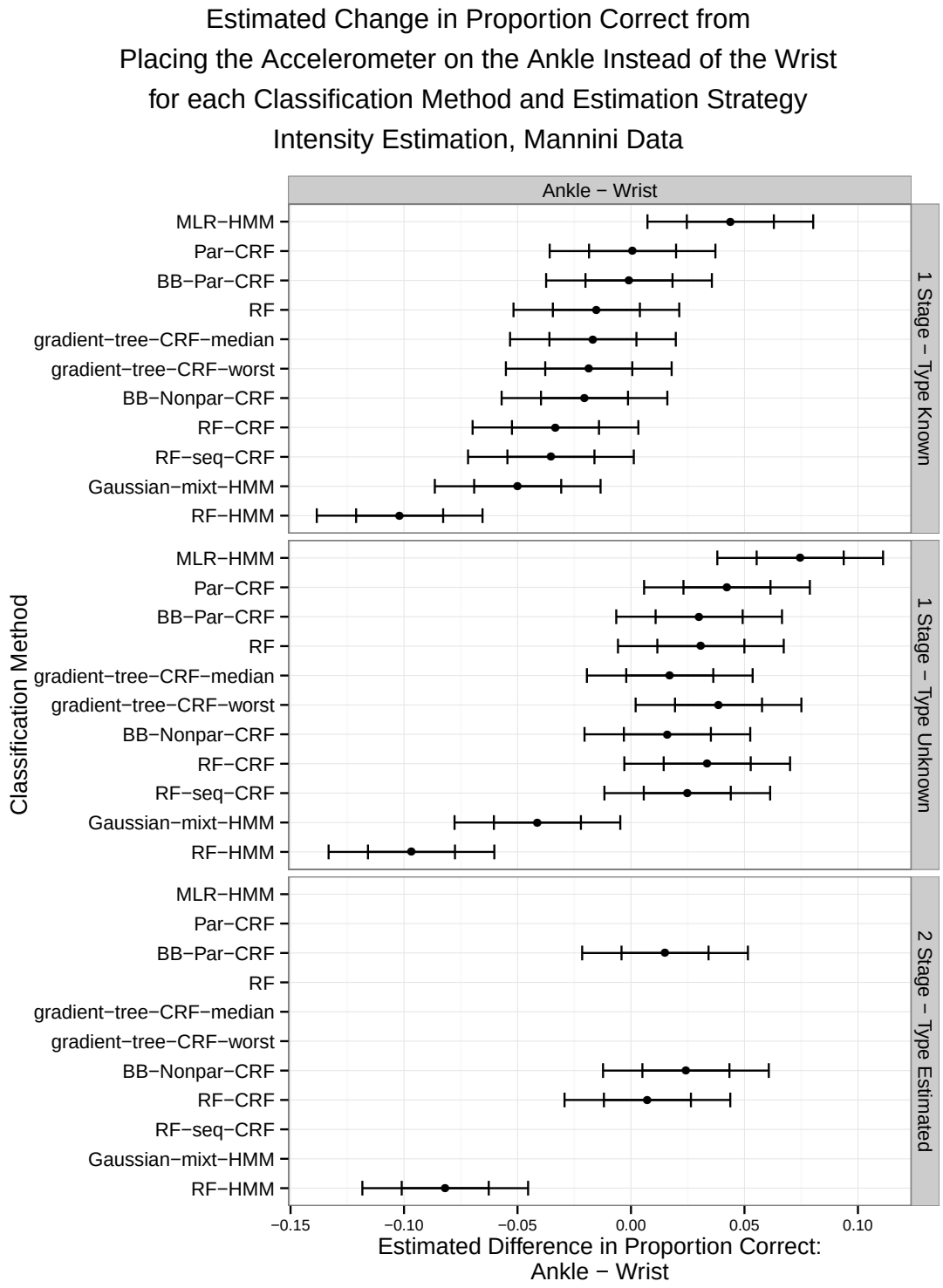


Figure 75. Point and interval estimates for the difference in performance between the ankle and wrist accelerometer locations for each classification method, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 74, 75, 76, 77, 78, and 79.**

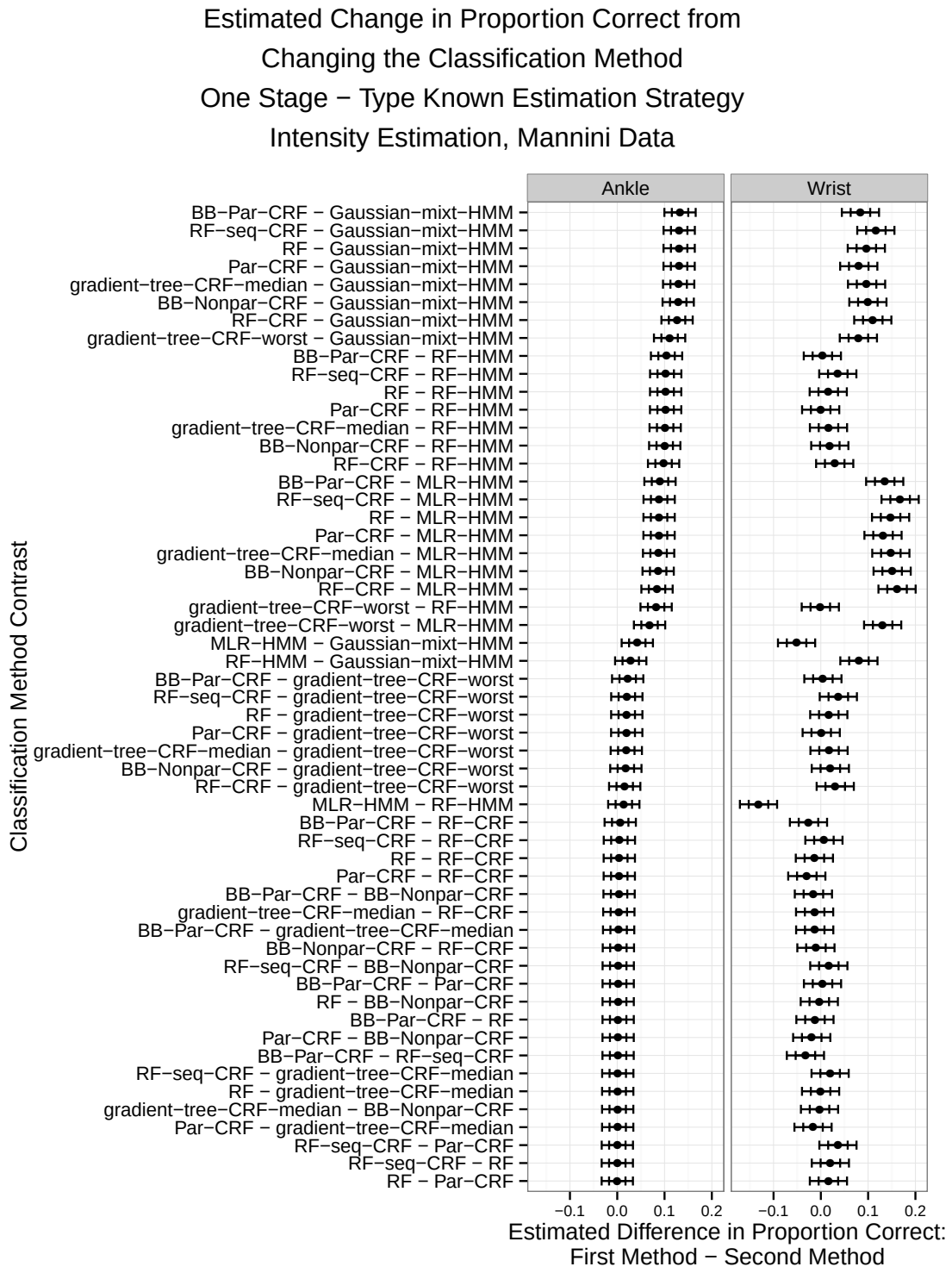


Figure 76. Point and interval estimates for the difference in performance between each pair of classification methods using a one stage estimation strategy where the true activity type in each window is known, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 74, 75, 76, 77, 78, and 79.**

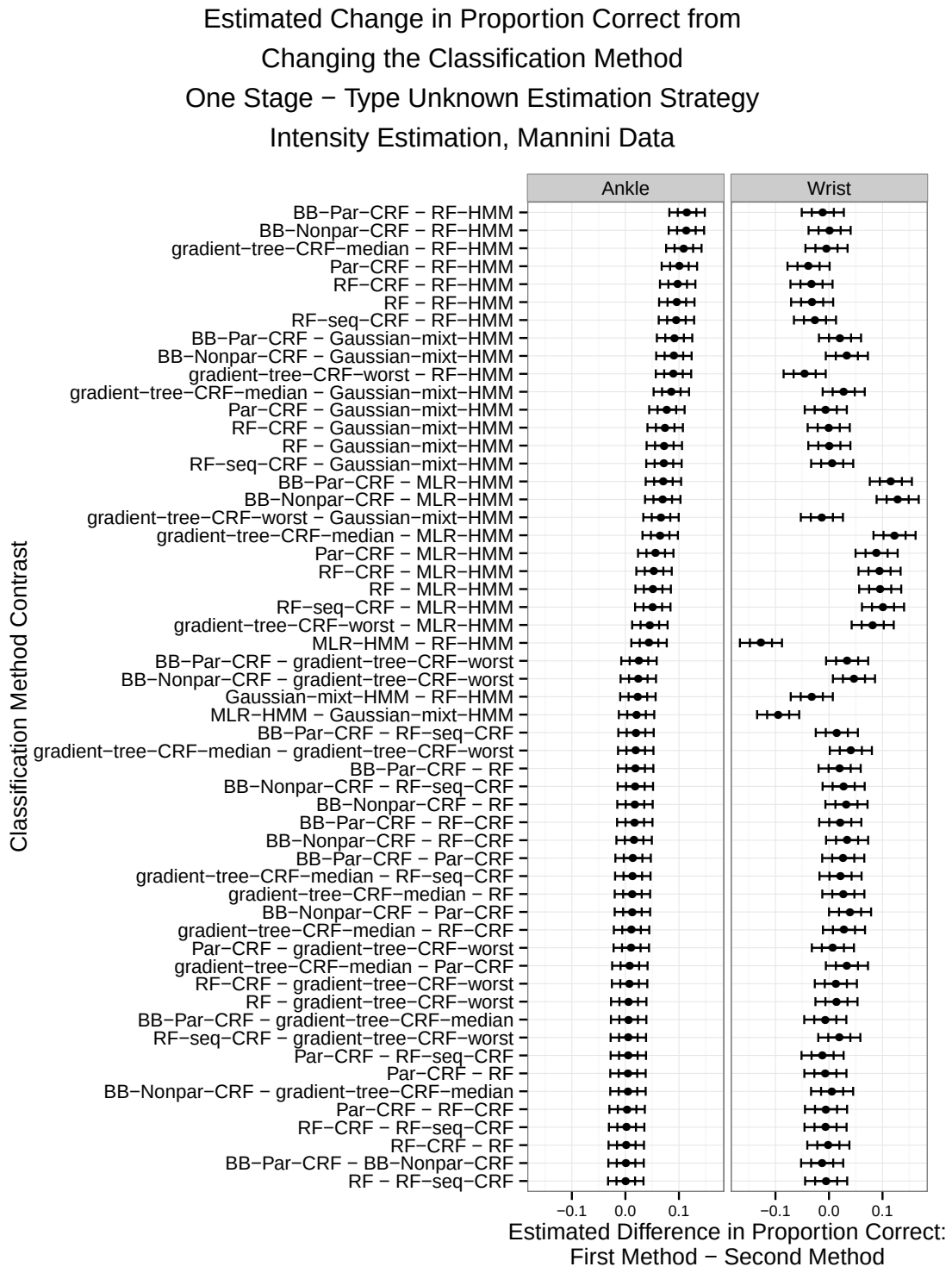


Figure 77. Point and interval estimates for the difference in performance between each pair of classification methods using a one stage estimation strategy where the true activity type in each window is unknown, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 74, 75, 76, 77, 78, and 79.**

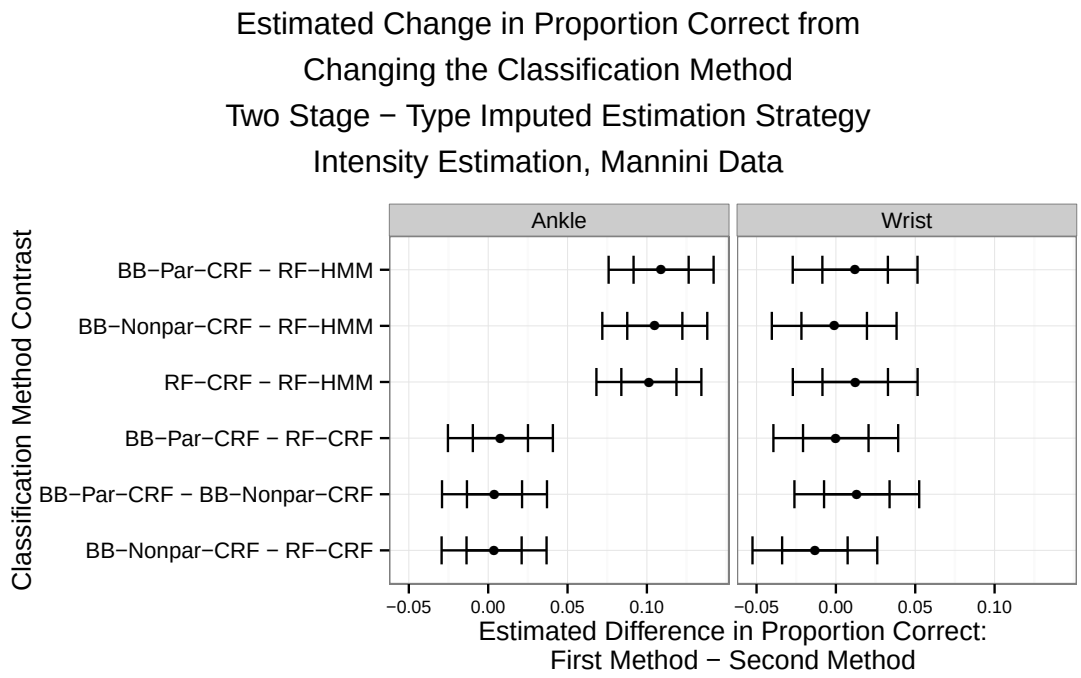


Figure 78. Point and interval estimates for the difference in performance between each pair of classification methods using a two stage estimation strategy where the activity type in each window is estimated in the first stage and that estimate is used as an input to intensity estimation in the second stage, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 74, 75, 76, 77, 78, and 79.**

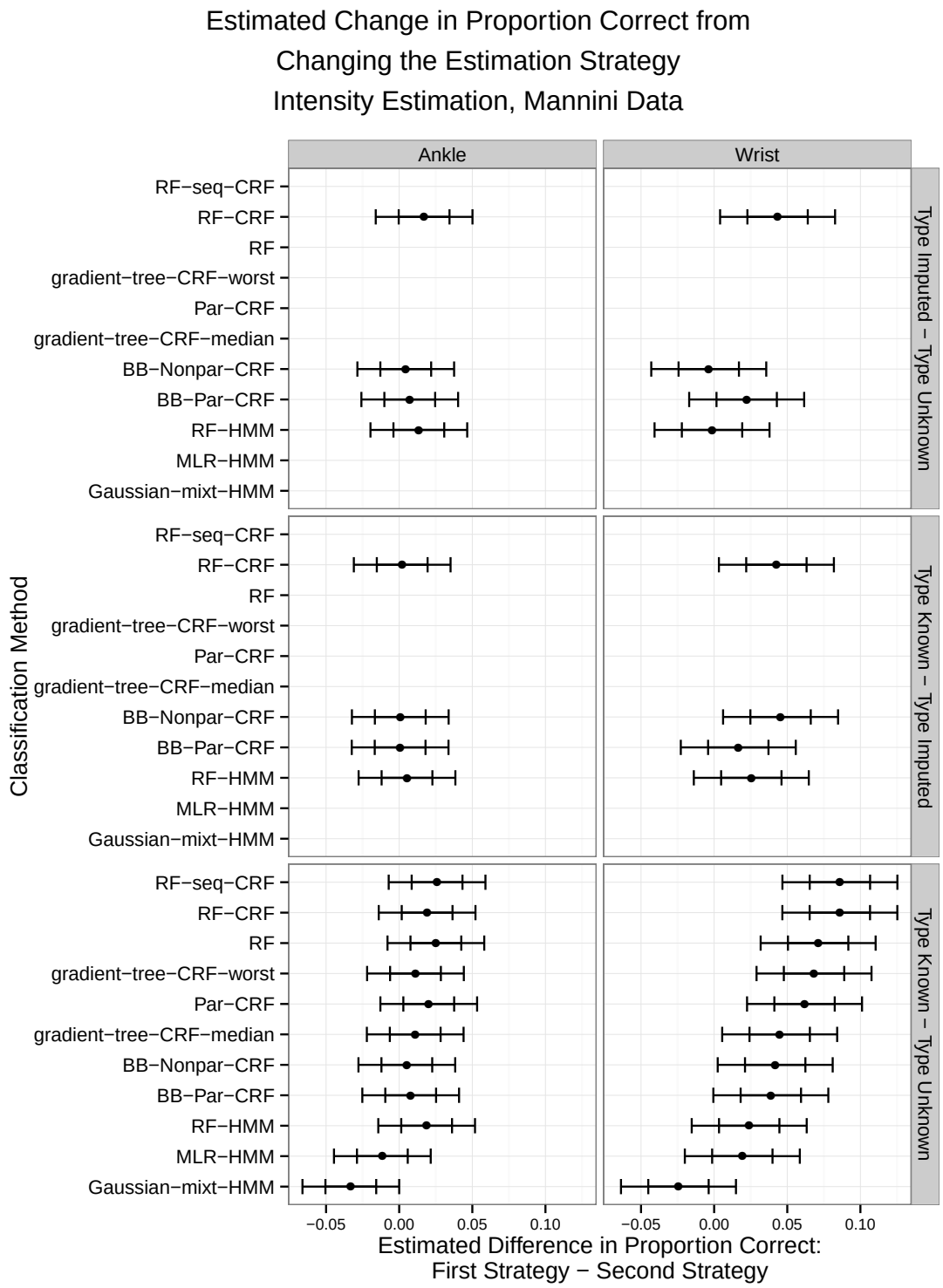


Figure 79. Point and interval estimates for the difference in performance between each pair of estimation strategies, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 74, 75, 76, 77, 78, and 79.**

ankle instead of the wrist. This improvement ranges from about 0.01 to 0.04, depending on the classification method. This is equivalent to an improvement of between about 10 and 40 minutes if we extrapolate to a 16 hour day while maintaining the same composition of the relative amount of time spent in each intensity category.

The estimates in Figures 76 and 77 confirm that when applied to the ankle data with a one stage estimation strategy, the **Gaussian-mixt-HMM**, **RF-HMM**, and **MLR-HMM** methods had the lowest proportion correct. Using data from the wrist, the **MLR-HMM** and **Gaussian-mixt-HMM** methods did the worst. The improvement in the proportion correct that could be realized by switching away from one of these methods to one of the other methods ranged from about 0.08 to about 0.17, or between about one hour and 15 minutes and 2 hours and 45 minutes in a 16 hour day. We remind the reader that these results are for the overall proportion of time points classified correctly, and the story would be somewhat different if we had examined the macro F_1 score instead.

The estimates in Figure 78 show that with the two stage estimation strategy and data from the ankle, the **RF-HMM** had an average proportion correct that was lower than the other methods by about 0.1. With the wrist data, the two-stage methods are essentially indistinguishable from each other.

The estimates in Figure 79 demonstrate that with the ankle data, the gains from including information about activity type in estimating intensity that we observed in the boxplots above are for the most part too small to detect with our linear model. All of the confidence intervals overlap 0 except for the one for the **Gaussian-mixt-HMM**, and that method did worse when it used information about activity type. The wrist data make a slightly stronger case for using information about activity type. When the true type is known, several methods showed gains in the proportion correct when they used that information. However, only the **RF-CRF** method showed an appreciable gain in the proportion correct when the class memberships were unobserved and had to be imputed.

8.4 Sasaki Laboratory Data

Figures 80, 81, and 82 display box plots summarizing the results of the application to intensity classification with the laboratory data from Sasaki [2013]. In this data set, the difference in

Labeled Class	Predicted Class			
	Sedentary	Light	Moderate	Transition
Sedentary	959	45	23	2
Light	55	2007	541	12
Moderate	0	376	1549	3
Transition	2	31	24	44

Table 26. Confusion matrix for intensity classification with the RF-HMM method and one stage estimation strategy with physical activity type unknown, applied to the lab wrist data from Sasaki [2013], all subjects combined.

proportion correct and mean squared error achieved by the first stage classifiers with and without use of the activity type labels is much less apparent than it was in the data from Mannini et al. [2013]. Overall, it does seem that using the observed activity type labels may improve estimates of intensity if that information is available, but the improvement is small. The differences in F_1 score are slightly larger however, and show that using the activity type in classifying activity intensity is helpful. Similarly, the two stage procedure does not seem to offer much of an advantage over the one-stage procedure that doesn't use any information about the activity type when we consider the proportion correct and mean squared error, but it does yield a small improvement in the F_1 score. The lack of a major difference in the estimation strategies is confirmed in Figure 83, where we see that for many subjects, all three estimation strategies resulted in exactly the same proportion correct.

The confusion matrices in Tables 26 through 31 yield similar conclusions. Including activity type information leads to only minor changes in the confusion matrices, suggesting that the activity type is not contributing much to the class estimates. The activity type adds very little information about intensity that was not already available in the accelerometer signal.

It appears that all of the classification methods achieve about the same proportion correct, with the exception of the **Gaussian-mixt-HMM** method with the one-stage estimation strategy using the true activity type and the **Par-CRF** method with data from the ankle or wrist. There is a little more variation in the F_1 scores and MSE. According to the F_1 score, the **BB-Par-CRF**, **RF-CRF**, **RF-seq-CRF**, **RF-HMM**, and **MLR-HMM** seem to offer better performance than the other methods. As we saw in the applications to activity type prediction, although the **RF** had a lower F_1 score than the other methods, its MSE is in line with the other methods and is less variable.

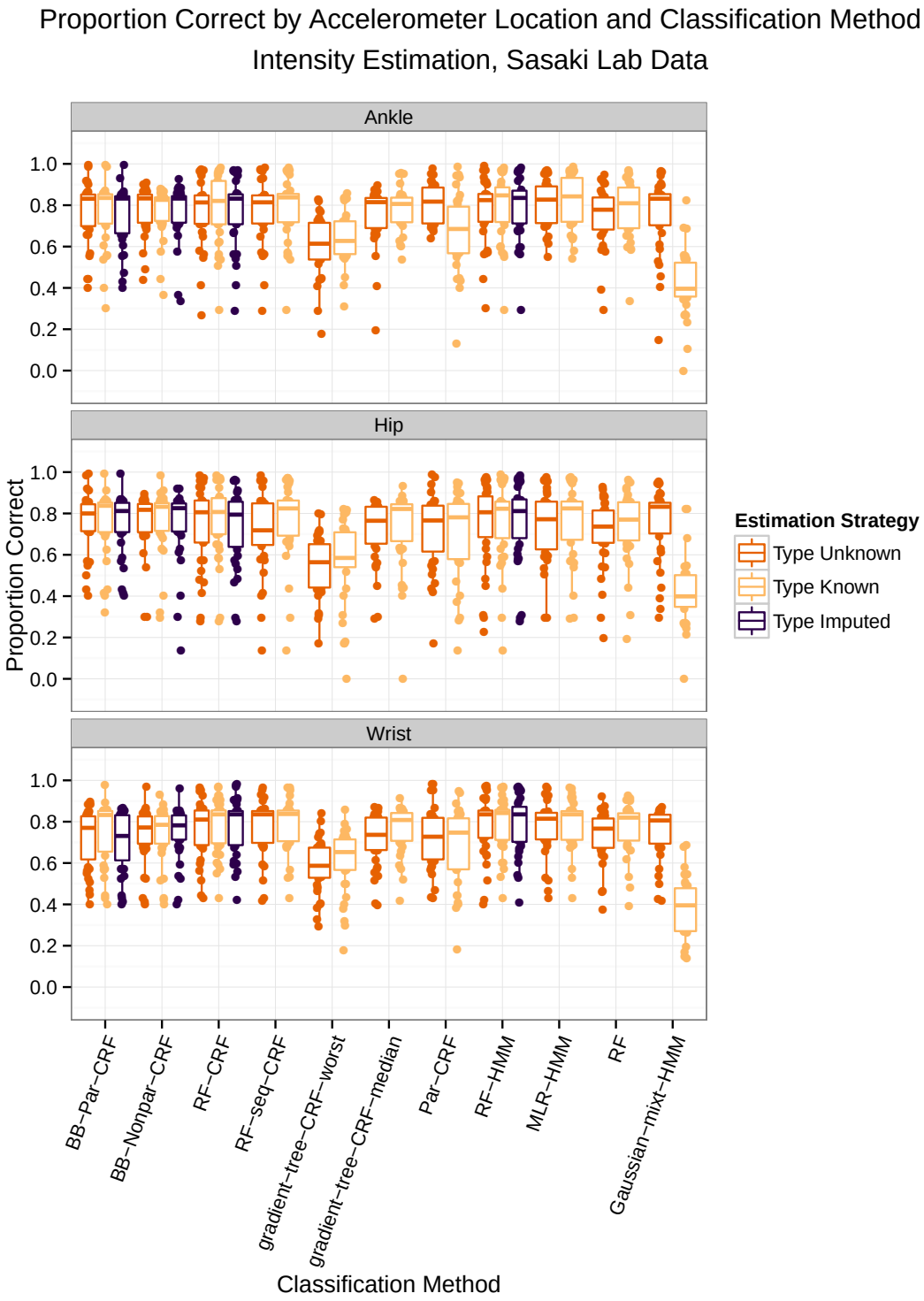


Figure 80. Box plots showing the proportion of windows classified correctly in the data from Sasaki [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

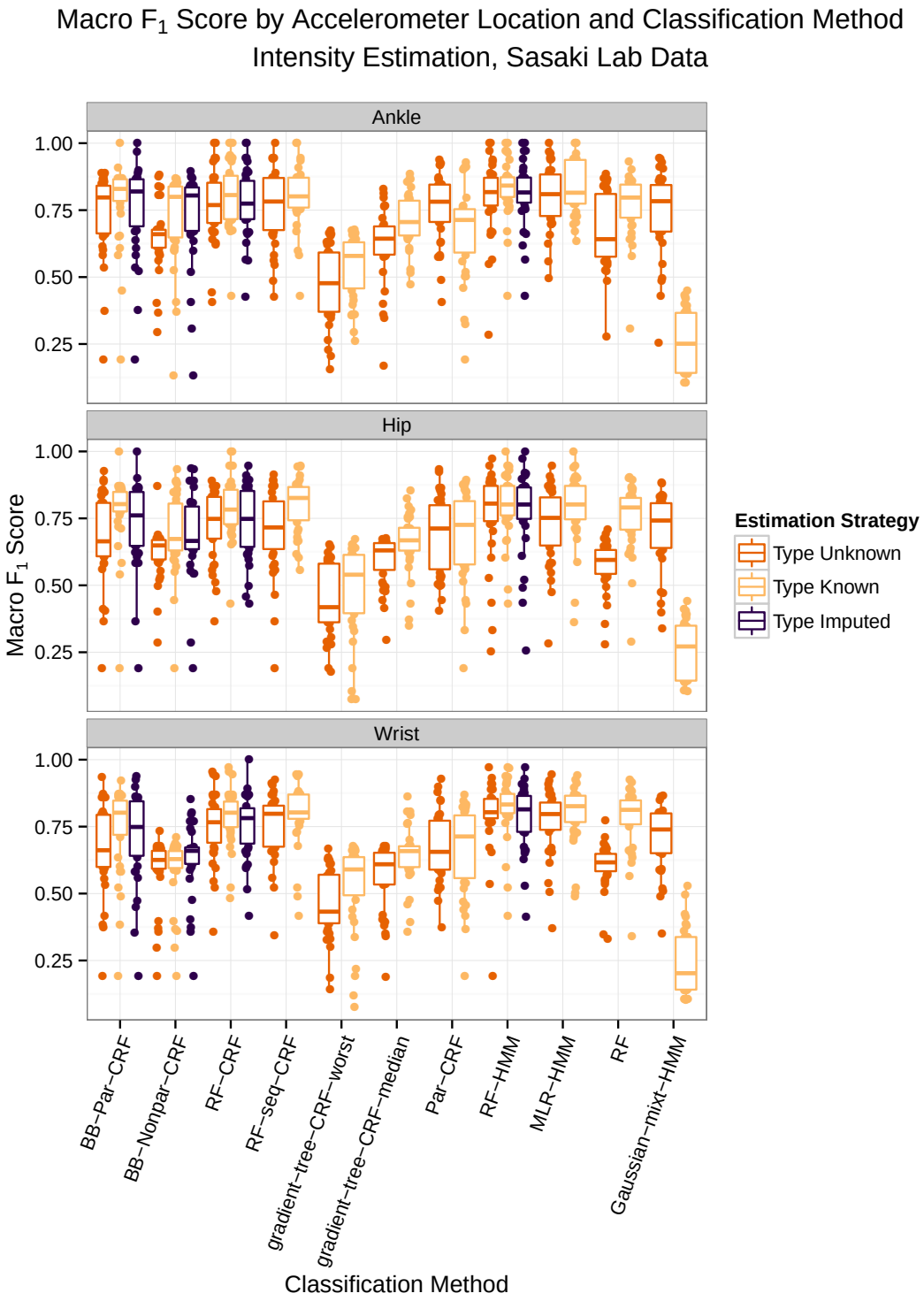


Figure 81. Box plots showing the macro F_1 score combining precision and recall across all physical activity intensity categories in the data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

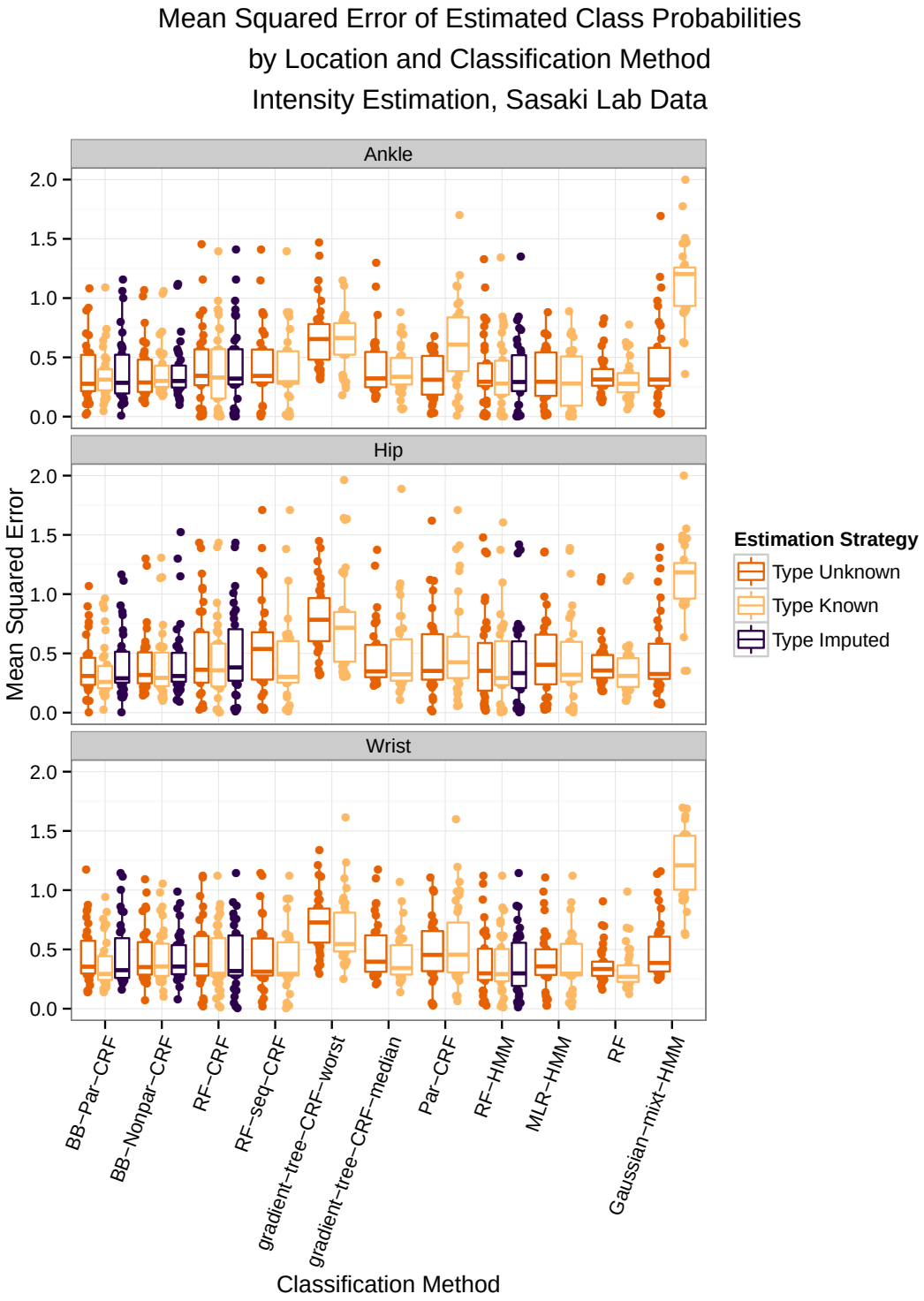


Figure 82. Box plots showing the mean squared error of the estimated class membership probabilities relative to the labeled class memberships in the data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

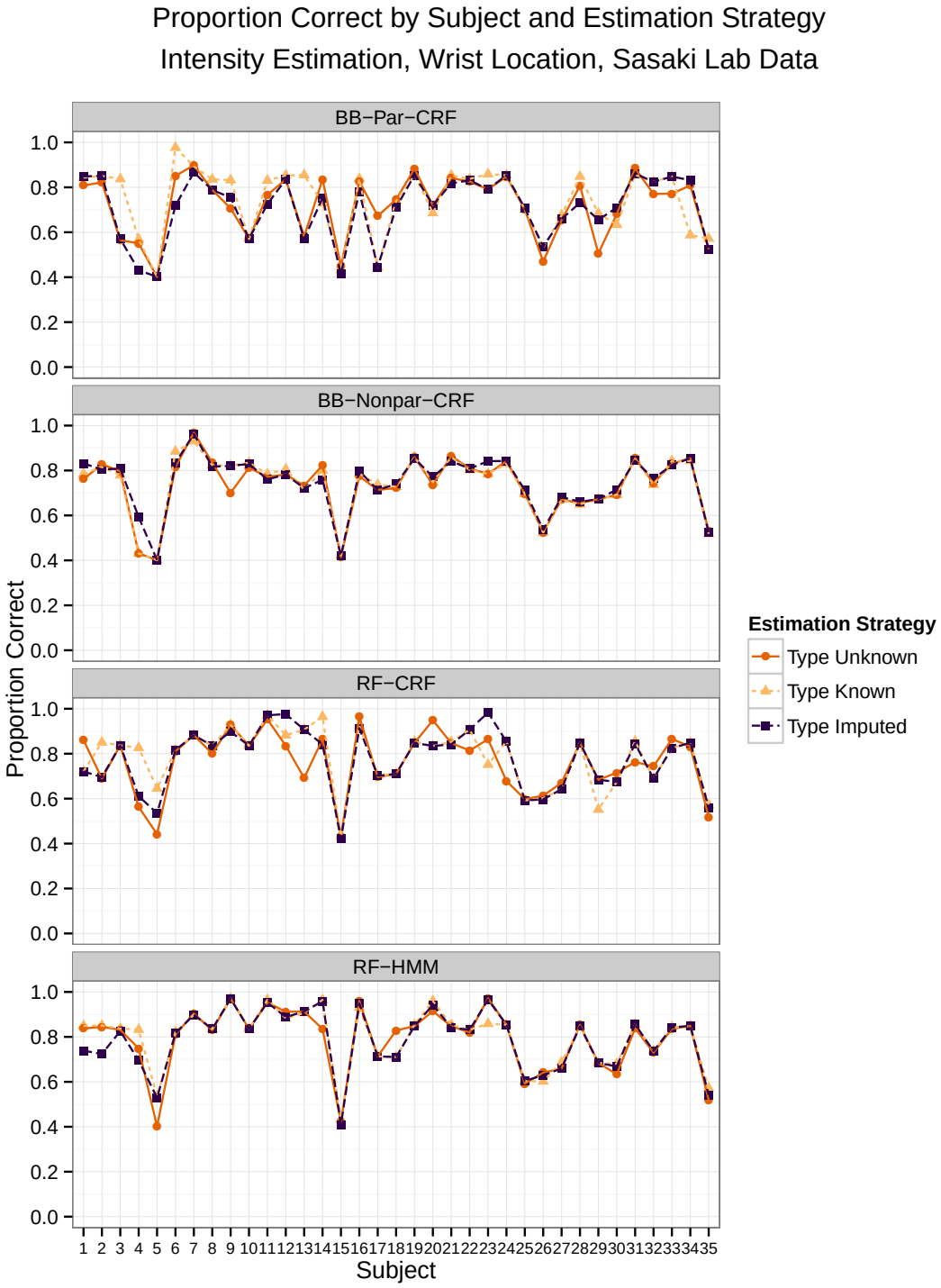


Figure 83. Proportion of time windows with intensity level classified correctly by subject in the data from Sasaki [2013], using a leave-one-subject-out procedure. We show these values for just the BB-Par-CRF, BB-Nonpar-CRF, RF-CRF, and RF-HMM classification methods applied to data collected from accelerometers placed at the wrist location.

Labeled Class	Predicted Class			
	Sedentary	Light	Moderate	Transition
Sedentary	960	45	23	1
Light	23	2033	546	13
Moderate	0	359	1564	5
Transition	0	29	16	56

Table 27. Confusion matrix for intensity classification with the RF-HMM method and one stage estimation strategy with physical activity type known, applied to the lab wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class			
	Sedentary	Light	Moderate	Transition
Sedentary	960	45	23	1
Light	46	1991	563	15
Moderate	0	374	1547	7
Transition	5	32	23	41

Table 28. Confusion matrix for intensity classification with the RF-HMM method and two stage estimation strategy with physical activity type imputed in the first stage, applied to the lab wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class			
	Sedentary	Light	Moderate	Transition
Sedentary	954	51	23	1
Light	72	1764	772	7
Moderate	23	529	1376	0
Transition	13	34	37	17

Table 29. Confusion matrix for intensity classification with the BB-Par-CRF method and one stage estimation strategy with physical activity type unknown, applied to the lab wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class			
	Sedentary	Light	Moderate	Transition
Sedentary	960	45	23	1
Light	23	1924	656	12
Moderate	0	530	1390	8
Transition	0	33	21	47

Table 30. Confusion matrix for intensity classification with the BB-Par-CRF method and one stage estimation strategy with physical activity type known, applied to the lab wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class			
	Sedentary	Light	Moderate	Transition
Sedentary	936	68	23	2
Light	68	1761	786	0
Moderate	23	576	1329	0
Transition	4	35	35	27

Table 31. Confusion matrix for intensity classification with the BB-Par-CRF method and two stage estimation strategy with physical activity type imputed in the first stage, applied to the lab wrist data from Sasaki [2013], all subjects combined.

We fit the same linear mixed effects model to the classification results that we used in Section 8.3 for the results of intensity classification with the data from Mannini et al. [2013]. We display point and interval estimates for the average proportion correct for each combination of classifier, accelerometer location, and number of classes, along with sets of contrasts between these quantities in Figures 84 through 89. These estimates confirm the observations we made from the box plots above. The estimates in Figure 85 show that in this data set, there is not an appreciable benefit from changing the accelerometer location. Figures 86, 87, and 88 show that most classification methods perform about as well as each other, except for the **Gaussian-mixt-HMM** method and the **Par-CRF** method for some combinations of accelerometer location and the estimation strategy. Figure 89 shows that if the true class were available it could provide information about physical activity intensity that is not in the accelerometer data from the wrist, but the activity type does not contain information about intensity that is not already available in the accelerometer signal if we use data from the ankle or the hip. In any case, when the true activity type is unknown, using an imputed activity type in a two-stage procedure does not lead to appreciable gains in the proportion correct.

8.5 Sasaki Free Living Data

Figures 90, 91, and 92 display box plots summarizing the results of the application to intensity classification with the free living data from Sasaki [2013]. As we saw with classification of activity type, the overall classification rates are lower with data from accelerometers placed at the wrist than they are with accelerometers at the ankle or hip. Also, the classification rates are lower with the free living data than they were with the lab data sets. In these plots we can see that if it is available, using the true activity type makes an appreciable improvement in classification of intensity with every classification method other than the **Gaussian-mixt-HMM**. However, the two stage methods that use imputed class memberships do about as well as the one stage methods that do not use the activity type. Figure 93 breaks down performance by subject, and shows in more detail that while using the true activity type is helpful if it is available, imputing the type does not lead to gains in performance. In terms of comparisons among the different classification methods, it appears that the **BB-Par-CRF** and **BB-Nonpar-CRF** methods offer the most consistently high performance, and the **Gaussian-mixt-HMM** method is consistently among the

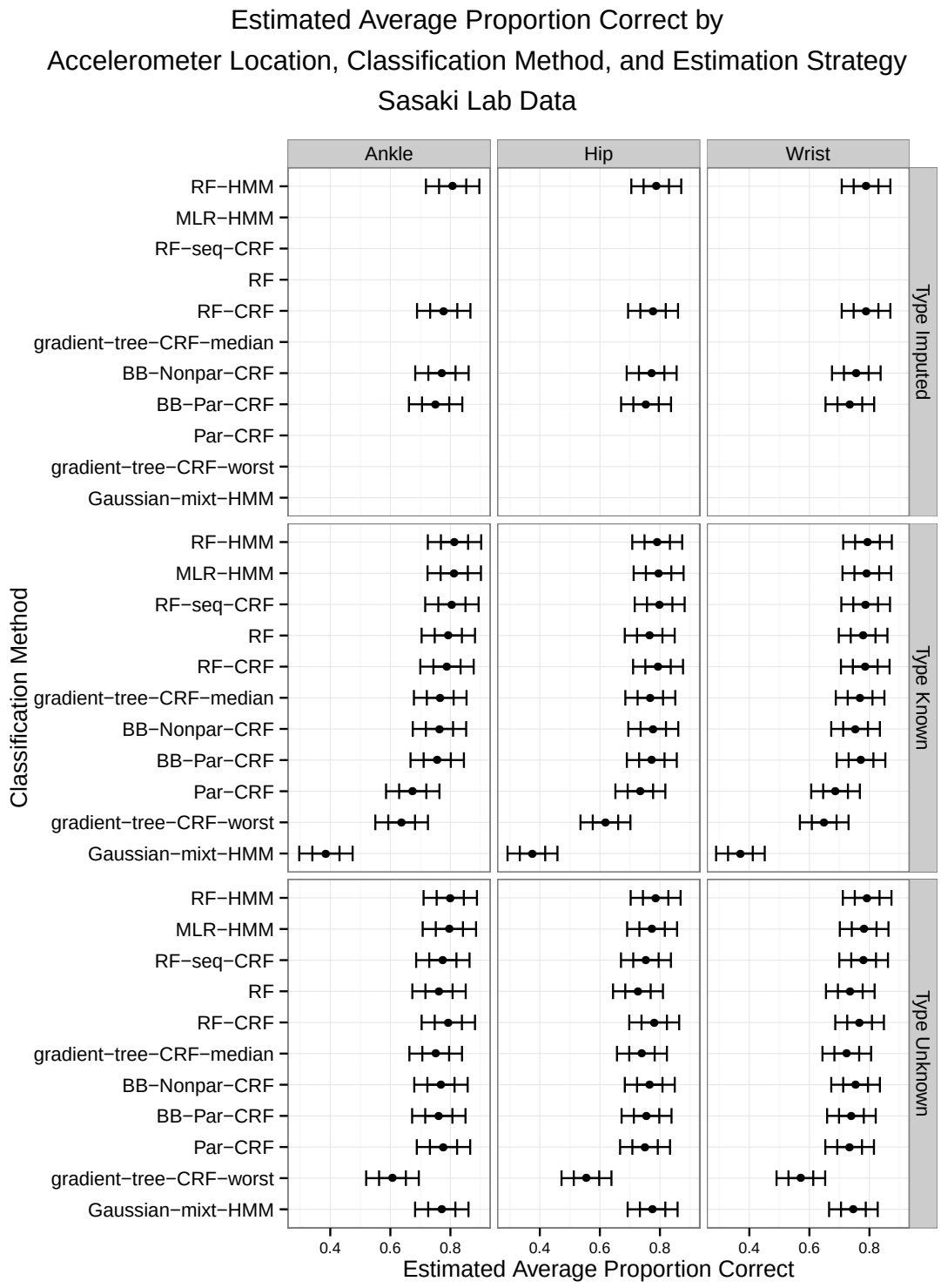


Figure 84. Point and interval estimates for the fixed effects parameters in model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 84, 85, 86, 87, 88, and 89.**

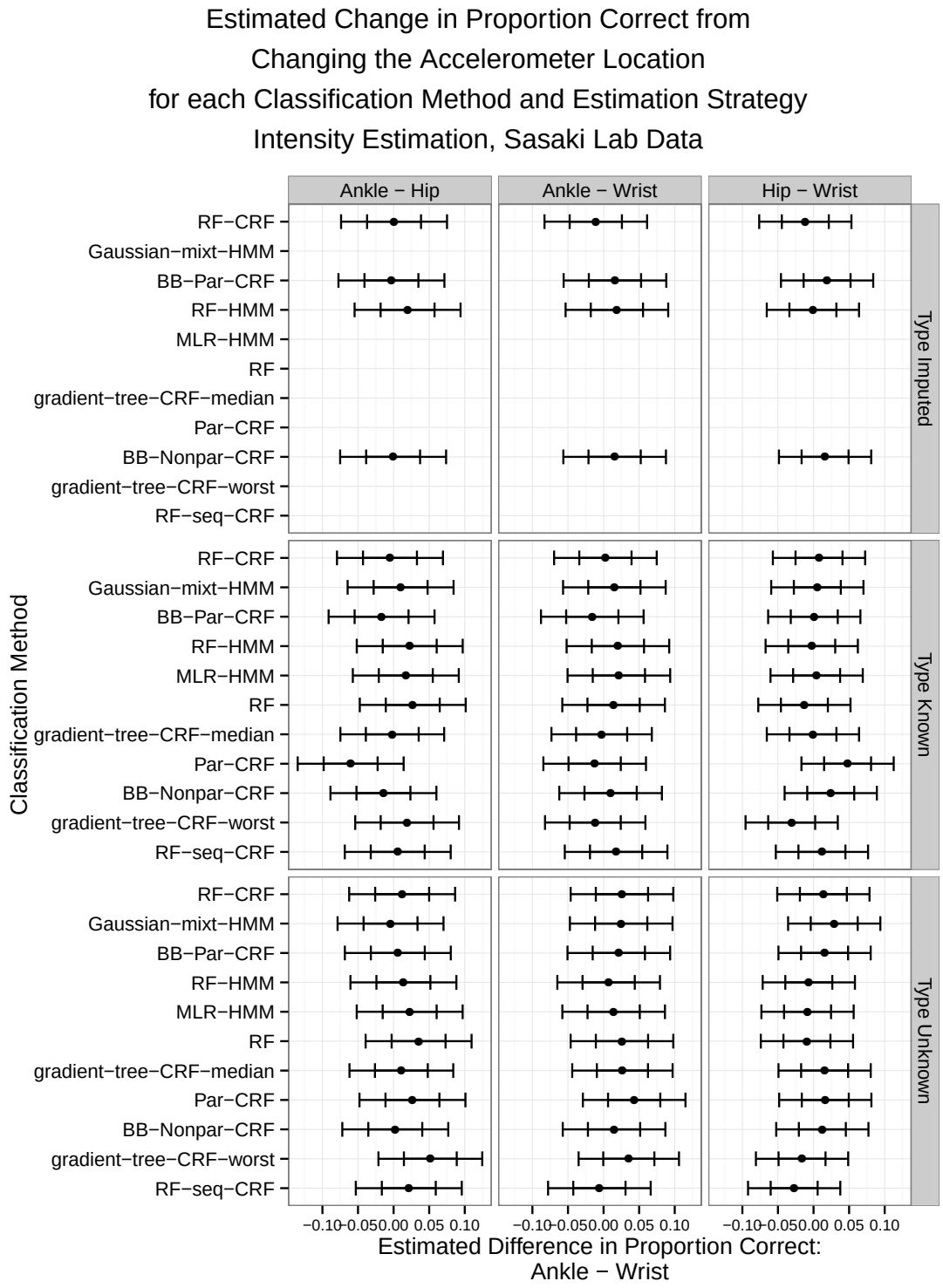


Figure 85. Point and interval estimates for the difference in performance between the ankle and wrist accelerometer locations for each classification method, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 84, 85, 86, 87, 88, and 89.**

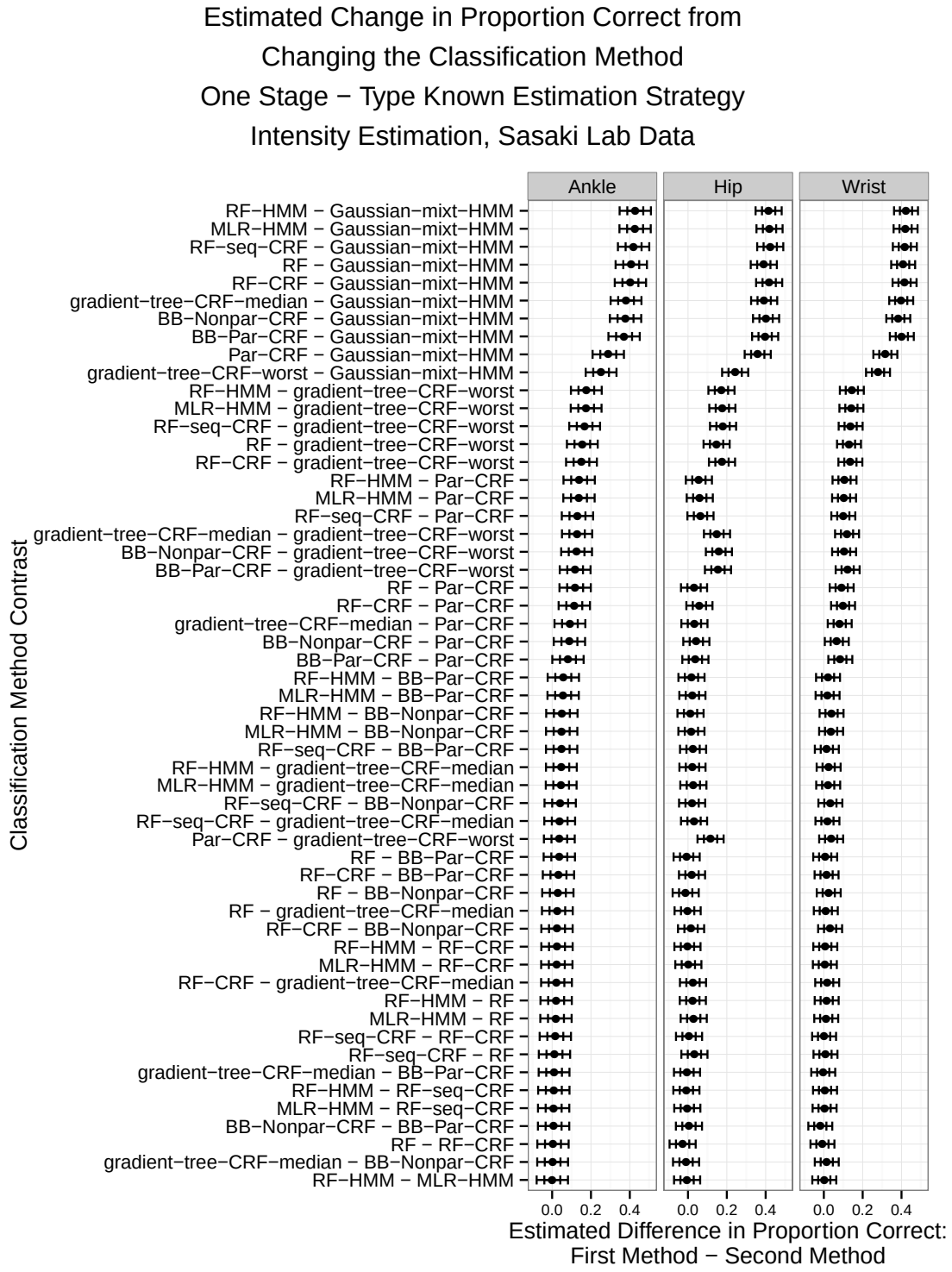


Figure 86. Point and interval estimates for the difference in performance between each pair of classification methods using a one stage estimation strategy where the true activity type in each window is known, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 84, 85, 86, 87, 88, and 89.**

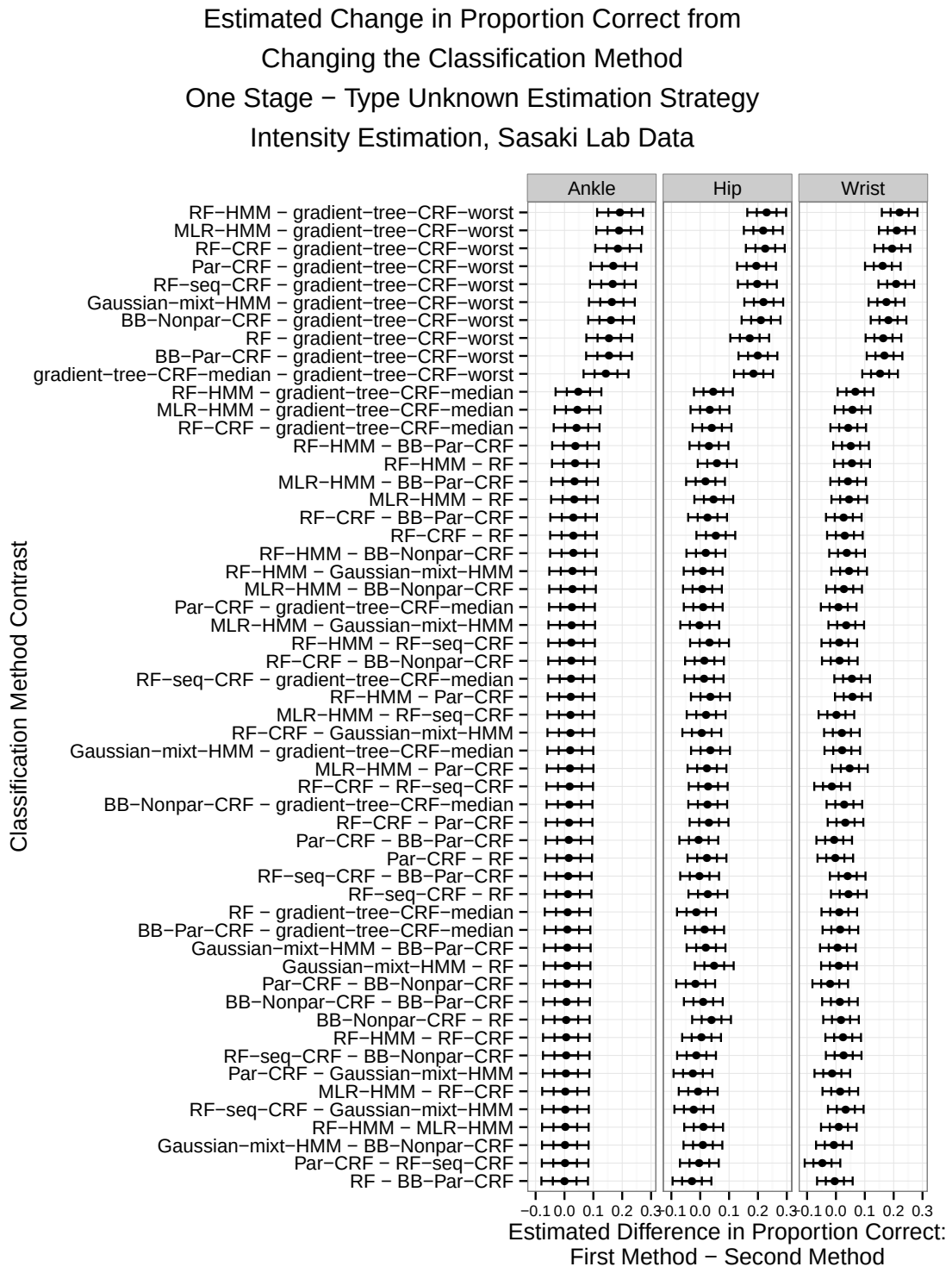


Figure 87. Point and interval estimates for the difference in performance between each pair of classification methods using a one stage estimation strategy where the true activity type in each window is unknown, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 84, 85, 86, 87, 88, and 89.**

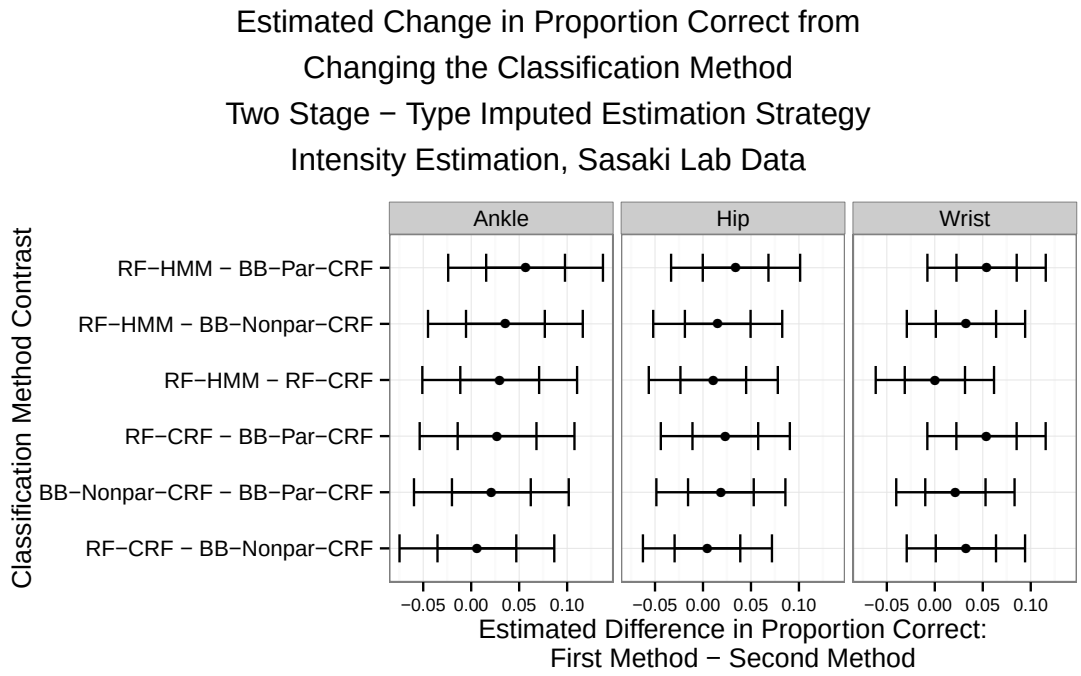


Figure 88. Point and interval estimates for the difference in performance between each pair of classification methods using a two stage estimation strategy where the activity type in each window is estimated in the first stage and that estimate is used as an input to intensity estimation in the second stage, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 84, 85, 86, 87, 88, and 89.**

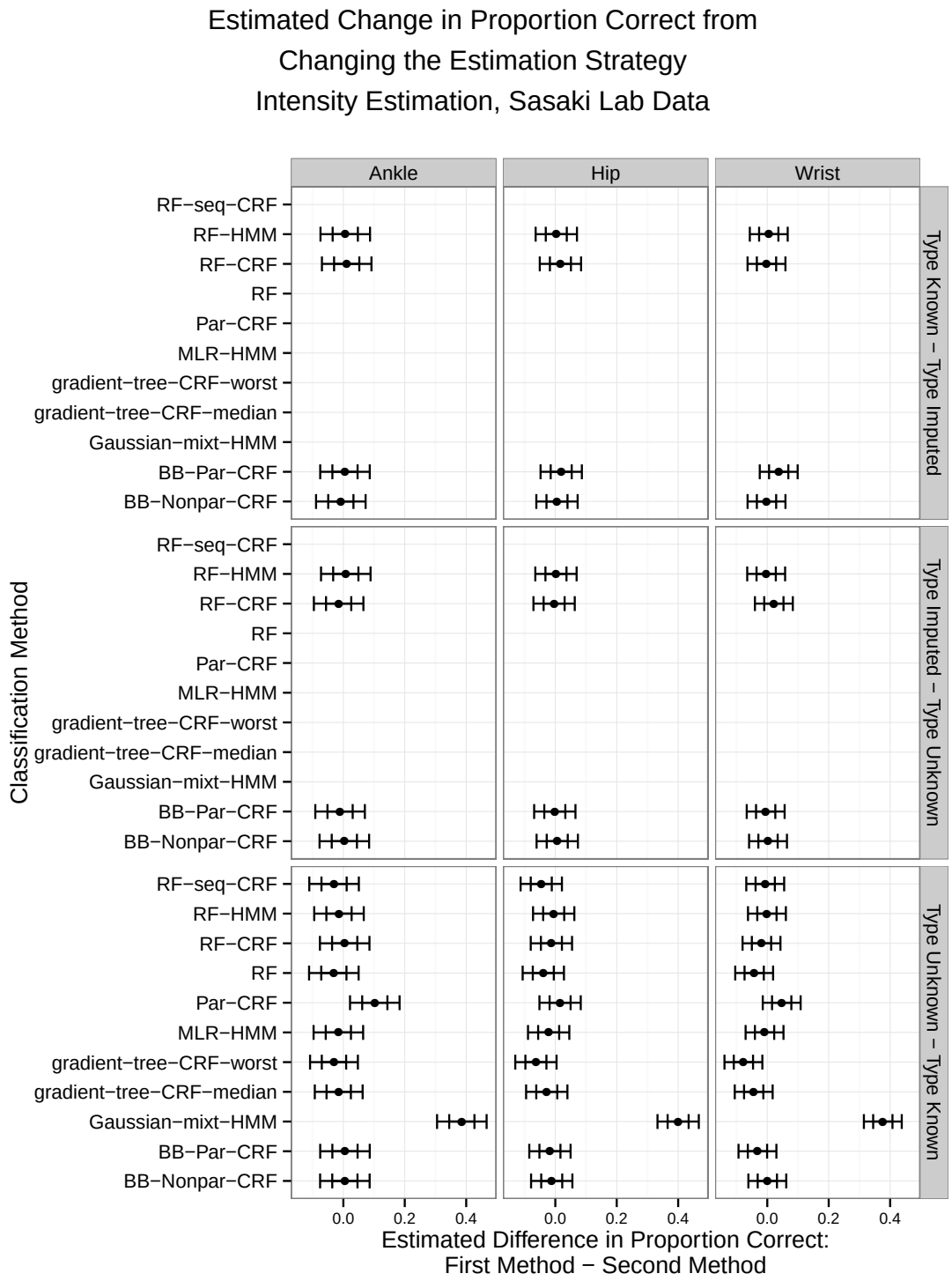


Figure 89. Point and interval estimates for the difference in performance between each pair of estimation strategies, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 84, 85, 86, 87, 88, and 89.**

Labeled Class	Predicted Class				
	Sedentary	Light	Moderate	Vigorous	Transition
Sedentary	1585	451	2	1	16
Light	603	2540	65	68	169
Moderate	49	318	1007	511	73
Vigorous	0	5	51	10	1
Transition	59	467	88	45	125

Table 32. Confusion matrix for intensity classification with the RF-HMM method and one stage estimation strategy with physical activity type unknown, applied to the free living wrist data from Sasaki [2013], all subjects combined.

worst-performing methods. However, most of the methods perform about as well as each other. These results are the same with all three summary statistics (proportion correct, F_1 score, and MSE).

The confusion matrices in Tables 32 through 37 reveal interesting behavior of the **RF-HMM** method. The free living component of the study included only 67 windows of vigorous intensity activity. As before, the **RF-HMM** method makes a trade-off in favor of improved recall rates in classes representing a lower proportion of the data overall, at the expense of lower precision. As a result, the **RF-HMM** method misclassifies a large number of points (including between 15 and 42% of windows where the true activity intensity is Moderate) as Vigorous activity. Using the true activity helps this situation somewhat. However, the estimation strategy using imputed values for physical activity type does much worse in terms of misclassifying Moderate activity as Vigorous activity than the strategy that ignores the activity type altogether. Although this issue with Vigorous activity does not arise with the **BB-Par-CRF**, we can see a similar pattern in terms of an overall increase in misclassification rates when we use the imputed activity types. Of the five on-diagonal entries in the confusion matrix, four decrease and the fifth remains fixed at 0 when we move from a one-stage model that ignores activity type to a two-stage model that imputes activity type. Thus, the overall proportion correct aggregated across all subjects decreased or stayed fixed at 0 for every intensity category when we moved from a one-stage process ignoring activity type to a two-stage process using imputed activity types.

We fit the same linear mixed effects model to the classification results that we used in Sections 8.3 and 8.4. We display point and interval estimates for the average proportion correct for

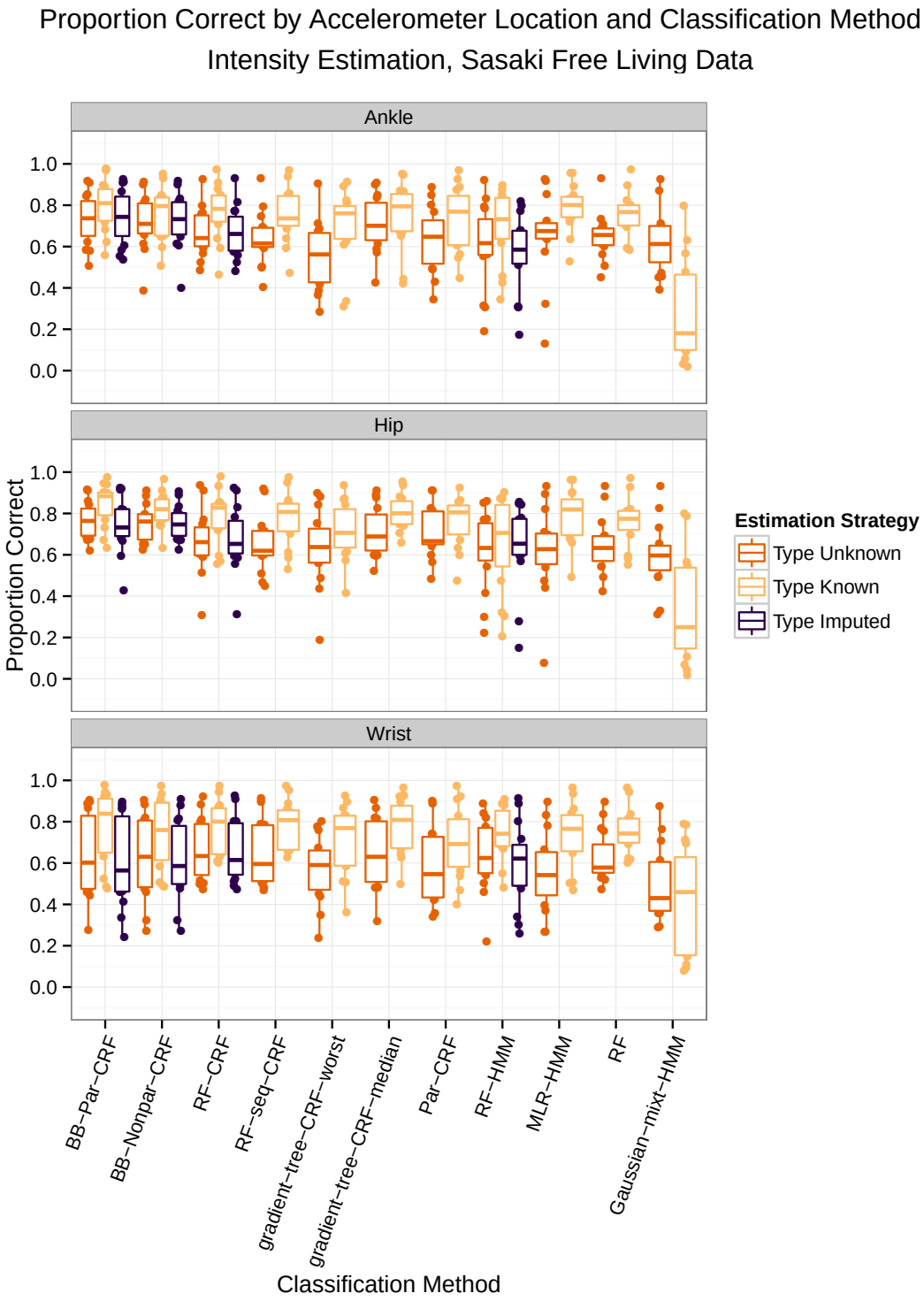


Figure 90. Box plots showing the proportion of windows classified correctly in the free living data from Sasaki [2013] using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

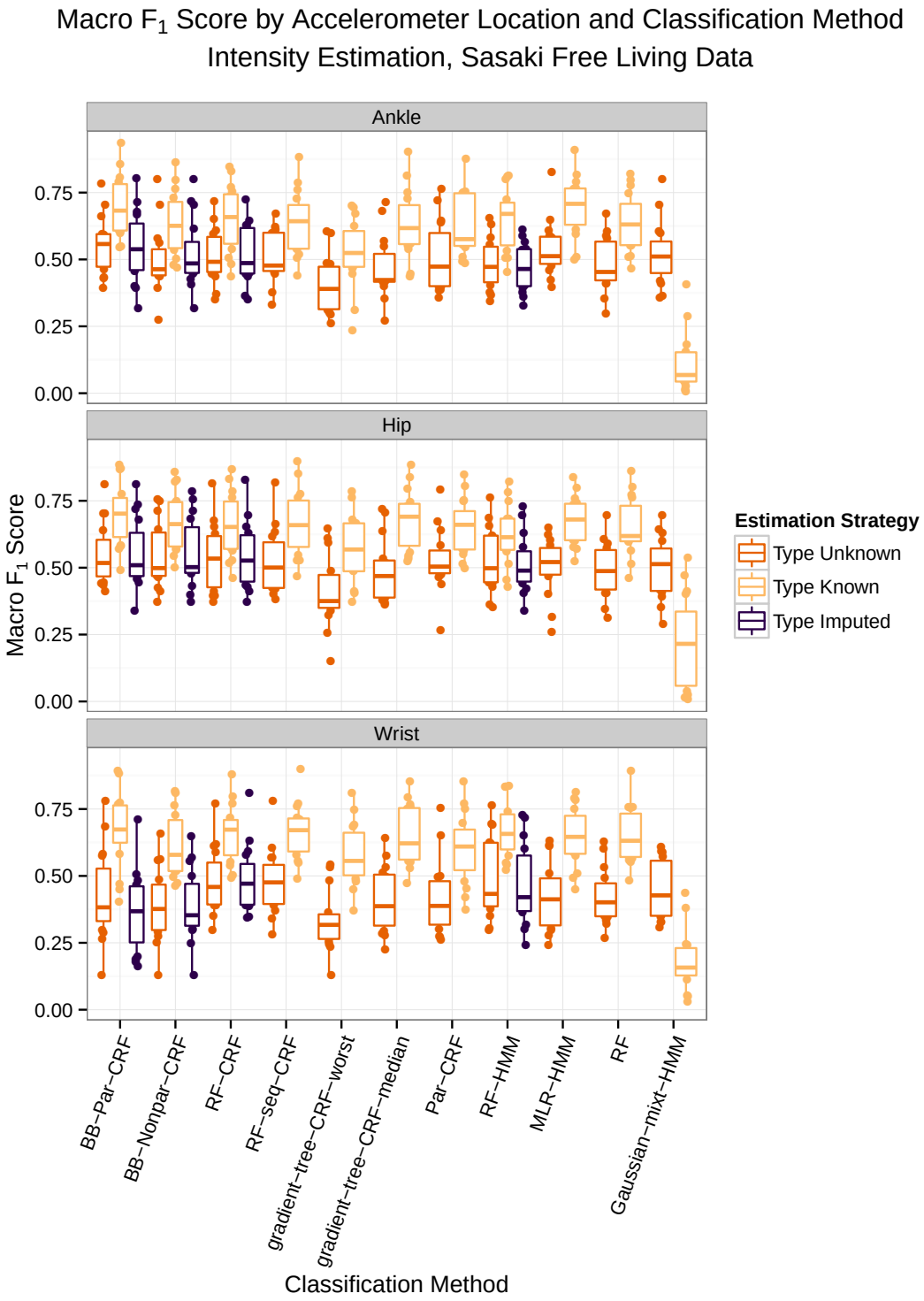


Figure 91. Box plots showing the macro F_1 score combining precision and recall across all physical activity intensity categories in the free living data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

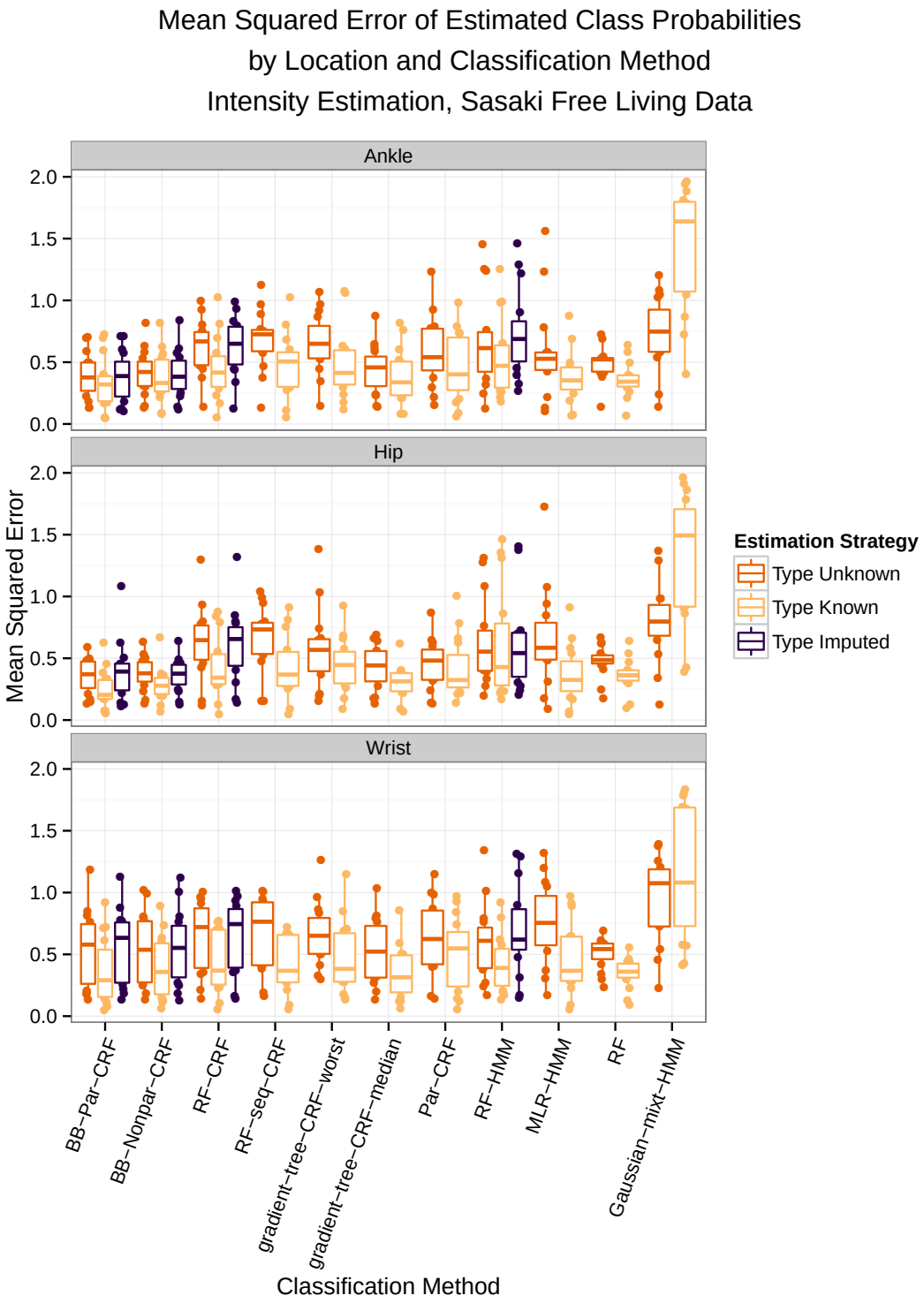


Figure 92. Box plots showing the mean squared error of the estimated class membership probabilities relative to the labeled class memberships in the free living data from Sasaki [2013], using a leave-one-subject-out procedure. A separate box plot is displayed for each combination of accelerometer location, classification method, and estimation strategy. Each point corresponds to a combination of accelerometer location, classification method, estimation strategy, and subject.

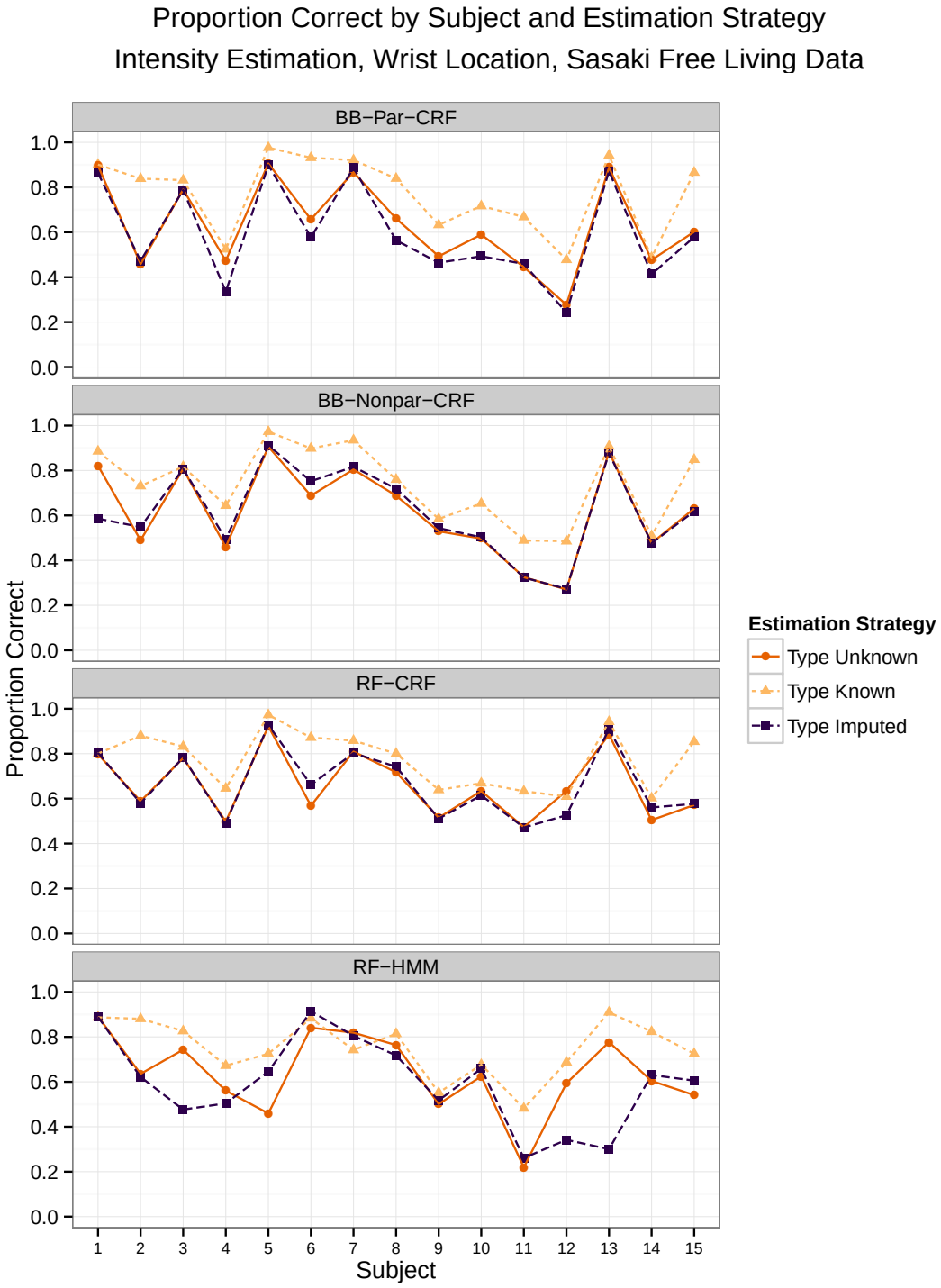


Figure 93. Proportion of time windows with intensity level classified correctly by subject in the free living data from Sasaki [2013], using a leave-one-subject-out procedure. We show these values for just the BB-Par-CRF, BB-Nonpar-CRF, RF-CRF, and RF-HMM classification methods applied to data collected from accelerometers placed at the wrist location.

Labeled Class	Predicted Class				
	Sedentary	Light	Moderate	Vigorous	Transition
Sedentary	1906	148	0	0	1
Light	696	2521	20	84	124
Moderate	3	250	1381	298	26
Vigorous	0	6	48	11	2
Transition	24	258	34	29	439

Table 33. Confusion matrix for intensity classification with the RF-HMM method and one stage estimation strategy with physical activity type known, applied to the free living wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class				
	Sedentary	Light	Moderate	Vigorous	Transition
Sedentary	1769	258	0	0	28
Light	928	2266	44	60	147
Moderate	48	337	680	825	68
Vigorous	0	5	48	13	1
Transition	70	471	71	66	106

Table 34. Confusion matrix for intensity classification with the RF-HMM method and two stage estimation strategy with physical activity type imputed in the first stage, applied to the free living wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class				
	Sedentary	Light	Moderate	Vigorous	Transition
Sedentary	1182	861	9	0	3
Light	693	2483	249	0	20
Moderate	26	437	1489	0	6
Vigorous	0	5	62	0	0
Transition	60	558	153	0	13

Table 35. Confusion matrix for intensity classification with the BB-Par-CRF method and one stage estimation strategy with physical activity type unknown, applied to the free living wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class				
	Sedentary	Light	Moderate	Vigorous	Transition
Sedentary	1532	523	0	0	0
Light	584	2714	49	0	98
Moderate	0	327	1618	0	13
Vigorous	0	11	56	0	0
Transition	13	289	31	0	451

Table 36. Confusion matrix for intensity classification with the BB-Par-CRF method and one stage estimation strategy with physical activity type known, applied to the free living wrist data from Sasaki [2013], all subjects combined.

Labeled Class	Predicted Class				
	Sedentary	Light	Moderate	Vigorous	Transition
Sedentary	1127	921	6	0	1
Light	868	2329	238	0	10
Moderate	68	488	1399	0	3
Vigorous	0	5	62	0	0
Transition	91	559	131	0	3

Table 37. Confusion matrix for intensity classification with the BB-Par-CRF method and two stage estimation strategy with physical activity type imputed in the first stage, applied to the free living wrist data from Sasaki [2013], all subjects combined.

each combination of classifier, accelerometer location, and number of classes, along with sets of contrasts between these quantities in Figures 94 through 99. These estimates confirm the observations we made from the box plots above.

The estimates in Figure 95 show that classification was easier with the ankle and hip data than it was with the wrist data. The average improvements in the proportion correct from using the ankle or hip instead of the wrist vary, but are as large as about 0.21. Figures 96, 97, and 98 show that most classification methods achieved about the same average proportion correct, except for the **Gaussian-mixt-HMM** method. Figure 99 shows that if the true class were available it could provide information about physical activity intensity that is not in the accelerometer signal. However, using a two-stage procedure is ineffective, and does not lead to increased success in classifying activity intensity.

8.6 Discussion

In this Section, we tie together the results of the applications to classification of physical activity intensity in our three data sets. We have focused here on how the accelerometer location, classification method, and strategy for using information about activity type impact classification performance.

Our results regarding the impact of the accelerometer location on classification of physical activity intensity are very similar to the results we saw in Chapter 7 when classifying according to activity type. For both the free living data from Sasaki [2013] and the lab data from Mannini et al. [2013], we saw that the accelerometer location can have a large impact on classification success, with classification generally being more difficult with data from the wrist than it is with data from the ankle or hip. The improvement that can be realized by switching away from the wrist ranged from about 0.03 to 0.21 in the proportion correct, and was larger in the free living data than it was in the data from Mannini et al. [2013]. Surprisingly, we did not observe an effect of accelerometer location on classification performance when classifying according to intensity level in the laboratory data from Sasaki [2013]. However, the direction of the improvement was consistent in the other two data sets, and consistent with the pattern we saw with classification according to activity type in Chapter 7.

We saw less variability among the different classification methods when classifying accord-

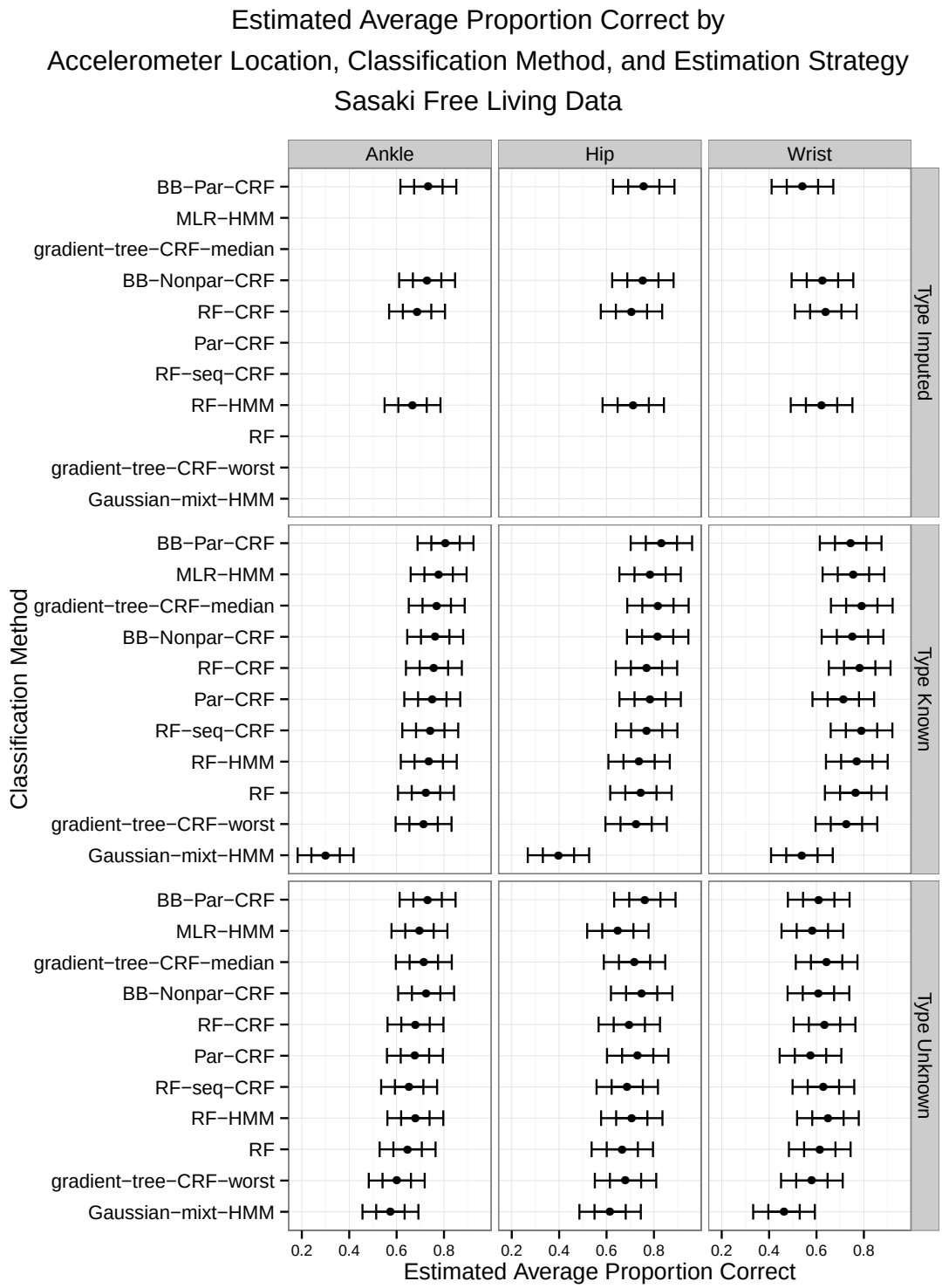


Figure 94. Point and interval estimates for the fixed effects parameters in model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 94, 95, 96, 97, 98, and 99.**

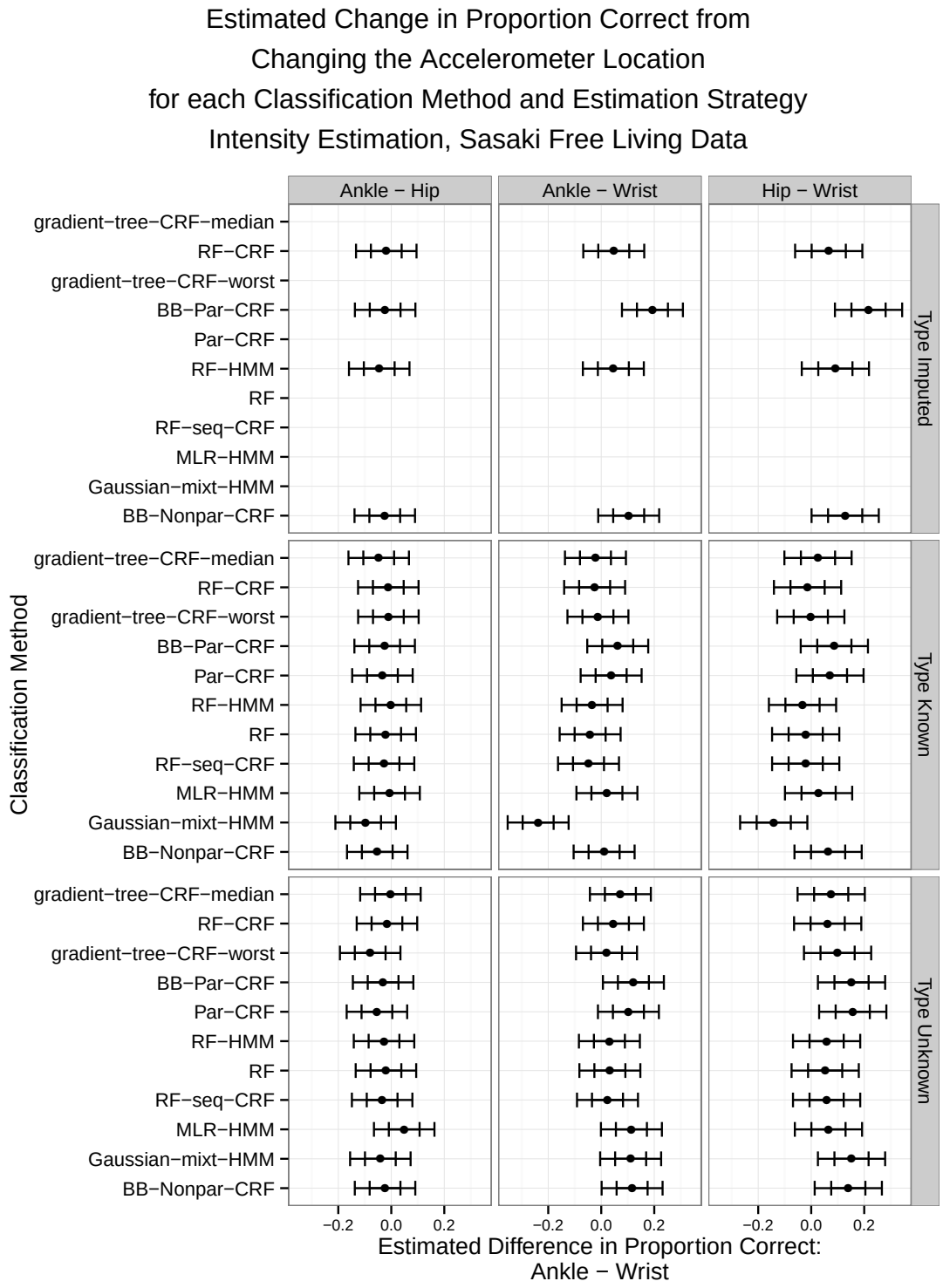


Figure 95. Point and interval estimates for the difference in performance between the ankle and wrist accelerometer locations for each classification method, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 94, 95, 96, 97, 98, and 99.**

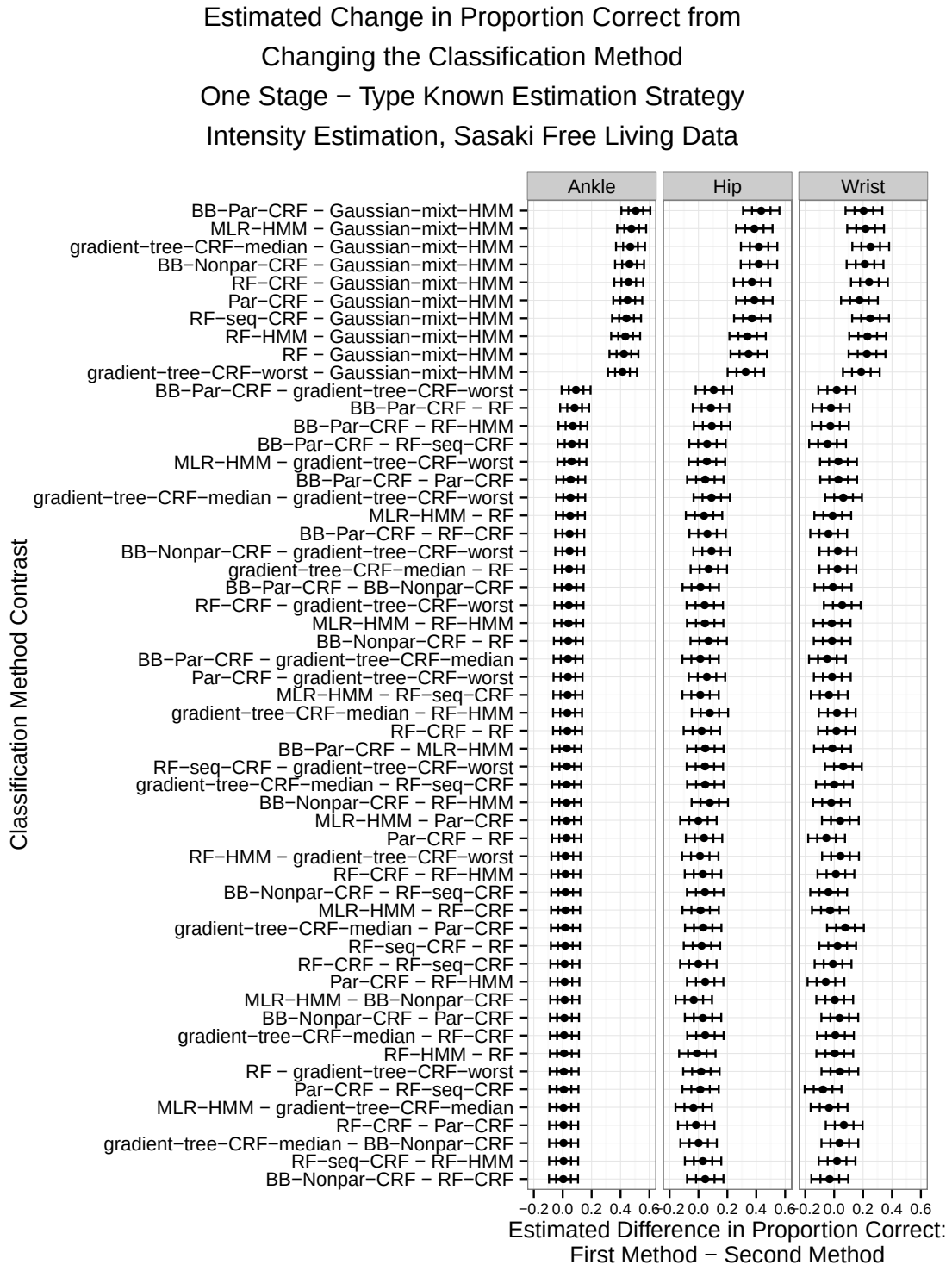


Figure 96. Point and interval estimates for the difference in performance between each pair of classification methods using a one stage estimation strategy where the true activity type in each window is known, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 94, 95, 96, 97, 98, and 99.**

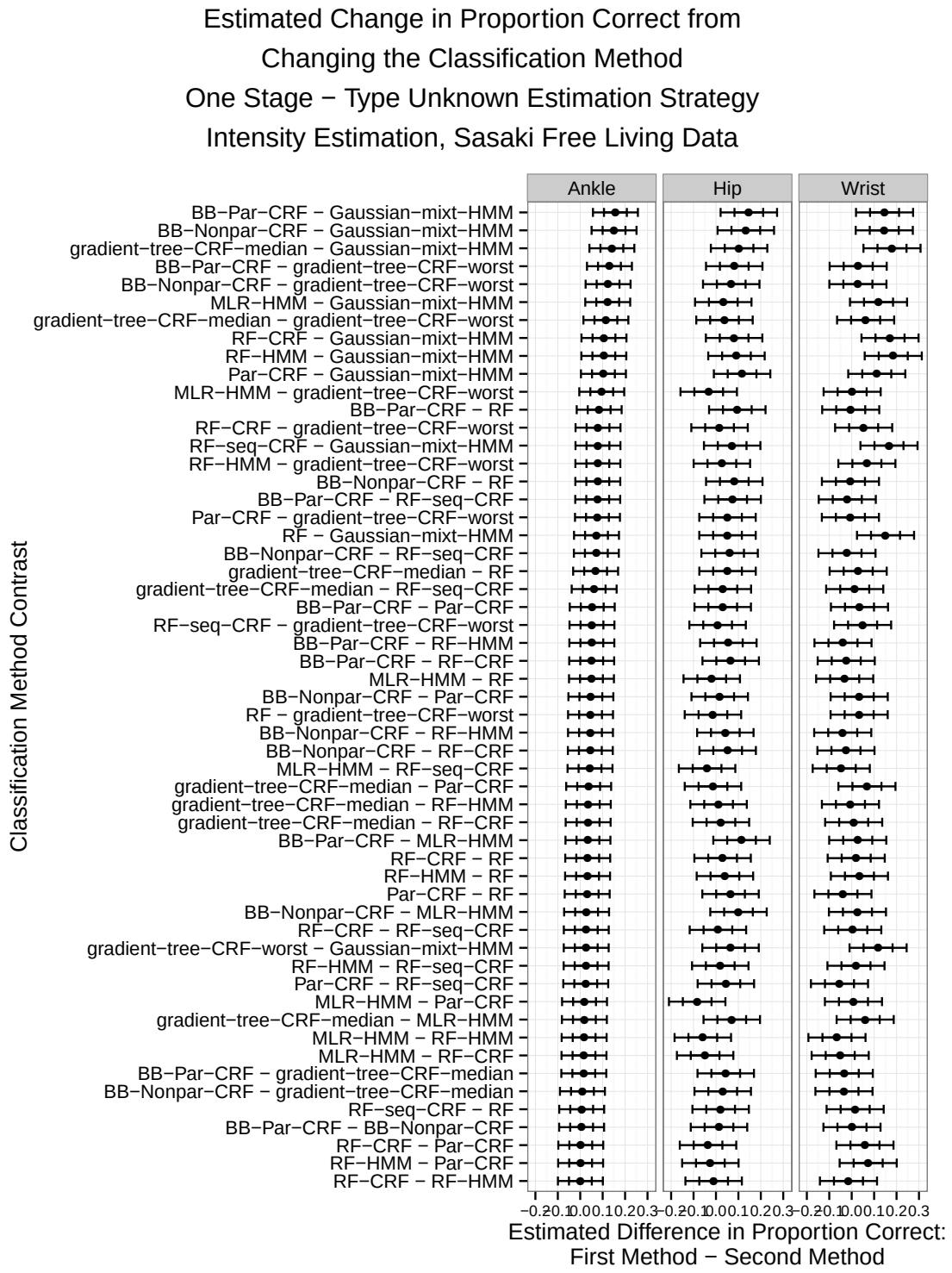


Figure 97. Point and interval estimates for the difference in performance between each pair of classification methods using a one stage estimation strategy where the true activity type in each window is unknown, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 94, 95, 96, 97, 98, and 99.**

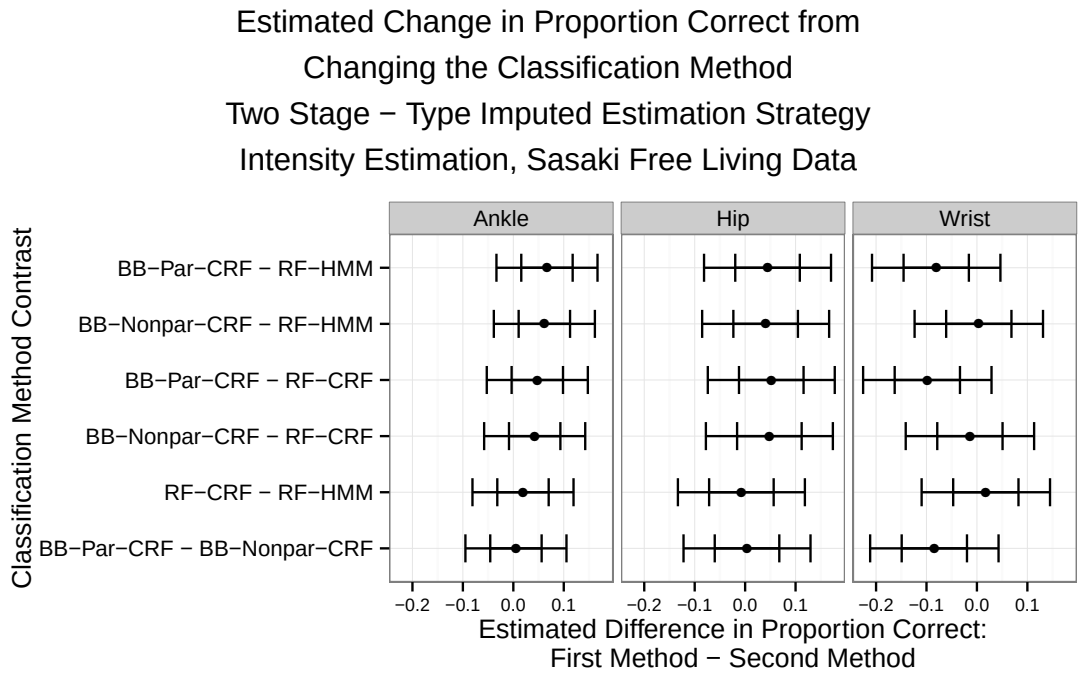


Figure 98. Point and interval estimates for the difference in performance between each pair of classification methods using a two stage estimation strategy where the activity type in each window is estimated in the first stage and that estimate is used as an input to intensity estimation in the second stage, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 94, 95, 96, 97, 98, and 99.**

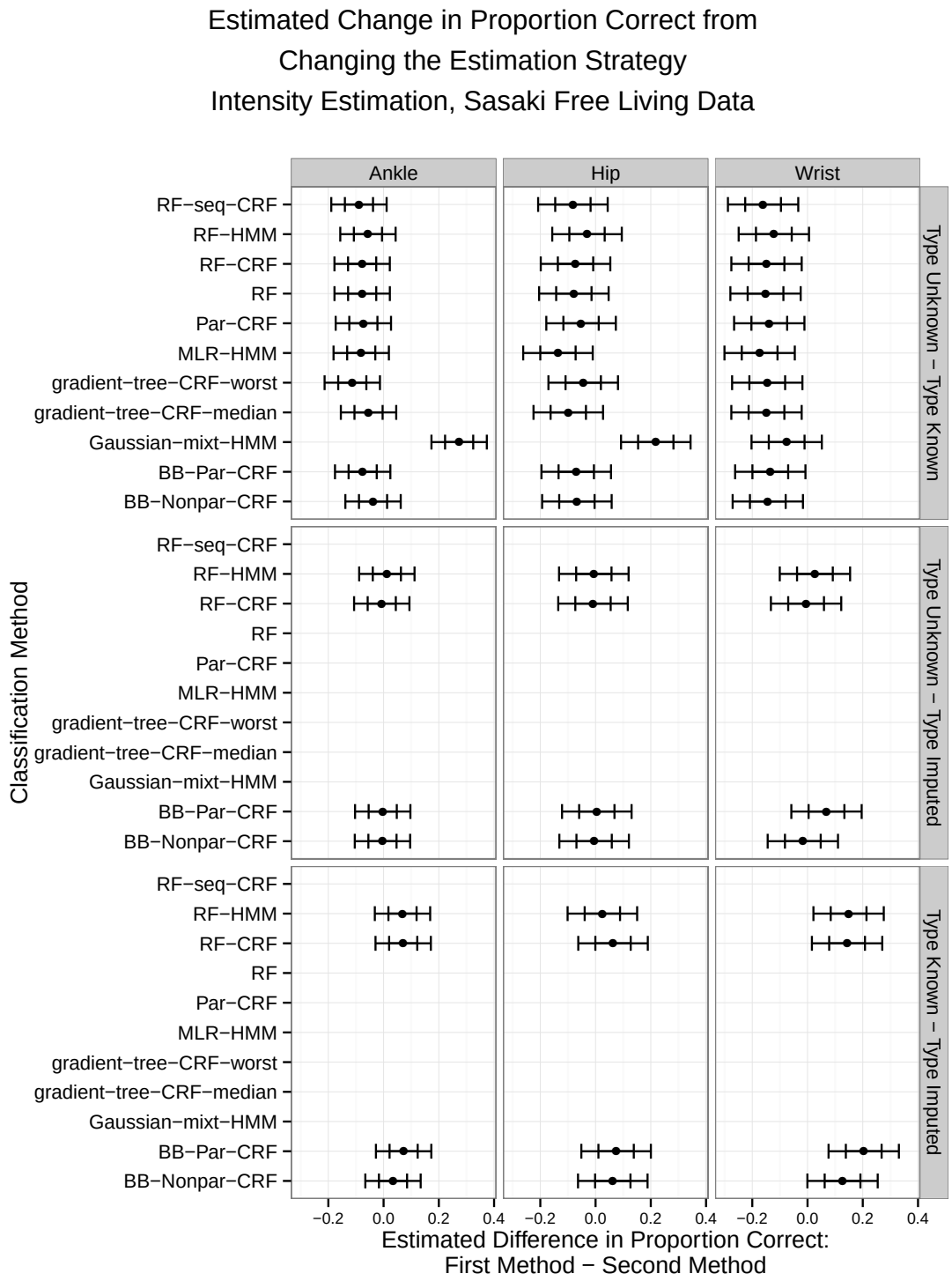


Figure 99. Point and interval estimates for the difference in performance between each pair of estimation strategies, based on model (8.3.1)**. The inner marks indicate endpoints for individual 95% confidence intervals; the outer marks indicate endpoints for familywise 95% confidence intervals for all intervals in Figures 94, 95, 96, 97, 98, and 99.**

ing to intensity level than we did when classifying according to activity type. However, we did consistently observe lower performance from the **Gaussian-mixt-HMM** method than we had with the other methods. The **Par-CRF**, **MLR-HMM**, and **RF-HMM** methods also had inconsistent performance, generally doing well but with a much lower proportion correct in a few cases. As we saw in the applications to activity type classification in Chapter 7, the performance of the **gradient-tree-CRF** method is sensitive to how the data are partitioned into training and validation data sets. The median performance of the method with respect to this partitioning is in line with the best of the other methods, but the worst case performance is much lower. The remaining methods, including the three methods we proposed in Chapter 5 and the static **RF**, all did about as well as each other.

We also compared several approaches to using information about the type of activity an individual is engaged in when classifying that activity according to its intensity level. In the data from Mannini et al. [2013] and the free living data from Sasaki [2013], we found that if the true activity type labels were observed, using them in models for intensity could lead to improved classification performance. Plots of the results of the application to the lab data from Sasaki [2013] indicated that a similar effect was likely present with those data, but the effect size was very small.

However, it is less clear whether it is useful to impute the activity type for use in an intensity classification model if it is not observed. Of all the applications of the two-stage estimation procedure across three data sets and multiple accelerometer locations with four classification methods, only one combination of these factors resulted in an appreciable performance gain relative to the one-stage procedure that did not use any information about activity type. The proportion correct achieved by the **RF-CRF** method using the wrist data from Mannini et al. [2013] improved by about 0.04. All other point estimates for the change in the proportion correct that is realized by using the two stage method instead of ignoring the activity type are less than 0.025 and have (approximate) confidence intervals that include 0. In fact, with the free living wrist data and the **BB-Par-CRF** method, the estimated effect size is about -0.07 – indicating that using the two stage procedure actually resulted in worse classification results.

There are several differences between our work and the work of Albinali et al. [2010] that could explain why we found that the two stage procedure results in little to no improvement in estimation of intensity while they found that the two stage procedure was better than the one

stage procedure. The differences between our studies are as follows:

1. **Data sets used for model estimation and evaluation**: A major limitation of the study by Albinali et al. [2010] is that in their work, the two-stage intensity estimation model is trained on the same data set on which it is evaluated, while the one-stage methods are trained on different data sets than the one that is used to evaluate their performance. They are careful to use cross-validation to evaluate the performance of classification methods, but this cannot control for the fact that the models were trained with data from different studies with different data collection protocols and different sets of activities performed by the subjects. This means that in their results, the impact of using a two-stage procedure is confounded with the impact of training the model on the same data set that it is evaluated with rather than training the model and evaluating it using different data sets. Our analysis resolves this problem by more systematically comparing models that do and do not use information about activity type, using the same data sets to estimate the model parameters and evaluate performance for all models and estimation strategies with a leave-one-subject-out procedure.

2. **Model flexibility**: The study by Albinali et al. [2010] also confounds model flexiblity with the use of imputed activity types in estimating activity intensity. The one-stage intensity estimation procedures they use consist of a single linear regression model relating a univariate summary of the signal from a single accelerometer in each second to intensity levels. On the other hand, their two-stage procedure uses a flexible decision tree to perform the stage 1 classification based on a set of several time and frequency domain features using information from all three axes of the recordings from three different accelerometers. They then fit a separate linear model analogous to that used in the one stage procedure for each activity type. The resulting two-stage model is much more flexible and uses much more information from the accelerometers than the one-stage procedure did. With this design, it is impossible to say how much of the observed improvement in intensity estimation is due to the use of imputed activity types, and how much is due to the increased modeling flexibility and use of a richer set of features from more accelerometers. Again, our work resolves these problems by fitting essentially the same models with and without imputed activity type information, thereby isolating the impact of using estimated activity types on

intensity estimation performance.

3. **Data collection setting**: Here we are referring to the distinction between the laboratory and free living settings. Subjects in the study by Albinali et al. [2010] performed one of two routines with specified activities in a fixed order. Although these routines included a component "in the field", data collection is still more similar to the laboratory data collection of Mannini et al. [2013] and Sasaki [2013] than the free living component of the study by Sasaki [2013]. This is important because we found in Chapter 7 that classification of physical activity type was much more successful with laboratory data than it was with free living data. In our results, the two stage procedure offered the largest gains relative to the one stage procedure in the laboratory data from Mannini et al. [2013]. In that case, all point estimates for the difference between these estimation strategies were positive. On the other hand, with the free living data from Sasaki [2013], in every case the difference in performance between the two estimation strategies was small; in several cases, the point estimates for these differences indicated that the two stage procedure did worse than the one stage procedure. One possible explanation for this behavior is that the first stage classification results are much better with laboratory data than they are with free living data, and therefore offer more in the way of useful information when estimating activity intensity.

4. **First stage classification categories**: Albinali et al. [2010] used a much finer system of activity type categories than we did. It therefore stands to reason that their first stage classification is more informative about intensity than ours. However, a finer system of first stage classification also makes implementation in the free living setting that much more difficult. First stage classification rates were very low in the free living data even with our more limited set of 4 categories.

5. **Problem solved**: Albinali et al. [2010] focused on regression models estimating numeric kilocalories, while we developed classification models for intensity. It may be that the information provided by imputed activity types yields larger performance gains for regression than it does for classification.

Unfortunately, these differences between our studies are confounded, so we cannot determine which of them is most important in explaining the differing conclusions drawn from our results and those of Albinali et al. [2010] regarding the merits of using the two stage procedure.

However, in our opinion, the limitations of the work of Albinali et al. [2010] discussed in points 1 and 2 above are very important and make it essentially impossible to draw the conclusion that imputed activity types are informative about intensity from their results. To be clear, we think their methods are interesting; the results they present simply do not allow us to isolate the impact of using a two stage procedure on intensity estimation. Our view is that the remaining factors discussed above combine to tell a consistent story: If we could improve the first stage classifiers sufficiently that we could achieve high first-stage classification rates with a detailed system of activity type classes in a free living setting, we might be able to achieve small gains in estimation of intensity by using a two stage procedure. However, with the current state of the art, activity type classification rates are too low in the free living setting and the two stage procedure may actually yield worse performance with free living data.

CHAPTER 9

CONCLUSION

In this work, we have discussed methods for classification of physical activity according to activity type or intensity level using CRF models. Our primary contributions to the statistical literature are three new estimation algorithms for CRF models. Our primary contributions to the literature on classification of physical activity using accelerometer data are evidence that models like CRFs can provide improved classification performance relative to other models in some settings, and a careful study of the impact of the treatment of activity type in estimation of activity intensity. We review these contributions, their limitations, and opportunities for future work addressing these limitations in this Chapter. We begin with a discussion of our model and estimation strategies in Section 9.1, and then review what we have learned about the impact of other factors such as the data collection setting, accelerometer placement, and the use of two-stage classification procedures on classification of physical activity in Section 9.2.

9.1 Discussion of Model and Estimation Strategies

We developed three new estimation procedures for linear chain CRFs. Of these, the **BB-Par-CRF** and **BB-Nonpar-CRF** methods performed well relative to the other methods we compared in our simulation study and applications to classification of physical activity type and intensity. The performance of the third approach, represented by the **RF-CRF** and closely related **RF-seq-CRF** methods, was less consistent.

Our most successful approach was the **BB-Par-CRF** method, in which we used a parametric model specification and combined bagging and boosting to estimate the model parameters. In many cases, this method had similar classification performance to the **Par-CRF** method, which

used the same parametric model specification and obtained parameter estimates via maximum likelihood. However, the **BB-Par-CRF** did outperform the **Par-CRF** in several cases for classification of both activity type and intensity. Although we do not have a formal proof that our estimation procedure acts as a regularizing technique, this is one way of interpreting similar boosting algorithms: the parameter estimates slowly converge to the maximum likelihood estimates, but we use early stopping with evaluation of classification performance on a validation subset to halt the estimation process before the maximum likelihood estimates have been reached and we have overfit the data. This view of our bagging and boosting procedure could explain the differences in performance between our method and the method using maximum likelihood estimation.

Our comparisons did not include an L_2 regularized parametric CRF, which is another common approach to reducing overfitting in CRF models. We do not necessarily believe that the **BB-Par-CRF** would outperform an L_2 regularized CRF in terms of classification performance, but boosting does offer the potential for reduced computation time. Estimation with an L_2 penalty requires selection of the penalty factor. This is usually done through cross validation. This can be very slow. Our boosting procedure also uses cross validation to select a tuning parameter: the number of boosting iterations. However, optimization with respect to this discrete tuning parameter is simpler than optimization of the continuous L_2 penalty.

The other method we proposed that worked well was the **BB-Nonpar-CRF** method. In this approach we used a non-parametric specification for the component of the model relating the accelerometer features to the activity types, and we again combined bagging and boosting to perform the parameter estimation. This method is similar to the ideas introduced by Dietterich et al. [2004], with two modifications: (1) the use of bagging and a LOP to combine the model fits obtained with each bagged training sample, and (2) the use of a random subset of features for each tree in the boosting process. As we discussed in Section 7.6, it appears that the most important of these changes was the use of a LOP to combine results from several fits using different partitions of the data into training and validation subsets. Bagging and random subset selection had limited impact on the predictive performance of the model fit. A simpler and faster method than our **BB-Nonpar-CRF** that would achieve similar classification performance would be to perform a procedure analogous to 10-fold cross validation to obtain 10 model fits based on different partitions of the data into training and validation sets, and then combine those 10 models using a LOP.

Our final methods combined the ideas behind estimation of static random forests with the CRF model. We proposed two variations on this idea using different strategies for bagging the observations: the **RF-CRF** and **RF-seq-CRF**. These methods performed well in many cases, but had large drops in classification performance for a few of the more difficult classification tasks. Additionally, we did not find evidence that bagging at the level of observations for individual time points as in the **RF-CRF** method led to improvements in classification performance relative to the simpler method of bagging complete observation sequences as in the **RF-seq-CRF** method.

Through simulation studies and applications to classification with real data, we have demonstrated that CRFs can offer better classification performance than a HMM that models the observation distributions with a mixture of Gaussians, even after an initial transformation of the features to approximate normality. In our view, this result holds for our particular application and data set because the accelerometer features have a complex joint distribution that is not well modeled by a mixture of Gaussians with a small number of mixture components, and our sample size is fairly small relative to the number of features we have used so that there is not enough information in the data to use a large number of mixture components. An alternative approach to modeling that we have not explored would be to select a small subset of the features that are most informative, and to model the joint distribution of those features more carefully. We have chosen to use a large number of features because we believe that these features are all likely to be informative about activity type and intensity.

Our simulation studies and data applications have also demonstrated that accounting for sequential dependence in activity type and intensity can lead to improved classification performance. The improvement in classification performance that was achieved by using a dynamic model instead of a static model was largest in the laboratory data sets, and in the cases of the simulation study with low Bayes error rate. As we have discussed, one possible explanation for this is that when the overall classification rates are low, the dynamic models have less to gain from sharing information across time. Under this theory, if the overall classification rates can be improved in the free living setting, the difference in performance between the static and dynamic methods with free living will be more apparent than they were in our applications. We believe that improved classification rates are possible in the free living setting; we will discuss this issue more in Section 9.2.

The method McShane et al. [2013] proposed for combining a static classification method with

the dependence structure of a HMM performed very well in almost every case, particularly when the static classification method was a random forest. This method has the major advantage over CRFs of being much faster to estimate. It can be viewed as an intermediate solution between HMMs and CRFs in which the parameters governing transitions between classes in adjacent time points are estimated via maximum likelihood in the HMM, and the parameters describing the relationship between the features and the classes are estimated discriminatively. The good performance of this approach suggests that in applications to physical activity classification, discriminative estimation of the parameters describing transitions between activity types is less beneficial than discriminative estimation of the parameters describing the relationship between the features and the activity classes.

We now turn to some criticisms of our methods. We focus first on limitations of the basic first-order Markov linear chain CRF model that forms the foundation of all three of our methods, and discuss several modifications that could be made to address those limitations. Some of these modifications would be feasible with our data, but others would likely require more data. Nevertheless, we mention them as possibilities for future research. We then discuss alternatives to the more general approach we have taken of dividing the accelerometer signal up into non-overlapping windows and summarizing the signal in each window with a vector of features.

One way of thinking about limitations of our model is to examine each of the assumptions we made in Section 5.1, and whether those assumptions can be relaxed. Our first simplifying assumption was that the different observation sequences are independent of each other given the observed features, and that the observation sequences for each subject are described by a CRF with the same parameter values. The second part of this assumption is clearly false. Different people people engage in different activity types with varying frequencies: for example, some people walk or ride a bike more often than others. Different people also have different patterns of motion, so that the locations in the space of features that are associated with each activity type vary across different individuals. We have not addressed this variation across individuals in our models because we believe that we would need data for more subjects and for more time per subject in order to model this variation. However, it would be interesting to explore the use of hierarchical models to allow for variation among individuals in future work.

Our second simplifying assumption was that activity types obey a first-order Markov structure, so that in the graph representing conditional dependence relationships, each node $Y_{i,t}$ is

connected only with its immediate neighbors $Y_{i,t-1}$ and $Y_{i,t+1}$. Again, this assumption is clearly incorrect. As we have discussed, we believe that using a more flexible model for sequential dependence would be inappropriate for the data gathered in the laboratory, where subjects were told the order and duration of activities to perform. We focused on the model specification that we used so that we could apply the methods we developed with all three of our data sets.

However, we believe that a more flexible model for time dependence would be appropriate in the free living setting. There are several ways that the first order structure used in our models could be relaxed. One idea would be to adapt the variable duration models that have been proposed for HMMs for use with CRFs. Roughly, these methods separate the modeling for transitions from a class back to the same class from the modeling for transitions between different classes, and use a more general specification for the transitions back to the same class. Another alternative would be to simply expand to a second order Markov structure. Ordinarily, this would involve a large increase in the number of model parameters. For instance, with 6 classes moving from a first order structure to a fully parameterized second order structure means the the number of parameters used in modeling transitions jumps from $6^2 = 36$ to $6^3 = 216$. However, we have coded the data in such a way that all moves from one activity category to another move through the Transition category. We could use that information to formulate a second order model with a relatively small number of transition parameters. A third alternative would be to use a specification that is common in CRF models in which the compatibility of observing states r and s at times t and $t + 1$ depends on the feature values at those times. For example, this possibility is allowed for by the parameterization used by Dietterich et al. [2004].

While any one of these approaches might be feasible to estimate with our free living data set, we believe that they would lead to only modest improvements in classification performance. This is because with the free living data, our models that accounted for time dependence models led to only negligible improvements in classification performance relative to the static methods. It seems likely that without more fundamental changes in the available data (which we will discuss in Section 9.2), small tweaks to the time dependence model will not lead to substantial improvements to classification performance with free living data.

Another restriction imposed in our formulation of the CRF is that the accelerometer features in each time window are only informative about the activity type and intensity at that time. In the graphical representation of the model, this restriction is reflected in the fact that there is an

edge connecting $Y_{i,t}$ with $\mathbf{x}_{i,t}$, but not with $\mathbf{x}_{i,t^*}$ for any other time index t^*. It seems likely that the accelerometer features are also informative about the activity type and intensity in nearby windows. This feature of the model could be easily changed by inserting links between $Y_{i,t}$ and $\mathbf{x}_{i,t^*}$ for values of t^* within $\pm w$ indices of t in the graph. In fact, most specifications of CRF models allow for this type of dependence. We adopted the formulation in which $Y_{i,t}$ depends only on $\mathbf{x}_{i,t}$ so that we could compare the different treatments of the feature vectors in a CRF and a HMM with similar structures. However, including these edges in the graph could lead to improved classification results.

All of the methods that we included in our comparisons are based on the same general approach: we divide the acceleration signal up into non-overlapping windows 12.8 seconds long, extract a vector of features summarizing the acceleration signal in each window, and fit a model that relates these features to the activity type in each window. However, this is not the only option for modeling these data, and other approaches we have not considered may offer superior performance. One option that was suggested by Zheng et al. [2013] and Lester et al. [2005] is to combine information from several overlapping windows of different lengths. This idea could be implemented within a single CRF by expanding $\mathbf{x}_{i,t}$ to include features from multiple window lengths, or by using a LOP to combine inferences from multiple CRFs operating at different time scales. Another option would be to abandon the windowing approach altogether and model the accelerometer signal directly. We are not aware of any previous methods that take this approach, but it may be worth exploring in future research. It may also be beneficial to explore the use of new features.

9.2 Discussion of Applications to Physical Activity Classification

In this Section, we discuss what we have learned about the impact of several different factors related to data collection and processing on classification performance. First, as with previous researchers, we found that classification is much more difficult with free living data than it is with data gathered in the laboratory. We discuss several reasons for this and suggest some ideas for obtaining improved classification rates with free living data. We also review our findings relating to the impact of the accelerometer location on classification performance, and on the use of information about physical activity type in estimating its intensity.

We observed a large drop in classification performance when using free living data instead of data collected in the laboratory. It is natural to consider what the causes of this drop in performance are and how they can be addressed. Moving from the laboratory to the free living setting introduces two related challenges: (1) individuals' movement patterns are more complex, and (2) data collection is more difficult.

The increased complexity of movement patterns in the free living setting means that there is more variability in the accelerometer signal associated with each activity class. For instance, locomotion in the laboratory setting might involve walking back and forth down an empty hallway with the arms free to move at the sides. This leads to a strong, uninterrupted cyclical pattern in the acceleration recordings. In the real world, an individual may need to carry objects, navigate around other pedestrians or obstacles, move up and down inclines or steps, and occasionally pause for short periods of time. This increased variability in the behaviors that constitute locomotion results in less consistent and easily identifiable patterns of acceleration, and in turn makes the classification problem fundamentally more challenging. This is an important factor leading to lower classification rates in free living data that cannot be solved, although some of the revised modeling options we discussed in the previous Section may help the situation.

Data collection is also much more difficult in the free living setting. In particular, as we discussed in Chapter 3 our recorded labels for physical activity type are less reliable. This impacts the training of the classification methods, since the classifiers may learn to associate certain patterns in the accelerometer signal with the incorrect labels. It also impacts our scoring of the success of the classifier through leave one subject out cross validation, since a predicted class label may give an accurate description of a subject's behavior but disagree with the recorded class. One way to limit this problem would be to incorporate a method of validating the class labels in the study design, for instance by recording videos of subjects while they are wearing the accelerometers.

Another challenge is that the free living data from Sasaki [2013] contain data for fewer than half as many subjects as we had in either of the data sets gathered in the laboratory. We believe that it would be helpful to have data from a larger number of subjects because of the variability in accelerometer signals across subjects illustrated in Figure 9. With only 15 subjects, the accelerometer features one particular subject sometimes lie in a region of the feature space that does not contain data from any of the other subjects. For example, in Figure 9, the data from

subject 9 include a cluster of points around 1.05 on the horizontal axis and 0 on the vertical axis but there are no data in this region for any of the other subjects. To some extent this is an irresolvable consequence of the fact that different subjects have different patterns of movement. However, this problem would be mitigated to some extent if we had data from more subjects. Also, although the free living data have observations over a longer time period for each subject than the laboratory data do, they do not necessarily have recordings for a variety of activities for each subject. For example, 6 of 15 subjects engaged in fewer than 5 minutes of time labeled as locomotion during the time they were observed, including one subject with no time in the locomotion category. Similarly, only three subjects engaged in Recreational activity in the free living component of the study, with a total of about 2 hours and 15 minutes spent in that activity category.

It is difficult to determine how much of the drop in classification performance from the laboratory setting to the free living setting is due to the inevitable increased difficulty of the classification problem and how much is due to other factors that we could address such as increased error rates in the recorded labels or the smaller number of subjects and limited amount of time spent in some activity categories. However, we are encouraged by the plot in Figure 63, which indicates to us that at least some of the discrepancies between the predicted and recorded class labels may be due to errors in data collection. We believe that better classification rates could be achieved in the free living setting if more training data were available and we could ensure that the activity type labels were accurate. It may also be helpful to use activity type categories describing specific postures, and avoid more ambiguous categories such as Moving Intermittently.

As part of our applications to classification with real physical activity data, we investigated the impact that the accelerometer placement has on the success of the classifiers. Our results confirm the earlier work by Mannini et al. [2013] and Sasaki [2013] indicating that the best classification results can be achieved using data collected using an accelerometer placed at the ankle. Aside from whether the data were collected in the laboratory or in a free living setting, the accelerometer location is often the most important determinant of classification success. For instance, in the free living data from Sasaki [2013] all of the classification methods we compared had substantially improved performance when using data from the ankle instead of the wrist, and most of these classification methods did about as well as each other.

We also contributed to the discussion in the literature regarding the merits of using informa-

tion about a subject's activity type when estimating the intensity of their activity. Our results suggest that if the true activity type were known, it would provide useful information about activity intensity that is not contained in the accelerometer signal. However, in most cases the activity type is unknown and must be estimated. In the laboratory setting, where activity type classification methods perform well, we found some evidence that small gains in classification of intensity could be achieved by using a two-stage procedure in which we first classify activity according to type and then use those type estimates as an input to methods that classify according to intensity. In the free living setting, where classification according to activity type was less successful, using this two stage procedure did not lead to improved classification according to intensity level. In some cases, the two stage procedure actually had worse classification rates than a one stage procedure that did not make use of activity type. As we have discussed, it may be possible to improve the activity type classification results in free living data. This could lead in turn to improved performance of the two-stage procedure in free living data.

BIBLIOGRAPHY

Barbara E Ainsworth, William L Haskell, Stephen D Herrmann, Nathanael Meckes, David R Bassett, Catrine Tudor-Locke, Jennifer L Greer, Jesse Vezina, Melicia C Whitt-Glover, and Arthur S Leon. 2011 compendium of physical activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, 43(8):1575–1581, 2011.

Fahd Albinali, Stephen Intille, William Haskell, and Mary Rosenberger. Using wearable activity type detection to improve physical activity energy expenditure estimation. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 311–320. ACM, 2010.

Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.

Michael M. Anderson. Physical activity recognition of free-living data using change-point detection algorithms and hidden Markov models. Master's thesis, Oregon State University, June 2013.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Hagai Attias. Independent factor analysis. *Neural computation*, 11(4):803–851, 1999.

Scott Axelrod, Ramesh Gopinath, and Peder A Olsen. Modeling with a subspace constraint on inverse covariance matrices. In *INTERSPEECH*, 2002.

Jason Baldridge and Miles Osborne. Active learning and logarithmic opinion pools for HPSG parse selection. *Natural Language Engineering*, 14(02):191–222, 2008.

Jeffrey D Banfield and Adrian E Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821, 1993.

Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, pages 1–17. Springer, 2004.

Maurice S Bartlett. *An introduction to stochastic processes: with special reference to methods and applications*. Cambridge University Press, Cambridge, 3 edition, 1978.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

Halima Bensmail and Jacqueline J Meulman. Model-based clustering with noise: Bayesian inference and estimation. *Journal of classification*, 20(1):049–076, 2003.

Halima Bensmail, Gilles Celeux, Adrian E Raftery, and Christian P Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.

Axel Bernal, Koby Crammer, Artemis Hatzigeorgiou, and Fernando Pereira. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS computational biology*, 3(3):488–497, 2007.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

Jeff A Bilmes. Factored sparse inverse covariance matrices. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II1009–II1012. IEEE, 2000.

Alberto G Bonomi, AH Goris, Bin Yin, and Klaas R Westerterp. Detection of type, duration, and intensity of physical activity using an accelerometer. *Med Sci Sports Exerc*, 41(9):1770–1777, 2009a.

Alberto G Bonomi, Guy Plasqui, Annelies HC Goris, and Klaas R Westerterp. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. *Journal of Applied Physiology*, 107(3):655–661, 2009b.

Charles Bouveyron, Stéphane Girard, and Cordelia Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007.

Leo Breiman. Some infinity theory for predictor ensembles. Technical report, Technical Report 579, Statistics Dept. UCB, 2000.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Wadsworth, Belmont, CA, 1984.

Andrew Brown and Geoffrey E Hinton. Products of hidden Markov models. In *Proceedings of Artificial Intelligence and Statistics*, pages 3–11. Citeseer, 2001.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5): 1190–1208, 1995.

Olivier Cappé, Christian P Robert, and Tobias Rydén. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):679–700, 2003.

Peter Carbonetto, Nando De Freitas, Paul Gustafson, Natalie Thompson, et al. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *Artificial Intelligence and Statistics (AI & Statistics' 03)*, 2003.

G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Gilles Celeux and Gérard Govaert. Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793, 1995.

Yu-Shu Chen and Yi-Ming Chen. Combining incremental hidden Markov model and Adaboost algorithm for anomaly intrusion detection. In *Proceedings of the ACM SIGKDD Workshop on Cybersecurity and intelligence informatics*, pages 3–9. ACM, 2009.

Scott E Crouter, Kurt G Clowers, and David R Bassett. A novel method for using accelerometer data to predict energy expenditure. *Journal of applied physiology*, 100(4):1324–1331, 2006.

Shalom Darmanjian and Jose C Principe. Boosted and linked mixtures of HMMs for brain-machine interfaces. *EURASIP Journal on Advances in Signal Processing*, 2008(1):216453, 2008.

Shalom Darmanjian, Antonio Paiva, Jose Principe, and Justin Sanchez. Hierarchal decomposition of neural data using boosted mixtures of hidden Markov chains and its application to a BMI. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages 3062–3067. IEEE, 2007.

Sanne I de Vries, Francisca Galindo Garre, Luuk H Engbers, Vincent H Hildebrandt, and Stef Van Buuren. Evaluation of neural networks to identify types of activity using accelerometers. *Med Sci Sports Exerc*, 43(1):101–7, 2011.

Petros Dellaportas and Ioulia Papageorgiou. Multivariate mixtures of normals with unknown number of components. *Statist. Comp*, 16:57–68, 2006.

Satya Dharanipragada and Karthik Visweswariah. Gaussian mixture models with covariances or precisions in shared multiple subspaces. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1255–1266, 2006.

Jean Diebolt and Christian P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B (Methodological)*, 56(2): 363–375, 1994.

Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40 (2):139–157, 2000.

Thomas G. Dietterich, Adam Ashenfelter, and Yaroslav Bulatov. Training conditional random fields via gradient tree boosting. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 28–, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015428. URL `http://doi.acm.org/10.1145/1015330.1015428`.

Christos Dimitrakakis and Samy Bengio. Boosting HMMs with an application to speech recognition. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–621. IEEE, 2004.

Carlos Domingo and Osamu Watanabe. MadaBoost: A modification of AdaBoost. In *COLT*, pages 180–189. Citeseer, 2000.

Glen E Duncan, Jonathan Lester, Sean Migotsky, Jorming Goh, Lisa Higgins, and Gaetano Borriello. Accuracy of a novel multi-sensor board for measuring physical activity and energy expenditure. *European journal of applied physiology*, 111(9):2025–2032, 2011.

Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

Narayanan U Edakunni, Gary Brown, and Tim Kovacs. Boosting as a product of experts. *arXiv preprint arXiv:1202.3716*, 2012.

Bradley Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

Shlomo Engelberg. *Random signals and noise: a mathematical introduction*. CRC Press, Boca Raton, FL, 2007.

Bahar Erar. Mixture model cluster analysis under different covariance structures using information complexity. Master's thesis, The University of Tennessee, Knoxville, 2011.

Miikka Ermes, Juha Parkka, Jani Mantyjarvi, and Ilkka Korhonen. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *Information Technology in Biomedicine, IEEE Transactions on*, 12(1):20–26, 2008.

F Foerster, M Smeja, and J Fahrenberg. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Computers in Human Behavior*, 15(5):571–583, 1999.

Say Wei Foo and Liang Dong. A boosted multi-HMM classifier for recognition of visual speech elements. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 2, pages II–285. IEEE, 2003.

Say Wei Foo, Yong Lian, and Liang Dong. Recognition of visual speech elements using adaptively boosted hidden Markov models. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):693–705, 2004.

G David Forney Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`.

Chris Fraley, Adrian E. Raftery, Thomas Brendan Murphy, and Luca Scrucca. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*, 2012.

Patty S Freedson, Edward Melanson, and John Sirard. Calibration of the computer science and applications, inc. accelerometer. *Medicine and science in sports and exercise*, 30(5):777–781, 1998.

Yoav Freund. An adaptive version of the boost by majority algorithm. *Machine learning*, 43(3):293–318, 2001.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science + Business Media, 2006.

Mark JF Gales. Semi-tied covariance matrices for hidden Markov models. *Speech and Audio Processing, IEEE Transactions on*, 7(3):272–281, 1999.

Mark JF Gales. Maximum likelihood multiple subspace projections for hidden Markov models. *Speech and Audio Processing, IEEE Transactions on*, 10(2):37–47, 2002.

Giuliano Galimberti, Angela Montanari, and Cinzia Viroli. Latent classes of objects and variable selection. In *COMPSTAT 2008*, pages 373–383. Springer, 2008.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Bernd Gutmann and Kristian Kersting. Stratified gradient boosting for fast training of conditional random fields. In *Proceedings of the 6th International Workshop on Multi-Relational Data Mining*, pages 56–68, 2007.

Illapha Cuba Gyllensten and Alberto G Bonomi. Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. *Biomedical Engineering, IEEE Transactions on*, 58(9):2656–2663, 2011.

Thomas Hain, Phil Woodland, Gunnar Evermann, Mark Gales, Andrew Liu, Gareth Moore, and Lan Wang. Automatic transcription of conversational telephone speech-development of the CU-HTK 2002 system. In *IEEE Transactions on Acoustics, Speech and Signal Processing*. Citeseer, 2003.

Trevor Hastie and Robert Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176, 1996.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Science + Business Media, 2 edition, 2009.

Xiaodong He, Li Deng, and Wu Chou. Discriminative learning in sequential pattern recognition. *Signal Processing Magazine, IEEE*, 25(5):14–36, 2008.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. *Advances in neural information processing systems*, pages 266–272, 1998.

Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

A. Jasra, CC Holmes, and DA Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

A. Jasra, D.A. Stephens, and C.C. Holmes. Population-based reversible jump Markov chain Monte Carlo. *Biometrika*, 94(4):787–807, 2007.

Shihao Ji, Balaji Krishnapuram, and Lawrence Carin. Variational Bayes for continuous hidden Markov models and its application to active learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):522–532, 2006.

Hui Jiang. Discriminative training of HMMs for automatic speech recognition: A survey. *Computer Speech & Language*, 24(4):589–608, 2010.

Adam Kalai and Rocco A Servedio. Boosting in the presence of noise. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 195–205. ACM, 2003.

A. Komárek. A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Computational Statistics & Data Analysis*, 53(12):3932–3947, 2009.

Sarah Kozey, Kate Lyden, John Staudenmayer, and Patty Freedson. Errors in met estimates of physical activities using 3.5 ml. kg-1min-1 as the baseline oxygen consumption. *Journal of physical activity & health*, 7(4):508, 2010.

Nagendra Kumar and Andreas G Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech communication*, 26(4):283–297, 1998.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, 2001.

Michael Lavine and Mike West. A Bayesian method for classification and discrimination. *Canadian Journal of Statistics*, 20(4):451–461, 1992.

Brian G. Leroux and Martin L. Puterman. Maximum-penalized-likelihood estimation for independent and Markov- dependent mixture models. *Biometrics*, 48(2):pp. 545–558, 1992. ISSN 0006341X. URL `http://www.jstor.org/stable/2532308`.

Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello, and Blake Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *IJCAI*, volume 5, pages 766–772, 2005.

Jonathan Lester, Carl Hartung, Laura Pina, Ryan Libby, Gaetano Borriello, and Glen Duncan. Validated caloric expenditure estimation using a single body-worn sensor. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 225–234. ACM, 2009.

Lin Liao, Tanzeem Choudhury, Dieter Fox, and Henry A Kautz. Training conditional random fields using virtual evidence boosting. In *IJCAI*, volume 7, pages 2530–2535, 2007.

Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Jane Liu and Mike West. *Combined parameter and state estimation in simulation-based filtering*. Springer, 2001.

Jun S. Liu, Junni L. Zhang, Michael J. Palumbo, and Charles E. Lawrence. Bayesian clustering with variable and transformation selections. In JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D. Heckerman, AFM Smith, and M. West, editors, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, page 249. Oxford University Press, USA, 2003.

Phil Long and Rocco Servedio. Adaptive martingale boosting. In *Advances in Neural Information Processing Systems*, pages 977–984, 2008.

Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010.

Kate Lyden, Sarah L Kozey, John W Staudenmeyer, and Patty S Freedson. A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *European journal of applied physiology*, 111(2):187–201, 2011.

Kate Lyden, Sarah L Kozey Keadle, John W Staudenmeyer, and Patty S Freedson. A method to estimate free-living active and sedentary behavior from an accelerometer. *Publication Forthcoming*, 2014.

Andrea Mannini and Angelo Maria Sabatini. Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*, 10(2):1154–1175, 2010.

Andrea Mannini, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. Activity recognition using a single accelerometer placed at the wrist or ankle. *Medicine and science in sports and exercise*, 2013. doi: 10.1249/MSS.0b013e31829736d6.

J.M. Marin, K. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.

MJ Mathie, Branko G Celler, Nigel H Lovell, and ACF Coster. Classification of basic daily movements using a triaxial accelerometer. *Medical and Biological Engineering and Computing*, 42(5):679–687, 2004.

Andrew McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc., 2003.

Geoffrey J McLachlan, David Peel, and RW Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3):379–388, 2003.

Blakeley B McShane, Shane T Jensen, Allan I Pack, and Abraham J Wyner. Statistical learning with time series dependence: An application to scoring sleep in mice. *Journal of the American Statistical Association*, 2013.

Arthur Nádas. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(4):814–817, 1983.

Arthur Nádas, David Nahamoo, and Michael A Picheny. On a model-robust training method for speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(9):1432–1436, 1988.

Radford M Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press, 2002.

Peder A Olsen and Ramesh A Gopinath. Modeling inverse covariance matrices by basis expansion. *Speech and Audio Processing, IEEE Transactions on*, 12(1):37–46, 2004.

Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8:1145–1164, 2007.

David M Pober, John Staudenmayer, Christopher Raphael, and Patty S Freedson. Development of novel techniques to classify physical activity mode using accelerometers. *Medicine and science in sports and exercise*, 38(9):1626, 2006.

Stephen J Preece, John Y Goulermas, Laurence PJ Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. Activity identification using body-mounted sensors: a review of classification techniques. *Physiological measurement*, 30(4):R1, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

Rajat Raina, Yirong Shen, Andrew Mccallum, and Andrew Y Ng. Classification with hybrid generative/discriminative models. In *Advances in neural information processing systems*, number 16, pages 545–552, 2004.

Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. Activity recognition from accelerometer data. In *AAAI*, pages 1541–1546, 2005.

Nalini Ravishanker and Dipak Kumar Dey. *A first course in linear model theory*. CRC Press, 2002.

Sylvia Richardson and Peter J Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

C.P. Robert, T. Ryden, and D.M. Titterington. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):57–75, 2000.

Hans Rosdahl, Lennart Gullstrand, Jane Salier-Eriksson, Patrik Johansson, and Peter Schantz. Evaluation of the oxycon mobile metabolic system against the douglas bag method. *European journal of applied physiology*, 109(2):159–171, 2010.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973, 2004.

Antti-Veikko Ilmari Rosti. *Linear Gaussian Models for Speech Recognition*. PhD thesis, Wolfson College, University of Cambridge, May 2004.

Megan P Rothney, Megan Neumann, Ashley Béziat, and Kong Y Chen. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *Journal of applied physiology*, 103(4):1419–1427, 2007.

Y Dan Rubinstein. *Discriminative vs informative learning*. PhD thesis, Stanford University, 1998.

Tobias Rydén, Timo Teräsvirta, and Stefan Åsbrink. Stylized facts of daily return series and the hidden Markov model. *Journal of applied econometrics*, 13(3):217–244, 1998.

Jeffer Eidi Sasaki. *Development and Validation of Accelerometer-Based Activity Classification Algorithms for Older Adults: A Machine-Learning Approach*. PhD thesis, University of Massachusetts, Amherst, September 2013.

Lawrence K Saul and Mazin G Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 8(2):115–125, 2000.

Robert E Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

Steven L Scott. Bayesian analysis of a two-state Markov modulated Poisson process. *Journal of Computational and Graphical Statistics*, 8(3):662–670, 1999.

Steven L Scott. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, 97(457):337–351, 2002. doi: 10.1198/016214502753479464. URL `http://www.tandfonline.com/doi/abs/10.1198/016214502753479464`.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *The Journal of Machine Learning Research*, 12:1865–1892, 2011.

Robert H Shumway. *Applied statistical time series analysis*. Prentice Hall Series in Statistics, Englewood Cliffs, NJ, 1988.

Andrew Smith. *Logarithmic Opinion Pools for Conditional Random Fields*. PhD thesis, University of Edinburgh, 2007.

Andrew Smith and Miles Osborne. Diversity in logarithmic opinion pools. *Lingvisticae Investigationes*, 30(1):27–47, 2007.

Andrew Smith, Trevor Cohn, and Miles Osborne. Logarithmic opinion pools for conditional random fields. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 18–25. Association for Computational Linguistics, 2005.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

L. Spezia. Reversible jump and the label switching problem in hidden Markov models. *Journal of Statistical Planning and Inference*, 139(7):2305–2315, 2009.

L. Spezia. Bayesian analysis of multivariate Gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis*, 31(1):1–11, 2010.

John Staudenmayer, David Pober, Scott Crouter, David Bassett, and Patty Freedson. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, 107(4):1300–1307, 2009.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000a.

Matthew Stephens. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics*, pages 40–74, 2000b.

Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Machine Learning*, 4(4):267–373, 2011.

Charles Sutton, Michael Sindelar, and Andrew McCallum. Reducing weight undertraining in structured discriminative learning. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 89–95. Association for Computational Linguistics, 2006.

Kazem Taghva, Jeffrey S Coombs, Ray Pereda, and Thomas A Nartker. Address extraction using hidden Markov models. In *Electronic Imaging 2005*, pages 119–126. International Society for Optics and Photonics, 2005.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin Markov networks. *Advances in neural information processing systems*, 16:25, 2004.

Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

Antonio Torralba, Kevin P Murphy, and William T Freeman. Contextual models for object detection using boosted random fields. In *Advances in neural information processing systems*, pages 1401–1408, 2004.

Stewart G Trost, Weng-Keen Wong, Karen A Pfeiffer, and Yonglei Zheng. Artificial neural networks to predict activity type and energy expenditure in youth. *Medicine and science in sports and exercise*, 44(9):1801–1809, 2012.

U.S. Department of Health and Human Services. 2008 Physical Activity Guidelines for Americans, 2008. URL `http://www.health.gov/paguidelines`.

Douglas L Vail, John D Lafferty, and Manuela M Veloso. Feature selection in conditional random fields for activity recognition. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 3379–3384. IEEE, 2007a.

Douglas L Vail, Manuela M Veloso, and John D Lafferty. Conditional random fields for activity recognition. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 235. ACM, 2007b.

Vincent Vanhoucke and Ananth Sankar. Mixtures of inverse covariances. *Speech and Audio Processing, IEEE Transactions on*, 12(3):250–264, 2004.

La The Vinh, Sungyoung Lee, Hung Xuan Le, Hung Quoc Ngo, Hyoung Il Kim, Manhyung Han, and Young-Koo Lee. Semi-Markov conditional random fields for accelerometer-based activity recognition. *Applied Intelligence*, 35(2):226–241, 2011.

Sijian Wang and Ji Zhu. Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448, 2008.

Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognition Letters*, 25(7):767–775, 2004.

Jing-Hao Xue and D Michael Titterington. Comment on on discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Neural processing letters*, 28(3):169–187, 2008.

In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

Liang-Guo Zhang, Xilin Chen, Chunli Wang, Yiqiang Chen, and Wen Gao. Recognition of sign language subwords based on boosted hidden Markov models. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 282–287. ACM, 2005.

Shaoyan Zhang, Alex V Rowlands, Peter Murray, and Tina L Hurst. Physical activity classification using the genea wrist-worn accelerometer. *Medicine and science in sports and exercise*, 44(4):742–748, 2012.

Zhihua Zhang, Kap Luk Chan, Yiming Wu, and Chibiao Chen. Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. *Statistics and Computing*, 14(4):343–355, 2004.

Yonglei Zheng, Weng-Keen Wong, Xinze Guan, and Stewart Trost. Physical activity recognition from accelerometer data using a multi-scale ensemble method. In *Twenty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence. IAAI*, 2013.

Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. Multi-class AdaBoost. *Statistics and Its*, 2009.

Walter Zucchini and Iain L MacDonald. *Hidden Markov models for time series: an introduction using R*. CRC Press, 2009.