Biometrics 64, 1–24 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Physical Activity Classification with Dynamic, Discriminative Methods

Evan L. Ray[1,*], Jeffer E. Sasaki[2], Patty S. Freedson[3], and John Staudenmayer[4]

[1]Department of Biostatistics and Epidemiology, University of Massachusetts,
Amherst, Massachusetts, U.S.A.

[2]Graduate Program in Physical Education, Universidade Federal do Triangulo Mineiro,
Uberaba, Minas Gerais, Brazil

[3]Department of Kinesiology, University of Massachusetts,
Amherst, Massachusetts, U.S.A.

[1]Department of Mathematics and Statistics, University of Massachusetts,
Amherst, Massachusetts, U.S.A.

email: elray@umass.edu

Summary: A person's physical activity has important health implications, so it is important to be able to measure aspects of physical activity objectively. One approach to doing that is to use data from an accelerometer to classify physical activity according to activity type (e.g., lying down, sitting, standing, or walking) or intensity (e.g., sedentary, light, moderate, or vigorous). This can be formulated as a labeled classification problem, where the model relates a feature vector summarizing the accelerometer signal in a window of time to the activity type or intensity in that window. These data exhibit two key characteristics: (1) the activity classes in different time windows are not independent, and (2) the accelerometer features have moderately high dimension and follow complex distributions. Through a simulation study and applications to three data sets, we demonstrate that a model's classification performance is related to how it addresses these aspects of the data. Dynamic methods that account for temporal dependence achieve better performance than static methods that do not. Generative methods that explicitly model the distribution of the accelerometer signal features do not perform as well as methods that take a discriminative approach to establishing the relationship between the accelerometer signal and the activity class. Specifically, Conditional Random Fields consistently have better performance than commonly employed methods that ignore temporal dependence or attempt to model the accelerometer features.

Key words: Accelerometers; Classification; Conditional Random Field; Hidden Markov Model; Physical Activity.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

The United States Department of Health and Human Services published the 2008 Physical Activity Guidelines (U.S. Department of Health and Human Services, 2008) recommending that adults accumulate at least 2 hours and 30 minutes of moderate intensity physical activity each week and that this activity should occur in continuous bouts of at least 10 minutes. These recommendations are based on a large literature review which found that increased physical activity leads to a reduction in all-cause mortality risk, weight-loss, prevention of certain types of cancer, and improvements in cardiorespiratory, metabolic, musculoskeletal, functional and mental health. To understand the dose response relationships between physical activity and aspects of health more precisely and to assess the effects of interventions to increase physical activity, it is important to be able to accurately measure physical activity.

One approach to the objective measurement of physical activity is through the use of an accelerometer worn by the individual. The accelerometer records the acceleration that it experiences in each of three axes at a high frequency; measurements were recorded at frequencies of 80 to 90 Hz in the data sets we work with in this article. These acceleration recordings do not directly measure the quantities of interest. Instead, statistical models must be used to infer descriptions of physical activity type or intensity from the accelerometer signal.

A number of methods to estimate physical activity type or intensity from accelerometer data have been developed, and the vast majority of these methods proceed by dividing time up into non-overlapping windows and extracting a vector of features summarizing the accelerometer signal in each window. A classification model is then developed to relate this feature vector to the activity type or intensity in each window. The methods are developed from training data where the accelerometer signals and the true activity types or intensities

are observed. The models are then used for prediction when only the accelerometer signals are observed.

The purpose of this article is to use both real data and simulations to investigate if there are general characteristics of the classification methods that associate with superior performance. We focus on two characteristics: (1) whether the method accounts for temporal dependence in the activity class, and (2) whether the method is based on a model for the accelerometer signal features or not. We refer to methods that do not account for temporal dependence as static and others as dynamic. We refer to methods that are based on models for the features as generative and others as discriminative. In general, we find that a dynamic and discriminative approach leads to superior performance. While this is consistent with the dynamic nature of physical activity and the fact that summaries of the accelerometer signals tend to have complex and high dimensional distributions, most efforts in the literature have been devoted to static and generative models.

The rest of this article is organized as follows. In Section 2, we briefly review the existing literature on estimating physical activity from accelerometer data. In Section 3 we develop static, dynamic, generative, and discriminative approaches to classification, and we apply the methods to simulated and real physical activity data sets in Sections 4 and 5. We conclude with a discussion in Section 6.

2. Literature Review

In this Section we present some necessary scientific background and review relevant literature on estimation of physical activity from accelerometer data. In general, two aspects of physical activity are estimated from accelerometers: energy expenditure (a measure of exercise intensity) and activity type (what the person is doing). We discuss methods to estimate energy expenditure first. An example of accelerometer data is displayed in Figure 1.

[Figure 1 about here.]

The Metabolic Equivalent of Task (MET) is a body-size independent measure of energy expenditure. One MET is the energy used by an individual at rest, and the energy expenditure of other activities is expressed as multiples of this resting rate (Ainsworth et al., 2011). For ease of interpretation, METs are often discretized into four levels of energy expenditure that define categories of exercise intensity: Sedentary ($\leqslant 1.5$ METs), Light (> 1.5 and < 3 METs), Moderate ($\geqslant 3$ and < 6 METs), and Vigorous ($\geqslant 6$ METs).

Simple linear regression is the most prevalent way to estimate METs from an accelerometer. This approach relates a univariate summary of the total acceleration magnitude in non-overlapping time intervals (referred to as *counts* and often considered per minute window) to METs (e.g. Freedson et al., 1998). While the relationship between counts and METs is approximately linear during locomotion, METs are not a mathematical function of counts when the wearer of the accelerometer does a variety of activities. For instance, activities with similar intensities may have quite different counts, and activities with different intensities may have similar counts (Staudenmayer et al., 2009).

These problems can be partially remedied by using a richer summary of the acceleration signal and a more flexible regression model for the relationship between the accelerometer signal and activity intensity (e.g. Rothney et al., 2007; Staudenmayer et al., 2009). Another common option is to use separate models for different types of activity (Crouter et al., 2006; Lyden et al., 2014; Bonomi et al., 2009; Albinali et al., 2010; Lester et al., 2009). We note that these methods are all static in the sense that regression models are applied to different windows independently.

At the expense of not estimating total energy expenditure, it is also possible to bypass the initial regression step and classify physical activity intensity directly. If this strategy is used, the same modeling tools can be used to classify activity type (what the person is

actually doing) or intensity. Many static and discriminative classification methods have been applied to these problems, including support vector machines (Anderson, 2013; Gyllensten and Bonomi, 2011; Mannini et al., 2013; Ravi et al., 2005; Zhang et al., 2012; Zheng et al., 2013), classification trees (Albinali et al., 2010; Anderson, 2013; Bonomi et al., 2009; Bao and Intille, 2004; Bonomi et al., 2009; 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; 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.

Previous papers have suggested that models that account for temporal dependence might have better classification accuracy than models that do not (e.g. Gyllensten and Bonomi, 2011; Mannini and Sabatini, 2010; Bao and Intille, 2004). To our knowledge no previous study has directly examined the impact of this characteristic of the model on classification performance with real physical activity data.

Hidden Markov models (HMMs) are generative dynamic models that provide one way to model temporal dependence. A straightforward way to use HMMs for physical activity data is to represent the true activity class by the hidden state, which is modeled as changing over time according to a Markov process. The observed acceleration features follow a distribution that depends on the state. This setup was used by Mannini and Sabatini (2010). Pober et al. (2006) used the same general idea but employed a HMM that assigned three hidden states to each activity class.

The difficulty with this approach is that it requires us to estimate the distribution of the accelerometer features associated with each hidden state. Applied studies and theoretical results have shown that classification performance of generative models such as HMMs suffers

when the model is badly misspecified, and discriminative approaches may be preferred in these cases (e.g. Ng and Jordan, 2002; Nádas et al., 1988; Xue and Titterington, 2008).

A second approach to using HMMs is to first use a discriminative model that does not incorporate temporal dependence, such as a support vector machine or random forest, to obtain an initial classification, and then use a HMM to smooth those initial classifications over time. Variations on this idea have been used by Lester et al. (2005), Anderson (2013) and Ellis et al. (2014). McShane et al. (2013) developed a more formally justified variation on this theme in the context of classifying sleep type in mice with video recordings. Their work re-expresses the HMM in terms of the class membership probabilities that are obtained from a static non-parametric classification model. This approach combines the benefits of using a discriminative approach for relating the feature vectors to the activity classes with the temporal dependence structure of the HMM.

The Conditional Random Field (CRF) is another discriminative approach that can capture temporal dependence. The CRF was originally proposed in the computer science literature by Lafferty et al. (2001) and has been applied to a variety of problems, including part-of-speech tagging (Lafferty et al., 2001) and gene prediction (Bernal et al., 2007) among many others. Two previous studies have applied CRFs to classification of physical activity with accelerometer data, although their specific classification tasks were fairly different from the tasks that are of interest to public health researchers studying the effectiveness of interventions designed to increase physical activity levels and the links between physical activity and health (Vinh et al., 2011; Adams et al., 2016). The CRF model uses a graphical structure to represent the conditional independence relationships among the activity classes at different times. It can be shown that the model that results from conditioning on the observed features in the HMM is a special case of the CRF (Lafferty et al., 2001; Sutton and McCallum, 2011).

3. Classification Methods

First, we introduce notation. We denote the acceleration feature vector in window t for subject i by $\mathbf{X}_{i,t} \in \mathbb{R}^D$, and the activity type or intensity by $Y_{i,t} \in \{1, \ldots, S\}$. Here S is the total number of activity type or intensity levels, which varies with the data set. We let N denote the total number of subjects and T_i denote the number of windows for subject i. In our applications in Section 5, we use either $D = 13$ or $D = 77$ features depending on the data set; we list the features used in Section 5. Discussion of the advantages and disadvantages of specific features is outside the scope of this article.

Our first four classification methods are a finite mixture model (**FMM**), hidden Markov model (**HMM**), multinomial logistic regression (**MLR**), and a conditional random field (**CRF**). The specifications of these models that we use are closely related to each other, and together they cover all four combinations of the static/dynamic and generative/discriminative model characteristics. The **FMM** is a static generative model; the **HMM** is dynamic generative model that can be obtained by adding temporal dependence to the **FMM**; **MLR** is a static discriminative model that can be obtained from a restricted version of the **FMM** by conditioning on the observed accelerometer features rather than modeling their distribution; and the **CRF** is a dynamic discriminative model that can be obtained either by adding temporal dependence to **MLR** or by conditioning on the accelerometer features in a restricted version of the **HMM**.

In more detail, the **FMM** is specified as follows:

$$P(Y_{i,t} = s; \boldsymbol{\pi}) = \pi_s, s \in \{1, \ldots, S\}, 0 \leqslant \pi_s \leqslant 1, \sum_{s=1}^{S} \pi_s = 1 \quad (1)$$

$$f(\mathbf{x}_{i,t} | Y_{i,t} = s; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \sum_{k=1}^{K_s} w_{s,k} g(\mathbf{x}_{i,t}; \boldsymbol{\mu}_{s,k}, \Sigma_{s,k}), 0 \leqslant w_{s,k} \leqslant 1, \sum_{k=1}^{K_s} w_{s,k} = 1 \, \forall s. \quad (2)$$

Here, $g(\cdot)$ is the probability density function of the multivariate normal distribution. This is a static model because the class membership probabilities at time t depend only on the

observed features at that time. Additionally, this is a generative model because it models the distribution for the observed features associated with activity type s with a mixture of K_s normals.

The second model is a first-order **HMM** with one state for each activity class:

$$P(Y_{i,1} = s; \boldsymbol{\pi}) = \pi_s, s \in \{1, \ldots, S\}, 0 \leqslant \pi_s \leqslant 1, \sum_{s=1}^{S} \pi_s = 1 \quad (3)$$

$$P(Y_{i,t} = s | Y_{i,t-1} = r; Q) = q_{r,s},\ r, s \in \{1, \ldots, S\}, 0 \leqslant q_{r,s} \leqslant 1\, \forall r, s, \sum_{s=1}^{S} q_{r,s} = 1\, \forall r \quad (4)$$

$$f(\mathbf{x}_{i,t} | Y_{i,t} = s; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \sum_{k=1}^{K_s} w_{s,k} g(\mathbf{x}_{i,t}; \boldsymbol{\mu}_{s,k}, \Sigma_{s,k}), 0 \leqslant w_{s,k} \leqslant 1, \sum_{k=1}^{K_s} w_{s,k} = 1\, \forall s. \quad (5)$$

This model introduces dependence in the activity class at nearby time points through the transition probabilities in Equation (4), and is therefore a dynamic model. It maintains the generative model for the accelerometer features as a mixture of normals that was used in the **FMM**.

The **MLR** model takes a discriminative approach and directly models the conditional distribution of the activity class given the accelerometer features:

$$P(Y_{i,t} = y_{i,t} | \mathbf{X}_{i,t} = \mathbf{x}_{i,t}; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_{i,t}; \boldsymbol{\theta})} \exp \left\{ \sum_{s=1}^{S} \mathbf{I}_{\{s\}}(y_{i,t}) \left(\beta_{s,0} + \sum_{d=1}^{D} \beta_{s,d} x_{i,t,d} \right) \right\}. \quad (6)$$

Here, $Z(\mathbf{x}_{i,t}; \boldsymbol{\theta})$ is a normalizing factor ensuring that the distribution sums to 1 and $\mathbf{I}_A(x)$ is the indicator function, taking the value 1 if $x \in A$ and 0 otherwise. This is a static model since the distribution of the activity class at time t depends only on quantities observed at that time. It can be shown that this **MLR** specification can be obtained by conditioning on $\mathbf{X}_{i,t}$ in a **FMM** where the number of mixture components associated with each activity class, K_s, is fixed equal to 1 (Efron (1975)).

The fourth classification method is a linear chain **CRF**. This model is specified as follows:

$$P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}_i; \boldsymbol{\theta})} \exp \left\{ \sum_{s=1}^{S} \mathbf{I}_{\{s\}}(y_{i,1}) \zeta_s \right.$$

$$+\sum_{t=2}^{T_i}\sum_{r=1}^{S}\sum_{s=1}^{S}\mathbf{I}_{\{r\}}(y_{i,t-1})\mathbf{I}_{\{s\}}(y_{i,t})\omega_{r,s}$$
$$+\sum_{t=1}^{T_i}\sum_{s=1}^{S}\mathbf{I}_{\{s\}}(y_{i,t})\left(\beta_{s,0}+\sum_{d=1}^{D}\beta_{s,d}x_{i,t,d}\right)\Bigg\}. \quad (7)$$

Again, $Z(\mathbf{x}_i;\boldsymbol{\theta})$ is a normalizing factor ensuring that the distribution sums to 1. This model differs from **MLR** in that it specifies a conditional distribution for the entire sequence of activity classes observed over time for a given subject given the accelerometer features at all times. This distribution can be marginalized to obtain distributions for activity classes at individual time points. The first two terms in Equation (7) allow the model to capture how likely a subject is to begin in each activity class at time $t = 1$, and how likely transitions between different activity types are over the course of the remaining observed times. The third term has a similar form to the **MLR** specification, but incorporates contributions from all time points. It can be shown that this **CRF** specification arises if we condition on $\mathbf{X}_i$ in the joint model for $(\mathbf{X}_i, \mathbf{Y}_i)$ specified by the **HMM** above if K_s is fixed to 1 for all s (Sutton and McCallum (2011)).

Our final classification method is a random forest (**RF**, Breiman (2001)). This is a commonly used static discriminative method which is more flexible than the **MLR** model outlined above. We use the implementation in the `randomForest` package (Liaw and Wiener, 2002) for `R` (R Core Team, 2016) with the default options for number of trees, node size, and number of variables considered for each split.

We employ similar estimation strategies for the two generative models and for the two discriminative models. The **HMM** has parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, Q, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)$, $Q = [q_{r,s}]$, and $\mathbf{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ contain the parameters for all mixture components; the FMM has similar parameters, but does not include the transition matrix Q. For the **HMM**, we estimate Q via maximum likelihood. We estimate $\boldsymbol{\pi}_s$ as the observed proportion of the sample with $y_{i,t} = s$. This is the maximum likelihood estimate for the **FMM**; it is not

the maximum likelihood estimate for the **HMM**, but use of this estimate is a standard procedure to reduce sampling variance of the estimates (e.g., McShane et al. (2013)). For the observation distributions, we use `R`'s `mclust` package (Fraley et al., 2012) to estimate the mixture weights $\mathbf{w}$ and the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the normal mixture components. This package allows for a number of possible restrictions on the parameterizations of the normal component covariance matrices. It uses an EM algorithm to obtain local maximum likelihood estimates of the normal component parameters, and BIC to select the covariance parameterization and number of components. Before fitting the model, we apply the Yeo-Johnson transformation (Yeo and Johnson, 2000) to each covariate to approximate normality. We use the implementation of this transformation that is available in the `car` package (Fox and Weisberg, 2011) for `R`.

Our estimation algorithm for the **MLR** and **CRF** models employs 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 (note that this same technique is used in the estimation of random forests). In the final model fit, the coefficient estimates ζ_s, $\omega_{r,s}$, and $\beta_{s,d}$ are the average of the coefficient estimates obtained from separate model fits to each of these bagged data sets. We use a boosting procedure to obtain these separate model fits. The boosting step can be interpreted as a random block coordinate ascent algorithm converging to the maximum likelihood parameter estimates based on the given training data set, with early stopping used to reduce overfitting. We use estimates of classification performance on the out-of-bag observation sequences to select the stopping point for the boosting procedure. The precise estimation procedure is given in Algorithm 1. Similar estimation strategies for CRFs have been employed previously (e.g., Smith and Osborne (2007); Dietterich et al. (2004)). In order to resolve problems with identifiability, we fix $\zeta_S = 0$, $\omega_{S,S} = 0$, and $\beta_{S,d'} = 0 \, \forall d' = 0, \ldots, D$.

4. Simulation Study

The objective of the simulation study we describe in this Section is to understand how the performance of the classification methods outlined in Section 3 depends on two factors: (1) dependence in activity classes at nearby time points and (2) the complexity of the distributions for the feature vectors derived from the accelerometer data in each window. 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, 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 these two factors because we believe they are the most useful in helping to explain differences in the performance of dynamic/static and discriminative/generative classification methods when applied to real physical activity data.

In order to study this, we generate data from one of four distributions, varying whether or not there is temporal dependence in the simulated activity class and whether the observed data follow a relatively simple distribution or a more complex distribution. For each combination of these factors, 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 $D = 50$.

In the cases without time dependence, 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 FMM as follows:

$$\begin{aligned} p(Y_{i,1} = s|\boldsymbol{\pi}) &= \frac{1}{3}, \\ f(\mathbf{x}_{i,t}|Y_{i,t} = s) &= \sum_{m=1}^{M_s} w_{s,m} f_{s,m}(\mathbf{x}_{i,t}; \boldsymbol{\theta}_{s,m}), 0 \leqslant w_{s,m} \leqslant 1 \, \forall m, \sum_{m=1}^{M_s} w_{s,m} = 1. \end{aligned}$$

In cases with time dependence, the data are generated from a first-order HMM:

$$\begin{aligned} p(Y_{i,1} = s|\boldsymbol{\pi}) &= \frac{1}{3}, \\ p(Y_{i,t+1} = y_{i,t+1}|Y_{i,1:t} = y_{i,1:t}; Q) &= p(Y_{i,t+1} = y_{i,t+1}|Y_{i,t} = y_{i,t}; Q) \\ &= \frac{4}{5} \text{ if } y_{i,t} = y_{i,t+1} \text{ and } \frac{1}{10} \text{ otherwise,} \\ f(\mathbf{x}_{i,t}|Y_{i,t} = s) &= \sum_{m=1}^{M_s} w_{s,m} f_{s,m}(\mathbf{x}_{i,t}; \boldsymbol{\theta}_{s,m}), 0 \leqslant w_{s,m} \leqslant 1 \, \forall m, \sum_{m=1}^{M_s} w_{s,m} = 1. \end{aligned}$$

In both cases, 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. In the cases with simple emission distributions, we use a mixture of normal distributions. Thus, in those cases the data are generated from either the **FMM** or the **HMM** that is used as one of the classification methods. In the cases with more complex emission distributions, each mixture component is 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 summarized performance of the classification methods in each trial of the simulation study with two statistics: the proportion of time windows classified correctly and the macro F_1 score (Sokolova and Lapalme, 2009). For the sake of brevity, we have only included plots of the proportion correct here. The qualitative story is similar when we consider the macro F_1 score; those results are deferred to the Supplementary Materials.

Figure 2 summarizes the results of the simulation study. In the case with simple emission distributions and no time dependence, the **FMM** and the **HMM** have the best performance. This is expected, as the **FMM** is the true data generating model and the **HMM** model can be viewed as an overparameterized version of the data generating model, so that restricting each row of Q to be equal to $\boldsymbol{\pi}$ yields the data generating model. Introducing time dependence, the **HMM** is now the data generating model and the method with the best performance. However, the **CRF** is also able to make use of the information provided by this dependence, and it offers better performance than **MLR** in this case. In the cases with complex emission

distributions, where the generative **FMM** and **HMM** models are misspecified, those models do not offer any advantages over the discriminative approaches. In the setting that most closely resembles our real data, with complex emission distributions and time dependence, the dynamic and discriminative **CRF** has the best performance. The static **RF** method is consistently outperformed by the dynamic **HMM** and **CRF** models in both settings when the data generating process includes time dependence. The relative performance of the methods is the same when measured by the macro F_1 score (Supplemental Figure 1).

[Figure 2 about here.]

5. Applications

In this Section, we present classification results for three physical activity data sets. We describe the data collection procedures in Subsection 5.1 and present the classification results in Subsection 5.2.

5.1 *Data Collection*

Our three data sets were collected in two studies. The first was described and analyzed in Mannini et al. (2013) and the last two in Sasaki et al. (2016). Table 1 contains descriptive statistics for the study participants.

[Table 1 about here.]

For the first data set, 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 the order and duration of activities were determined by the researchers. While subjects performed these activities, they wore Wocket accelerometers on their ankle, thigh, wrist, and hip. The accelerometers recorded acceleration in each of 3 orthogonal axes at a frequency of 90 Hz. In this work, we analyze only the data from the ankle and the wrist;

these are the data that were previously discussed in Mannini et al. (2013). In that paper, researchers developed static discriminative statistical learning models (specifically, support vector machines [SVM]) to classify activities into one of several groups of similar activity types and to classify intensity. Classification performance was evaluated relative to the actual activity type by cross-validation.

Our second and third data sets were collected using healthy elderly subjects and used the ActiGraph GT3X+ accelerometer (3 axes at a frequency of 80 Hz) to measure acceleration. The second dataset was was collected using a laboratory protocol that was similar to the one used in Mannini et al. (2013), and the third dataset was conducted under free-living conditions. Staff followed the participants as they went about their normal activities and recorded what was done and when in terms of the type of activity performed and a categorical assessment of the intensity of the activity (Sedentary, Light, Moderate, or Vigorous). Similar to Mannini et al. (2013), Sasaki et al. (2016) used static discriminative statistical learning methods (RF and SVM) to classify activity and type, and evaluated performance using leave-one-subject-out cross-validation.

For all datasets, we summarized the accelerometer signals using non-overlapping 12.8 second windows as was done in Mannini et al. (2013). For each window, we computed a vector of statistics to summarize time and frequency domain features of the accelerometers signals. For the data from Mannini et al. (2013), we used the vector of 13 features in the recommended feature set from that work. For the data sets from Sasaki et al. (2016), we used a vector of 77 features; these statistics are similar to those used by Mannini et al. (2013), Sasaki et al. (2016) and others. Tables 2 and 3 list the accelerometer features used for the data from Mannini et al. (2013) and Sasaki et al. (2016) respectively.

[Table 2 about here.]

[Table 3 about here.]

In addition to summarizing acceleration in each window, we also assigned an activity type label and intensity for each window. For the first data set, we use the same system for classifying activity type as Mannini et al. (2013), with four categories: sedentary, locomotion, cycling, and non-locomotion movement. For the other two data sets, we used four categories: sedentary, locomotion, non-locomotion movement, and transition. The transition category occurred when a window included more than one activity. Activity intensity was represented by four categories differentiated on the basis of energy expenditure (METs): Sedentary ($\leqslant 1.5$ METs), Light (> 1.5 and < 3 METs), Moderate ($\geqslant 3$ and < 6 METs), and Vigorous ($\geqslant 6$ METs). We also used a Transition category for activity intensity in applications to the data from Sasaki et al. (2016). Intensity was measured using indirect calorimetry (e.g. (Kozey et al., 2010)) in the second data set. For the other datasets, we assigned an intensity category to each time point using the MET classification of activities in the Compendium of Physical Activity (Ainsworth et al. (2011)). The prevalence of different levels of activity type and intensity for each dataset are in Table 4.

[Table 4 about here.]

5.2 *Results*

Figure 3 summarizes the results from the activity type and intensity classification tasks. The left panel shows average percent correct for activity intensity classifications and the right are for activity type classifications. The reported means and confidence intervals are obtained from a mixed effects model and represent average performance over the three studies and the two accelerometer locations per study. All were estimated using leave-one-subject-out cross-validation. For the activity type classifications, the results for the **RF** method here are very similar to the results from a support vector machine (**SVM**) reported in Mannini et al. (2013); those authors did not attempt to classify activity intensity. The published results from Sasaki et al. (2016) are not comparable to our results because of differences in how the

data were preprocessed; however, the **RF** presented here is similar to the **RF** used in that article.

[Figure 3 about here.]

The figure indicates that the discriminative dynamic model (**CRF**) offers the best performance overall, and the generative static model (**FMM**) offers the worst performance. The three other approaches, which are either discriminative static models (**MLR** and **RF**) or a generative dynamic model (**HMM**) offer performance in the middle of this range. The improvement in performance of the **CRF** relative to the other methods is statistically significant for all cases other than the **MLR** and **RF** methods in the application to classification of activity intensity (tests conducted simultaneously with construction of the confidence intervals in Figure 3, family-wise significance level of 0.05). Table 5 contains the average percent correct for each response, location, data set, and model. The **CRF** achieved the highest performance level of any model in nine out of twelve classification tasks, and was competitive in the remaining three tasks. The **FMM** had the worst performance in every case but one. Summarizing classification performance with the F_1 score did not change these general trends, though the difference between mean F_1 scores achieved by the discriminative and generative methods was not statistically significant (see Supplement).

[Table 5 about here.]

6. Discussion

In this work, we have considered two general characteristics of methods that can be used to classify physical activity type or intensity with accelerometer data: (1) whether or not they handle temporal dependence in activity class, and (2) whether they take a generative or discriminative approach. Through a simulation study and applications to three data sets, we have demonstrated that using a dynamic, discriminative approach can yield consistent

gains in the proportion correct relative to static or generative methods. The argument in favor of dynamic, discriminative models is that they are a better representation of the data than static or generative methods are. Dynamic models are superior to static models because they capture a key feature of the data: activity types at nearby times are dependent. Discriminative methods are superior to generative methods because they do not require a specification for the distribution of the feature vector; this distribution is too complex to model well with simple parametric approaches, and is too high-dimensional to be handled easily by non-parametric methods. The complexity of the accelerometer features is illustrated in Figure 4.

[Figure 4 about here.]

We have discussed just these two characteristics of the classification method in this article, but there are other aspects of the modeling that we think may be fruitful to press in the future. For instance, 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 combining inferences from multiple CRFs operating at different time scales. It may also be beneficial to explore the use of new features. Another option would be to abandon the windowing approach altogether and model the accelerometer signal directly. An example of one approach to doing this is in Bai et al. (2012).

Another alternative 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. With this approach, generative approaches might be more successful than they are when a high-dimensional feature vector is used.

More flexible models than those used in this article can be formulated. We have selected these model specifications because of the connections between them outlined above; specifically, each discriminative model can be obtained from a corresponding generative model by conditioning on the accelerometer features, and each dynamic model can be obtained from a corresponding static model by adding a simple model for time dependence. This allows us to compare the relative benefits of using generative or discriminative approaches and static or dynamic approaches in a controlled manner. However, improvements on each of these models could be made.

The dynamic models we discussed in this article employed very simple first-order Markov time dependence structures. These structures could be expanded to capture more complex dependence in the activity type and intensity. McShane et al. (2013) explored several avenues for this using a discriminative HMM, and similar ideas have been developed with CRFs (e.g. Vinh et al. (2011)).

Our models also implicitly assume that the observation sequences for different subjects follow the same distributions. This assumption is likely false, since different individuals have different movement patterns. As a result, the locations in the space of features that are associated with each activity type likely 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.

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. 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; in fact, most specifications of CRF models allow for this type of dependence. We adopted the formulation used in this article so that we could compare the different treatments of the feature vectors in discriminative and generative models with similar structures. However, a more flexible CRF specification might lead to improved classification performance.

While any one of these approaches might be feasible to estimate, we believe that they would lead to only modest improvements in classification performance in the free living setting. We believe that the quality of the data is a more important limiting factor. While free living data is important to collect because laboratory data tends to lead to models that perform poorly outside of the laboratory, data collection is much more difficult in the free living setting than it is in the laboratory (Sasaki et al. (2016)). A result of this is that 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.

Acknowledgements

The authors thank Dr. Stephen Intille (Northeastern University) for making his data available to us. This work was partially supported by National Cancer Institute grant (R01-CA121005).

Supplementary Materials

Supplementary Materials, referenced in Section 4 and Section 5, are available with this paper at the Biometrics website on Wiley Online Library.

References

Adams, R. J., Saleheen, N., Thomaz, E., Parate, A., Kumar, S., and Marlin, B. M. (2016). Hierarchical span-based conditional random fields for labeling and segmenting events in wearable sensor data streams. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 334–343.

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., and Leon, A. S. (2011). 2011 compendium of physical activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise* **43,** 1575–1581.

Albinali, F., Intille, S., Haskell, W., and Rosenberger, M. (2010). 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.

Anderson, M. M. (2013). Physical activity recognition of free-living data using change-point detection algorithms and hidden Markov models. Master's thesis, Oregon State University.

Bai, J., Goldsmith, J., Caffo, B., Glass, T. A., and Crainiceanu, C. M. (2012). Movelets: A dictionary of movement. *Electronic journal of statistics* **6,** 559–578.

Bao, L. and Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, pages 1–17. Springer.

Bernal, A., Crammer, K., Hatzigeorgiou, A., and Pereira, F. (2007). Global discriminative learning for higher-accuracy computational gene prediction. *PLoS computational biology* **3,** 488–497.

Bonomi, A. G., Goris, A., Yin, B., and Westerterp, K. R. (2009). Detection of type, duration, and intensity of physical activity using an accelerometer. *Med Sci Sports Exerc* **41,** 1770–1777.

Bonomi, A. G., Plasqui, G., Goris, A. H., and Westerterp, K. R. (2009). Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. *Journal of Applied Physiology* **107,** 655–661.

Breiman, L. (2001). Random forests. *Machine learning* **45,** 5–32.

Crouter, S. E., Clowers, K. G., and Bassett, D. R. (2006). A novel method for using accelerometer data to predict energy expenditure. *Journal of applied physiology* **100,** 1324–1331.

de Vries, S. I., Garre, F. G., Engbers, L. H., Hildebrandt, V. H., and Van Buuren, S. (2011). Evaluation of neural networks to identify types of activity using accelerometers. *Med Sci Sports Exerc* **43,** 101–7.

Dietterich, T. G., Ashenfelter, A., and Bulatov, Y. (2004). 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. ACM.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70,** 892–898.

Ellis, K., Kerr, J., Godbole, S., and Lanckriet, G. (2014). Multi-sensor physical activity recognition in free-living. In *Proceedings of the 2014 ACM International Joint Conference*

on Pervasive and Ubiquitous Computing: Adjunct Publication, UbiComp '14 Adjunct, pages 431–440, New York, NY, USA. ACM.

Ermes, M., Parkka, J., Mantyjarvi, J., and Korhonen, I. (2008). Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *Information Technology in Biomedicine, IEEE Transactions on* **12,** 20–26.

Foerster, F., Smeja, M., and Fahrenberg, J. (1999). Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Computers in Human Behavior* **15,** 571–583.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*.

Freedson, P. S., Melanson, E., and Sirard, J. (1998). Calibration of the computer science and applications, inc. accelerometer. *Medicine and science in sports and exercise* **30,** 777–781.

Gyllensten, I. C. and Bonomi, A. G. (2011). Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. *Biomedical Engineering, IEEE Transactions on* **58,** 2656–2663.

Kozey, S., Lyden, K., Staudenmayer, J., and Freedson, P. (2010). 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,** 508.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Lester, J., Choudhury, T., Kern, N., Borriello, G., and Hannaford, B. (2005). A hybrid discriminative/generative approach for modeling human activities. In *IJCAI*, volume 5, pages 766–772.

Lester, J., Hartung, C., Pina, L., Libby, R., Borriello, G., and Duncan, G. (2009). Validated caloric expenditure estimation using a single body-worn sensor. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 225–234. ACM.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News* **2,** 18–22.

Lyden, K., Kozey Keadle, S. L., Staudenmeyer, J. W., and Freedson, P. S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Publication Forthcoming* .

Mannini, A., Intille, S. S., Rosenberger, M., Sabatini, A. M., and Haskell, W. (2013). Activity recognition using a single accelerometer placed at the wrist or ankle. *Medicine and science in sports and exercise* .

Mannini, A. and Sabatini, A. M. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors* **10,** 1154–1175.

Mathie, M., Celler, B. G., Lovell, N. H., and Coster, A. (2004). Classification of basic daily movements using a triaxial accelerometer. *Medical and Biological Engineering and Computing* **42,** 679–687.

McShane, B. B., Jensen, S. T., Pack, A. I., and Wyner, A. J. (2013). Statistical learning with time series dependence: An application to scoring sleep in mice. *Journal of the American Statistical Association* .

Nádas, A., Nahamoo, D., and Picheny, M. A. (1988). On a model-robust training method for speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* **36,** 1432–1436.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press.

Pober, D. M., Staudenmayer, J., Raphael, C., and Freedson, P. S. (2006). Development of novel techniques to classify physical activity mode using accelerometers. *Medicine and science in sports and exercise* **38,** 1626.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ravi, N., Dandekar, N., Mysore, P., and Littman, M. L. (2005). Activity recognition from accelerometer data. In *AAAI*, pages 1541–1546.

Rothney, M. P., Neumann, M., Béziat, A., and Chen, K. Y. (2007). An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *Journal of applied physiology* **103,** 1419–1427.

Sasaki, J., Hickey, A., Staudenmayer, J., John, D., Kent, J., and Freedson, P. S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine and Science in Sports and Exercise* To appear.

Smith, A. and Osborne, M. (2007). Diversity in logarithmic opinion pools. *Lingvisticae Investigationes* **30,** 27–47.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management* **45,** 427–437.

Staudenmayer, J., Pober, D., Crouter, S., Bassett, D., and Freedson, P. (2009). An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology* **107,** 1300–1307.

Sutton, C. and McCallum, A. (2011). An introduction to conditional random fields. *Machine*

Learning **4,** 267–373.

U.S. Department of Health and Human Services (2008). 2008 Physical Activity Guidelines for Americans.

Vinh, L. T., Lee, S., Le, H. X., Ngo, H. Q., Kim, H. I., Han, M., and Lee, Y.-K. (2011). Semi-Markov conditional random fields for accelerometer-based activity recognition. *Applied Intelligence* **35,** 226–241.

Xue, J.-H. and Titterington, D. M. (2008). Comment on on discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Neural processing letters* **28,** 169–187.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87,** 954–959.

Zhang, S., Rowlands, A. V., Murray, P., and Hurst, T. L. (2012). Physical activity classification using the genea wrist-worn accelerometer. *Medicine and science in sports and exercise* **44,** 742–748.

Zheng, Y., Wong, W.-K., Guan, X., and Trost, S. (2013). Physical activity recognition from accelerometer data using a multi-scale ensemble method. In *Twenty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence. IAAI.*

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

ALGORITHM 1: **CRF** Estimation Algorithm
Method: estimate_CRF
Inputs: Labeled data $\{(\mathbf{y}_i, \mathbf{x}_i),\ i = 1, \ldots, N\}$
Outputs: CRF parameter estimates.

(1) Initialize all parameter estimates $\widehat{\boldsymbol{\zeta}}$, $\widehat{\boldsymbol{\omega}}$, and $\widehat{\boldsymbol{\beta}}$ to $\underline{0}$.
(2) For $b = 1, \ldots, M_{bag}$, repeat the following:

(a) Draw a sample of N observation sequences with replacement from the set of all observation sequences. Collect the sampled sequences in $\mathcal{B}^b$ and the unsampled sequences in $\mathcal{O}^b$.
(b) Call **boost_CRF**$(\mathcal{B}^b, \mathcal{O}^b)$; the return value is the vector $(\widehat{\boldsymbol{\zeta}}^b, \widehat{\boldsymbol{\omega}}^b, \widehat{\boldsymbol{\beta}}^b)$.
(c) Set $\widehat{\boldsymbol{\zeta}} = \widehat{\boldsymbol{\zeta}} + \frac{1}{M_{bag}}\widehat{\boldsymbol{\zeta}}^b$, $\widehat{\boldsymbol{\omega}} = \widehat{\boldsymbol{\omega}} + \frac{1}{M_{bag}}\widehat{\boldsymbol{\omega}}^b$, and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \frac{1}{M_{bag}}\widehat{\boldsymbol{\beta}}^b$.

(3) Return the combined parameter estimates $(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\beta}})$.

Method: boost_CRF
Inputs: Labeled 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: CRF parameter estimates.

(1) Initialize $m = 0$, `validation_score`$[0] = -\infty$, $\widehat{\boldsymbol{\beta}} = \underline{0}$, $\widehat{\zeta}_s = log(\frac{n_s}{n_S})$ and $\widehat{\omega}_{r,s} = log(\frac{n_{r,s}}{n_{S,S}})$ for all $r, s = 1, \ldots, S$. 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.
(2) Repeat the following until the first occurrence of the largest element of `validation_score` is not within the last `M_search_threshold` values stored in `validation_score`:

(a) Set $m = m + 1$, `attempt_num` $= 0$, and `validation_score`$[m]$ = `validation_score`$[m - 1]$.
(b) Repeat the following until `validation_score`$[m]$ > `validation_score`$[m - 1]$ or `attempt_num` = `max_attempts`:

(1) Set `attempt_num` = `attempt_num` $+ 1$, $\tilde{\boldsymbol{\omega}} = \widehat{\boldsymbol{\omega}}$ and $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$.

(2) Randomly select the set $\mathcal{A}^m \subset \{1, \ldots, D\}$ of active features for the mth update. The number of active features is a user specified parameter.

(3) Using a numerical optimization routine, update $\tilde{\boldsymbol{\omega}}$ and $\tilde{\boldsymbol{\beta}}$ to the constrained local maximum likelihood estimates based on the training data, holding the parameter estimates for elements of $\tilde{\boldsymbol{\beta}}$ not in the active feature set fixed.

(4) Using the estimates from step (2)((b))(3), 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 `validation_score`$[m]$, store it in `validation_score`$[m]$ and set $\widehat{\boldsymbol{\omega}} = \tilde{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$.

(3) Return $(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\beta}})$.

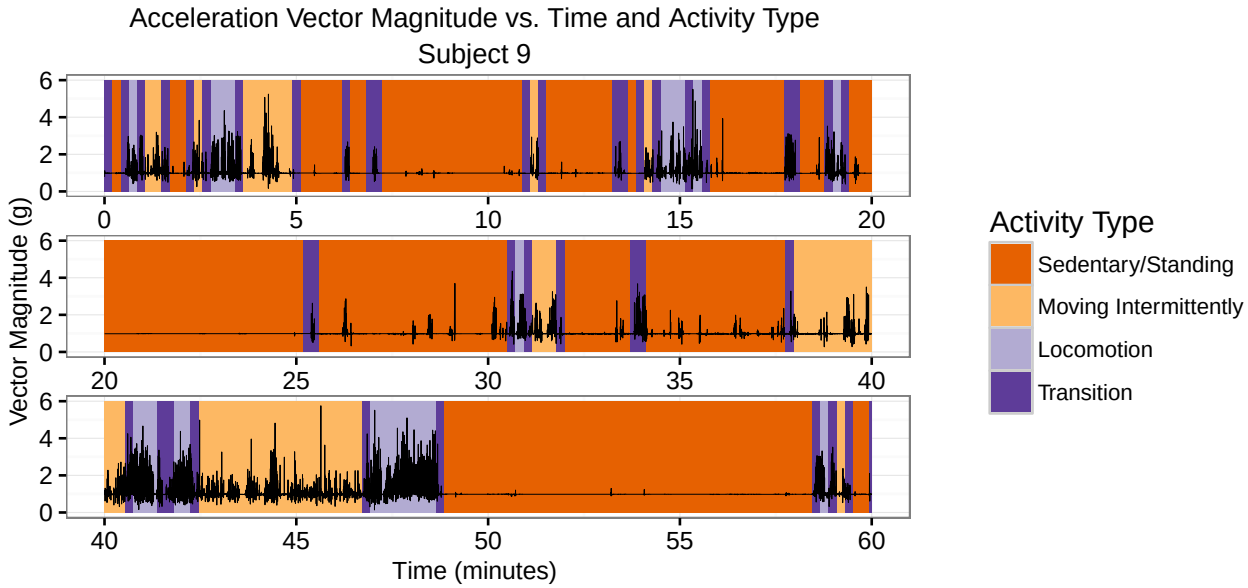

Figure 1. Plot of acceleration recordings from an accelerometer placed at the ankle for one subject in the free living data set from Sasaki et al. (2016); we describe the data further in Section 5. Time is on the horizontal axis and the vector magnitude of the accelerometer recordings at each point in time is on the vertical axis. The background shade indicates the activity type label.

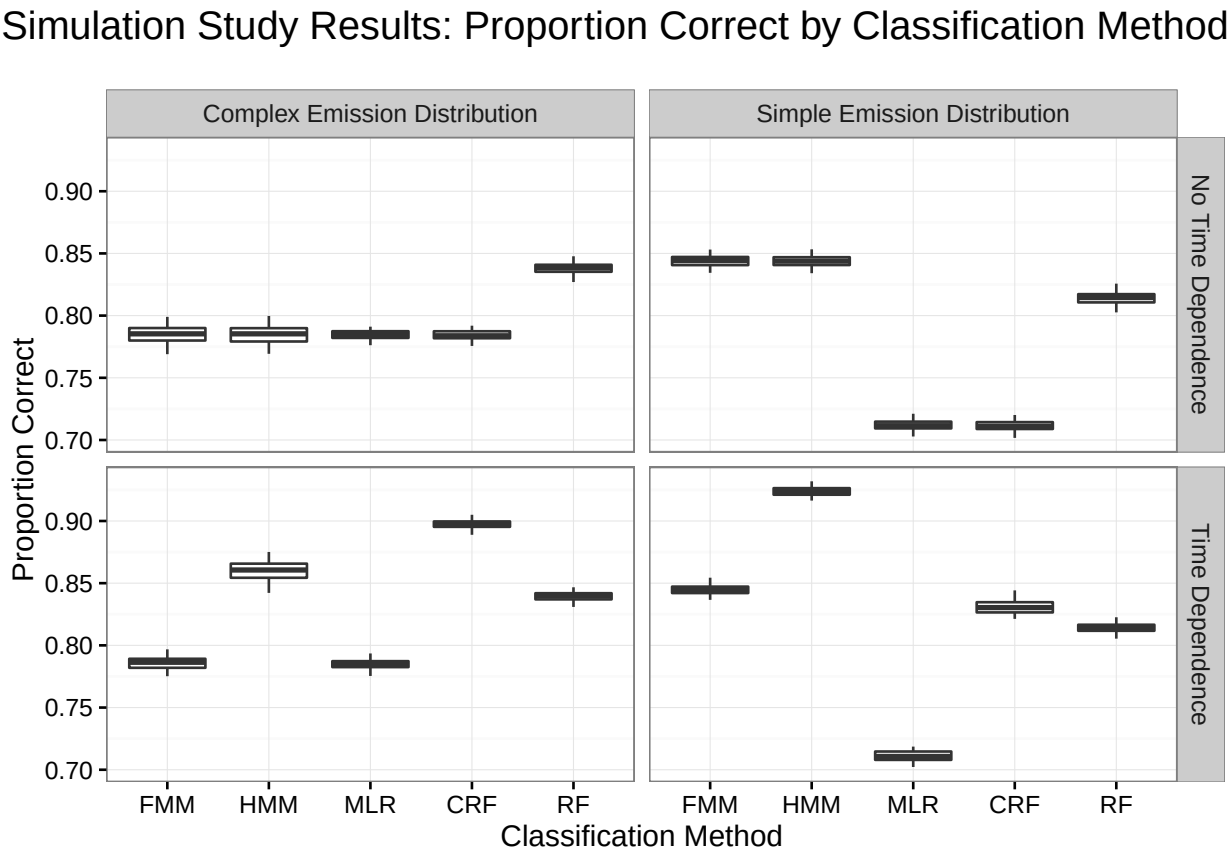

Figure 2. Box plots showing the proportion of time points classified correctly in the simulation study. A separate box plot is displayed for each combination of the complexity level of the feature emission distributions and the classification method. Each point corresponds to a combination of distribution complexity, classification method, and simulation index.

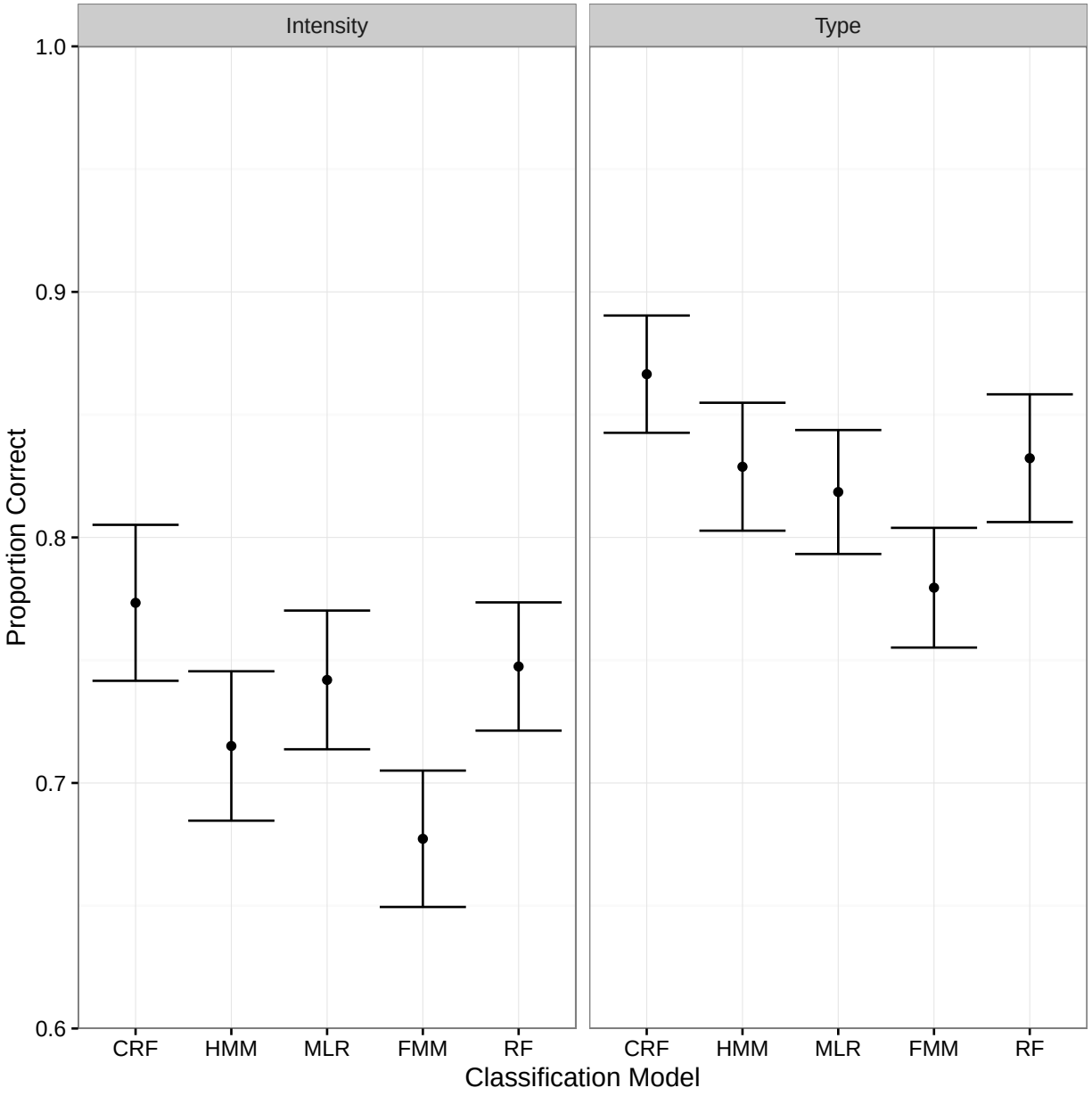

Figure 3. Results from activity type and intensity classification tasks in data from Mannini et al. (2013) and Sasaki et al. (2016), averaged across the three data sets and two accelerometer locations. The joint confidence intervals are from a linear mixed effects model and have a familywise confidence level of 95%.

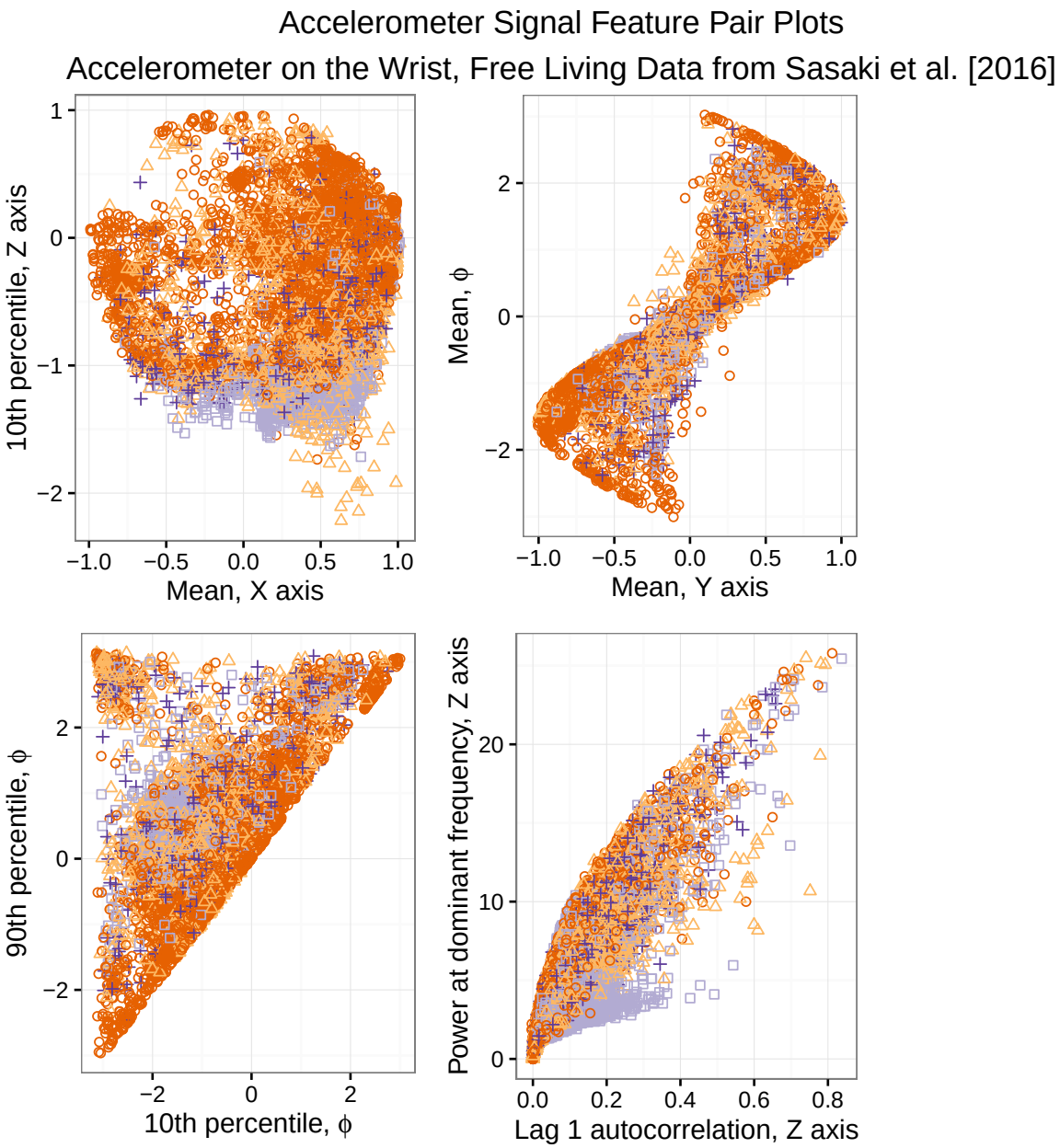

Figure 4. Plots of pairs of accelerometer features. Each point represents one window of length 12.8 seconds in the free living data from Sasaki et al. (2016) with the accelerometer placed at the wrist. Features plotted include means, percentiles, lag one autocorrelation, and power at the dominant frequency of acceleration recorded along each axis, and means and percentiles of the azumithal angle ϕ in a spherical coordinates representation of the signal. The azimuthal angle indicates the relative amounts of acceleration recorded along each coordinate axis in the horizontal plane relative to the accelerometer. There is complex structure in the feature distributions, with multiple modes and a variety of constraints.

	Data Set 1 Mannini Lab	Data Set 2 Sasaki Lab	Data Set 3 Sasaki 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 ± sd)	168.5 ± 9.3 cm	168.6 ± 9.8 cm	169.8 ± 9.8 cm
Weight (mean ± sd)	70.0 ± 15.6 kg	76.4 ± 14.2 kg	74.5 ± 11.4 kg

Table 1
Descriptive statistics for study participants.

Domain	Feature
Time	Mean
	Standard deviation
	Minimum and maximum
Frequency	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.

Domain	Feature	X	Y	Z	VM	θ	ϕ
Time	Mean	Y	Y	Y	Y	Y	Y
	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 place the observed VM values into 10 bins of equal size and calculate 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	Frequency and power of the first dominant frequency	Y	Y	Y	Y	N	N
	Frequency and power of the second dominant frequency	Y	Y	Y	Y	N	N
	Total power: The sum of the estimated power for all frequencies.	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.	Y	Y	Y	Y	N	N

Table 3

Features extracted from the accelerometer signal in preprocessing the data from Sasaki et al. (2016). The right-hand 6 columns indicate whether the listed feature was computed for each of the three axes on which acceleration was measured, vector magnitude, polar angle, and azimuthal angle.

Response	Data Set	Activity Class and Prevalence				
Intensity		Sedentary	Light	Moderate	Vigorous	Transition
	Mannini Lab, Ankle	35.2%	4.8%	55.5%	4.5%	-
	Mannini Lab, Wrist	35.5%	4.8%	55.3%	4.4%	-
	Sasaki Lab	18.4%	44.2%	35.7%	0%	1.7 %
	Sasaki Free Living	24.7%	41.5%	23.6%	0.8%	9.4%
Type		Sedentary	Ambulation	Cycling	Other	
	Mannini Lab, Ankle	39.9%	33.2%	13.8%	13.1%	
	Mannini Lab, Wrist	40.3%	32.9%	14.0%	12.9%	
Type		Sedentary/ Standing	Moving Intermittently	Locomotion	Transition	
	Sasaki Lab	17.2%	53.7%	27.4%	1.7%	
	Sasaki Free Living	45.5%	26.4%	18.9%	9.3%	

Table 4

Prevalence of activity intensity and type labels in the data from Mannini et al. (2013) and Sasaki et al. (2016). In the data from Mannini et al. (2013), prevalence varies slightly across accelerometer locations since different time windows were dropped in the cleaning process they used to handle missing data due to wireless transmission problems with the accelerometers.

Response	Location	Data Set	CRF	HMM	MLR	FMM	RF
Intensity	Ankle	Mannini	**0.890**	0.804	0.868	0.754	0.874
Intensity	Ankle	Sasaki Free Living	**0.732**	0.623	0.689	0.610	0.654
Intensity	Ankle	Sasaki Lab	**0.806**	0.766	0.755	0.710	0.744
Intensity	Wrist	Mannini	**0.870**	0.844	0.796	0.784	0.845
Intensity	Wrist	Sasaki Free Living	0.617	0.505	0.615	0.493	**0.631**
Intensity	Wrist	Sasaki Lab	0.725	**0.750**	0.729	0.711	0.737
Type	Ankle	Mannini	**0.990**	0.982	0.941	0.930	0.956
Type	Ankle	Sasaki Free Living	**0.732**	0.649	0.664	0.630	0.666
Type	Ankle	Sasaki Lab	**0.977**	0.955	0.937	0.876	0.935
Type	Wrist	Mannini	**0.895**	0.893	0.797	0.816	0.868
Type	Wrist	Sasaki Free Living	0.629	0.556	**0.639**	0.534	0.636
Type	Wrist	Sasaki Lab	**0.975**	0.938	0.934	0.891	0.932

Table 5

Estimated mean proportion correct for the activity type and intensity classification tasks in data from Mannini et al. (2013) and Sasaki et al. (2016) by response variable, accelerometer location and data set.