BIOMETRICS 64, 1–20 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 that found that increased physical activity leads to a large number of physical and mental health benefits. 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, 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 can 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; for example, a feature might summarize the mean acceleration recorded along one axis during that window. Features used in this work are described in Section 5). 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 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, such as the mixture of multivariate normal distributions that we consider, 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 previous work has used 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. METs are often discretized into four levels 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 has most commonly been used to relate a univariate discretization of the acceleration experienced over the course of non-overlapping time intervals (referred to as *counts* and often per minute windows) 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 (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. Staudenmayer et al., 2009). Another common option is to use separate models for different types of activity (e.g. Crouter et al., 2006). These methods are static in the sense that the regression models treat activity intensity levels in different windows as conditionally independent given the observed accelerometer data.

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 using the same modeling tools as are used to classify activity type (what the person is actually doing). Many static and discriminative classification methods have been applied to these problems, including support vector machines (e.g. Mannini et al. (2013)), classification trees (e.g. Bonomi et al. (2009)), artificial neural networks (e.g. Staudenmayer et al. (2009)), and nearest neighbors (e.g. Bao and Intille (2004)), 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. 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 approach was used by Mannini and Sabatini (2010). A difficulty with this approach is that it requires us to estimate the distribution of the accelerometer features associated with each hidden state. In general though, 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).

A second approach to using HMMs is to first use a discriminative model that does not incorporate temporal dependence to obtain an initial classification, and then smooth those initial classifications over time with an HMM (e.g. Lester et al. (2005)). McShane et al. (2013) developed a more formally justified variation on this theme. Their work re-expresses the HMM using class membership probabilities that are obtained from a static classification model. This 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 proposed in the computer science literature by Lafferty et al. (2001) in the context of natural language processing. 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 physical activity and health (Vinh et al., 2011; Adams et al., 2016).

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**). Theese models 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 a 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. A

mixture of K_s multivariate normal distributions is used to model the distribution of the accelerometer features for each activity type s. This is a static model because the class membership probabilities at time t depend only on the observed features at that time. Also, this is a generative model because it models the distribution for the observed features.

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 temporal dependence in activity class through the transition probabilities in Equation (4) and is therefore a dynamic model. Like the **FMM**, it uses a mixture of normals as a generative model for the accelerometer features within each activity class.

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 given the accelerometer features at all times. For the linear chain CRF considered here, a forward-backward algorithm can be used to marginalize this distribution to obtain distributions for activity classes at individual time points; similar methods are also available for more complex CRFs (Sutton and McCallum (2011)). 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)). We include the **RF** to enable rough comparison of the other approaches described above with a commonly used class of methods in the activity classification literature: static, discriminative methods that are more flexible than **MLR**. For example, the previously published article analyzing one of the data sets we will work with in Section 5 used several static discriminative methods. We use the implementation of random forests in the `randomForest` package (Liaw and Wiener, 2002) for `R` (R Core Team, 2016) with the default options for tuning parameters.

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. 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 `R`'s `car` package (Fox and Weisberg, 2011).

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. 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 the supplement. Similar estimation strategies for CRFs have been employed previously (e.g., Smith and Osborne (2007)). In order to resolve problems with identifiability, we fix $\zeta_S = 0$, $\omega_{S,S} = 0$, and $\beta_{S,d'} = 0 \, \forall d' = 0, \ldots, D$.

None of the estimation algorithms described here are guaranteed to yield parameter estimates at the global maximum of the likelihood. The likelihood for a mixture of normals has multiple local maxima, and the EM algorithms used in estimating these models are only guaranteed to converge to a local maximum (Fraley et al. (2012)). The **CRF**'s likelihood

poses similar challenges (Sutton and McCallum (2011)). Also, because of the use of early stopping to prevent overfitting, we can be confident that even a local maximum of the likelihood is not reached. It is possible that the performance of any of these methods could be improved with more refinement to these estimation procedures.

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 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 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. We used only 50 simulations because on average across the simulations, estimation of the CRF took about 8 hours to complete using 16 cores. 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:

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

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

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

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.

In designing this simulation study, our goal was that the setting with a complex emission distribution and time dependence would recreate the general characteristics of the real physical activity data that we work with in Section 5. The amount of training data available for model estimation and dimension of the feature vectors are in line with the data sets we will work with in Section 5. In the setting with time dependence, the probability of remaining in the same activity class is 0.8 in the simulation study; in the free living data set we will examine in Section 5, this probability was about 0.84. Similarly, the "complex" emission distributions attempt to capture the sorts of non-linear dependencies in the feature distributions that we observed with physical activity data. See Section 5.

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 differences in mean model performance discussed here are both statistically and practically significant (Supplemental Figures 1 and 2). The relative performance of the methods is the same when measured by the macro F_1 score (Supplemental Figure 3).

[Figure 2 about here.]

5. Applications

In this Section, we present classification results for three physical activity data sets. Data collection procedures are described in Subsection 5.1 and results are in Subsection 5.2.

5.1 *Data Collection*

Our three data sets were collected in two studies, Mannini et al. (2013) and Sasaki et al. (2016), and participant descriptive statistics are in Table 1. For the first dataset (Mannini et al. (2013)), each participant performed a subset of 26 activities in the laboratory. These activities were designed to be representative of activities people engage in in real life, but the order and duration of activities were determined by the researchers. The participants wore accelerometers on their ankle and wrist which recorded acceleration in each of 3 orthogonal axes at 90 Hz. In their analyses, Mannini et al. (2013) developed static discriminative statistical learning models (support vector machines [SVM]) to classify activities into one of several groups of similar activity types.

[Table 1 about here.]

Our second and third datasets (Sasaki et al. (2016)) were collected using healthy elderly participants 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 first dataset, and the third dataset was collected 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.

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 recommended in that work as making a good trade-off between computational complexity and classification performance. 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. We took different approaches to these datasets because we did not have access to the raw accelerometer files from the first dataset. Tables 1 and 2 in the supplement list the features. Although a more detailed consideration of the merits of different feature sets is beyond the scope of this article, we refer to Mannini et al. (2013), who examined that question.

In addition to summarizing acceleration in each window, we also assigned an activity type label and intensity to 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 using the four categories described previously in Section 2. We also used a Transition category for activity intensity. Classification performance was evaluated relative to the actual activity type by leave-one-subject-out cross-validation.

Figure 3 displays the distributions of four pairs of wrist accelerometer features for each activity type classification in the free living data from Sasaki et al. (2016). The figure illustrates that these distributions have complex structure, including multiple modes and a variety of linear and non-linear constraints.

[Figure 3 about here.]

5.2 *Results*

Figure 4 summarizes the results. The left panel shows average percent correct for activity intensity classifications and the right shows the same for activity type classifications. The 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 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 **RF** used in Sasaki et al. (2016) is similar to the **RF** presented here; their classification results are not identical because of differences in how the data were preprocessed.

[Figure 4 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. Table 2 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 (see Supplement).

[Table 2 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. We examined these relationships using a structured family of related models where the dynamic models (**HMM** and **CRF**) can be obtained by adding a model for temporal dependence to a static model (**FMM** and **MLR**), and the discriminative models (**MLR** and **CRF**) can be obtained by conditioning on the observed accelerometer data in a generative model (**FMM** and **HMM**). Additionally, we compared to a random forest, which is broadly representative of flexible static and discriminative models that are commonly used in the field.

The argument in favor of dynamic, discriminative models is that they are a better representation of the data than static or generative methods are. Many activities such as sedentary behavior and bouts of purposeful exercise occur in contiguous blocks of time, so the type and intensity of activity that an individual engages in at nearby times tend to be similar. Dynamic models explicitly model this temporal dependence and "borrow strength" over time to achieve improved classification. We believe that is why they performed better in our analyses of actual datasets. That said, we note that some activity types are less likely than others to be found in contiguous blocks of time; for example, it has been found that about 60 percent of human walking bouts have duration less than 30 seconds, roughly the time scale where dependence across time windows of length 12.8 seconds is relevant (Orendurff et al. (2008)). Dynamic methods may be less helpful for improving classification of these sorts of activities in free living data.

We believe that discriminative methods are superior to generative methods because they do not require a specification for the distribution of the feature vector. That distribution is

too complex to model well with simple parametric approaches, and is too high-dimensional to be handled easily by non-parametric methods.

Although the **CRF** consistently had the best performance among the methods presented here, there is still room for improvement. The proportion of time windows classified correctly was lower in the free living settings than in the corresponding laboratory settings, by an amount between about 5% and 30% depending on the response and the accelerometer location. Classification performance was also generally lower when the response was intensity than when the response was activity type. For the free living data, the best classification results were achieved using a **CRF** and data from an accelerometer placed at the ankle; yet in this case, only about 73% of time points were classified correctly. These performance levels will probably be sufficient for some applications and insufficient for others. That aspect of this work deserves further study, but is outside the scope of this work.

Our results echo previously published findings that the location of the accelerometer is an important determinant of classification performance, as was seen in both Mannini et al. (2013) and Sasaki et al. (2016). The results in Table 2 confirm these findings: across every combination of response variable and data set, and for all of the CRF, HMM, MLR, and RF models, the proportion of time points classified correctly was always higher when using data from the ankle than when using data from the wrist. We have not examined the possibility of using information from multiple accelerometers to inform classifications, but it is possible that such an approach might lead to better classification performance.

We did not explore the influence of non-overlapping window size in this paper. This issue was examined in Mannini et al. (2013) and Sasaki et al. (2016) who found that longer window sizes were generally associated with increased classification accuracy. Another option is 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).

Finally, data collection procedures are extremely important in the free living setting. Free living data is important to collect because laboratory data tends to lead to models that perform poorly outside of the laboratory, but data collection is much more difficult in the free living setting than it is in the laboratory Lyden et al. (2014). A result of this is that recorded labels for physical activity type are likely less reliable for data from free living participants. 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. This challenge could also be addressed by using unsupervised or partially supervised classification methods; this could also simplify the data collection process and facilitate the use of much larger data sets since accurately recorded activity type labels would not be required. One example along these lines is in Trabelsi et al. (2013), who use an unsupervised hidden Markov model to perform physical activity classification.

Acknowledgements

The authors thank Dr. Stephen Intille (Northeastern University) for making his data available. 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. All code is available at `www.github.com/elray/PACw`

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.

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.

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.

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.

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

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.

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.

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.

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* .

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.

Orendurff, M. S., Schoen, J. A., Bernatz, G. C., and Segal, A. D. (2008). How humans walk: bout duration, steps per bout, and rest duration. *Journal of rehabilitation research and development* **45,** 1077.

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

Foundation for Statistical Computing, Vienna, Austria.

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.

Trabelsi, D., Mohammed, S., Chamroukhi, F., Oukhellou, L., and Amirat, Y. (2013). An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE Transactions on Automation Science and Engineering* **10,** 829–835.

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.

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

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

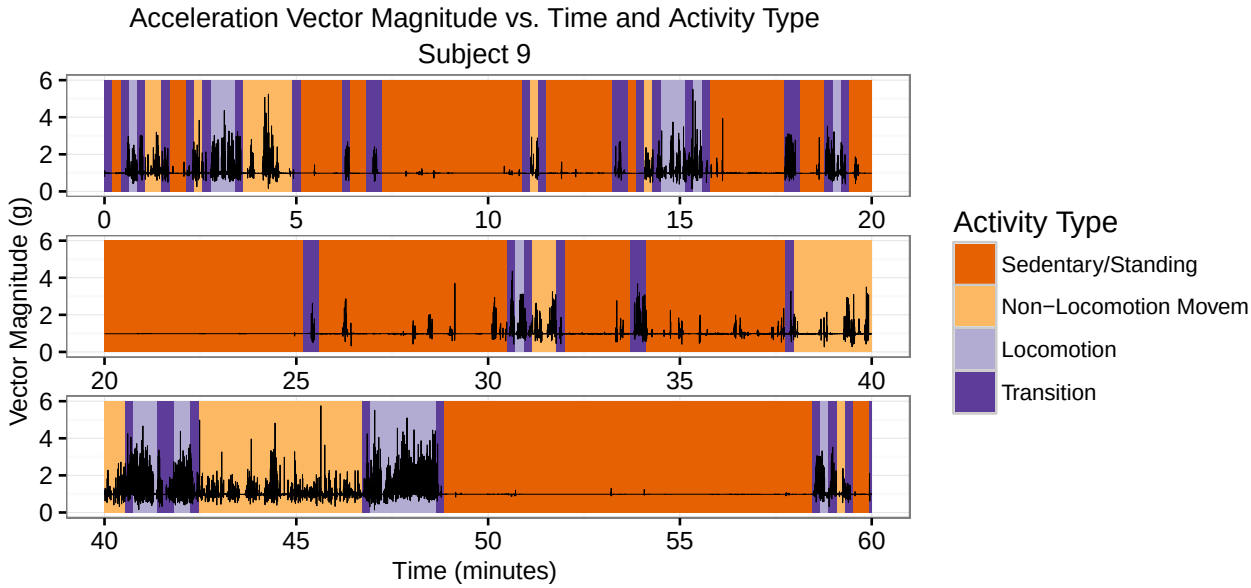

Figure 1. Plot of acceleration recordings from an accelerometer placed at the ankle for one subject in the free living data set from Sasaki et al. (2016); we describe the data further in Section 5. Time is on the horizontal axis and the vector magnitude of the accelerometer recordings at each point in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

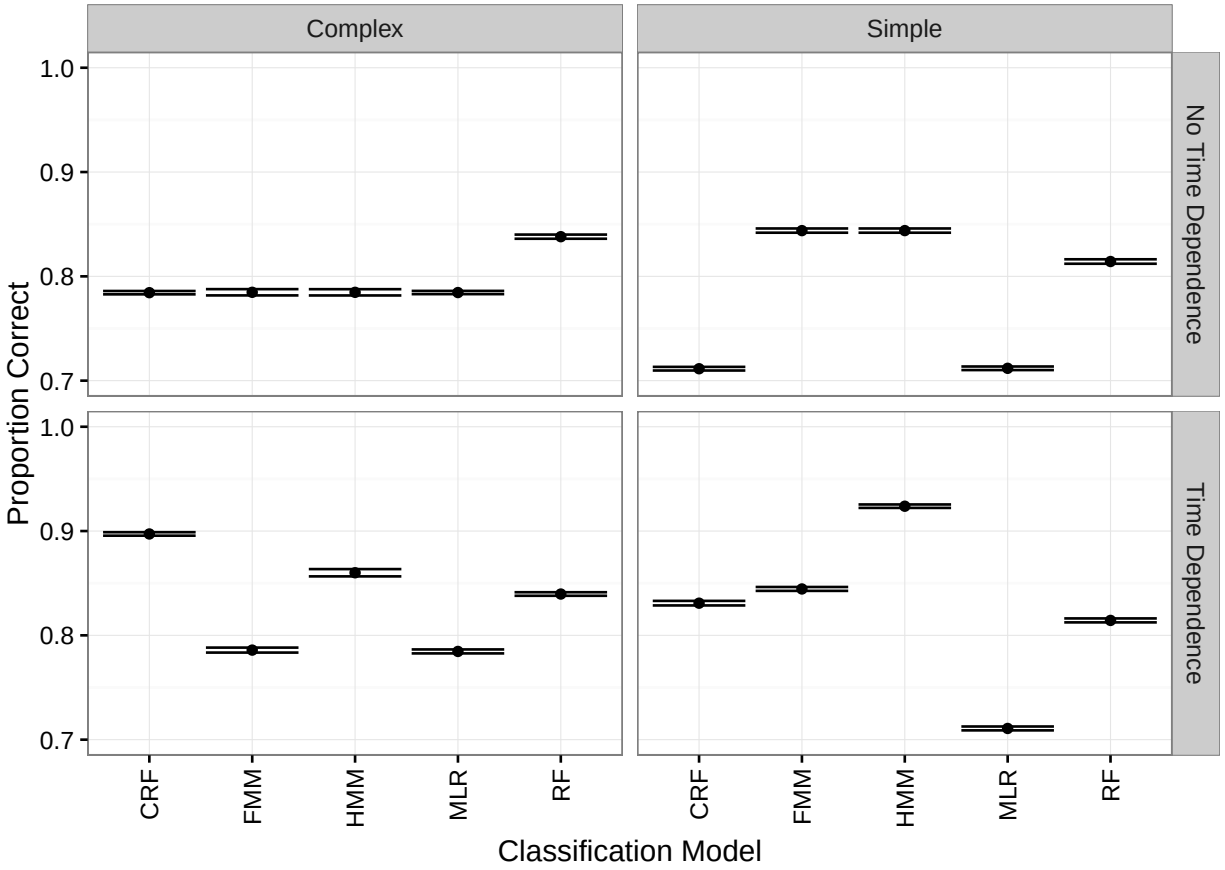

Figure 2. Estimates of the mean proportion of time points classified correctly for each combination of complexity level of the feature emission distributions, presence or absence of temporal dependence in the data generating process, and classification method in the simulation study. The confidence intervals are from a linear mixed effects model and have a familywise confidence level of 95%.

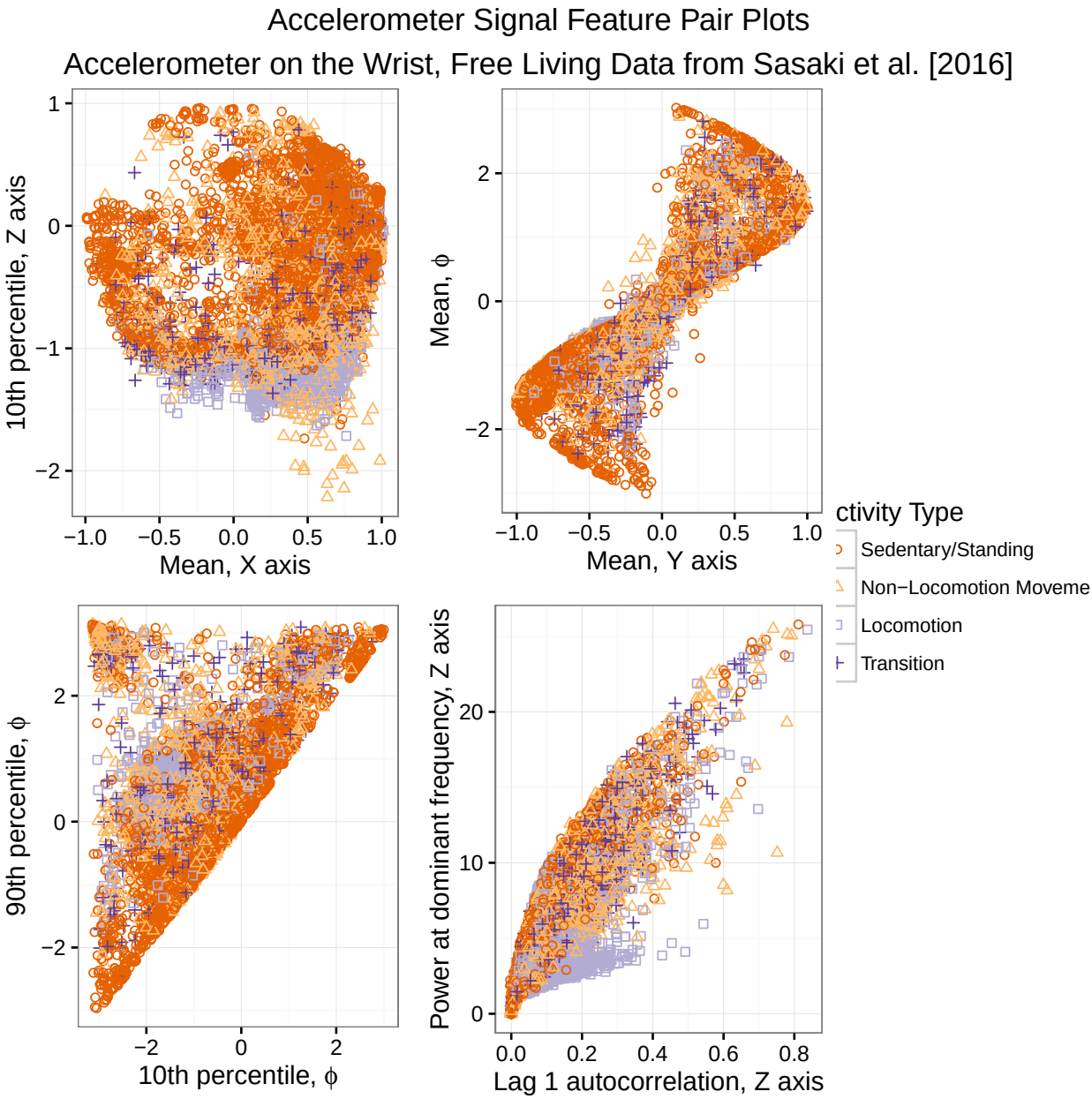

Figure 3. Plots of pairs of accelerometer features. Each point represents one window of length 12.8 seconds in the free living data from Sasaki et al. (2016) with the accelerometer placed at the wrist. Features plotted include means, percentiles, lag one autocorrelation, and power at the dominant frequency 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.

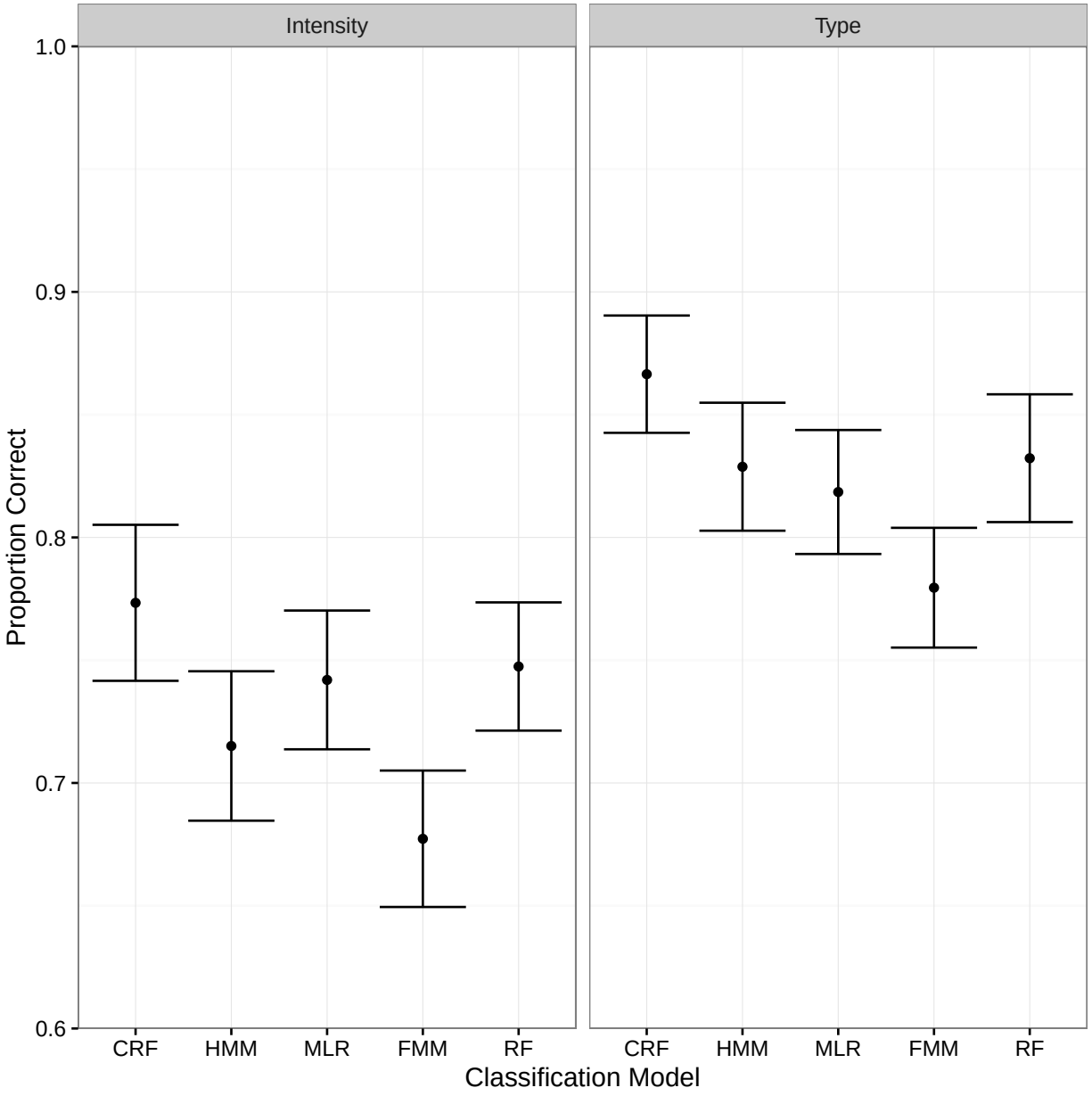

Figure 4. Results from activity type and intensity classification tasks in data from Mannini et al. (2013) and Sasaki et al. (2016), averaged across the three data sets and two accelerometer locations. The joint confidence intervals are from a linear mixed effects model and have a familywise confidence level of 95%.

	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.

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 2

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.