Supplement to Physical Activity Classification with Dynamic, Discriminative Methods

Evan Ray
Department of Mathematics and Statistics, University of Massachusetts,
Amherst, MA 01003-9305, USA
Evan.L.Ray@gmail.com

Jeffer Sasaki
Graduate Program in Physical Education, Universidade Federal do Triangulo Mineiro,
Uberaba, Minas Gerais, Brazil

Patty Freedson
Department of Kinesiology, University of Massachusetts,
Amherst, MA 01003-9305, USA

John Staudenmayer
Department of Mathematics and Statistics, University of Massachusetts,
Amherst, MA 01003-9305, USA

1 Introduction

In this supplement we present additional results showing the performance of each method in the simulation study and in the applications as measured by the macro F_1 score, as well as a description of the mixed effects models used to estimate mean performance for each model and assess whether the differences in model performance were statistically significant.

2 Simulation Study

In the main manuscript, we summarized results for the simulation study using the proportion of windows that were classified correctly. Here we display summaries of the macro F_1 scores achieved by each method. The macro F_1 score is defined as

$$\begin{aligned} F_1 &= 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \text{ where} \\ \text{Precision} &= \frac{1}{S}\sum_{s=1}^{S} \frac{\text{TP}_s}{\text{TP}_s + \text{FP}_s} \text{ and} \\ \text{Recall} &= \frac{1}{S}\sum_{s=1}^{S} \frac{\text{TP}_s}{\text{TP}_s + \text{FN}_s} \end{aligned}$$

Here, TP_s, FP_s, and FN_s are respectively the true positive rate, false positive rate, and false negative rate for class i. This score is a useful complement to the overall proportion correct because it incorporates both precision and recall and gives equal weight weight to all classes, whereas the proportion correct gives more weight to more prevalent classes [Sokolova and Lapalme, 2009].

For the simulation study, the relative performance of the methods as measured by the macro F_1 score was the same as it was when the methods were evaluated using the proportion correct (Supplemental Figure 1).

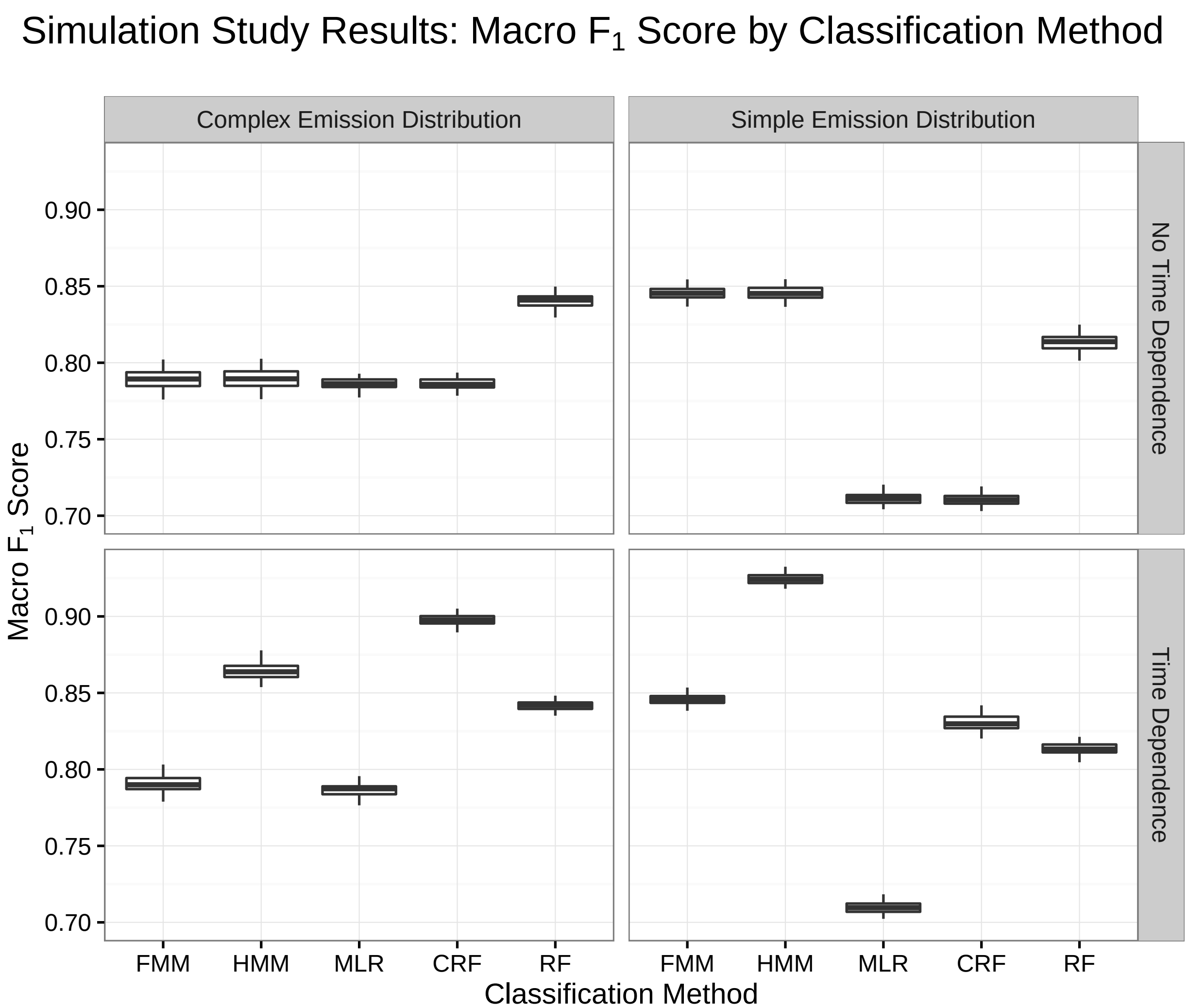
3 Applications

Here we present the classification results in the applications, as summarized by the macro F_1 score (Supplemental Figure 2). With the F_1 score, the differences in performance between the dynamic models and the corresponding static models are statistically significant at the $\alpha = 0.05$ level. The differences in mean F_1 score between the generative and discriminative models are not statistically significant or consistent in direction across classification of activity type or intensity. This is different from the measure of proportion correct discussed in the main manuscript, where discriminative models generally outperformed their static counterparts by a statistically significant margin. These results are consistent with Supplemental Table 1, where we present the mean F_1 score separately for each combination of response, location, and data set. Across all of these combinations, the dynamic models tended to achieve higher F_1 scores than the static models, and the **CRF** had the most consistent performance as measured by the F_1 score.

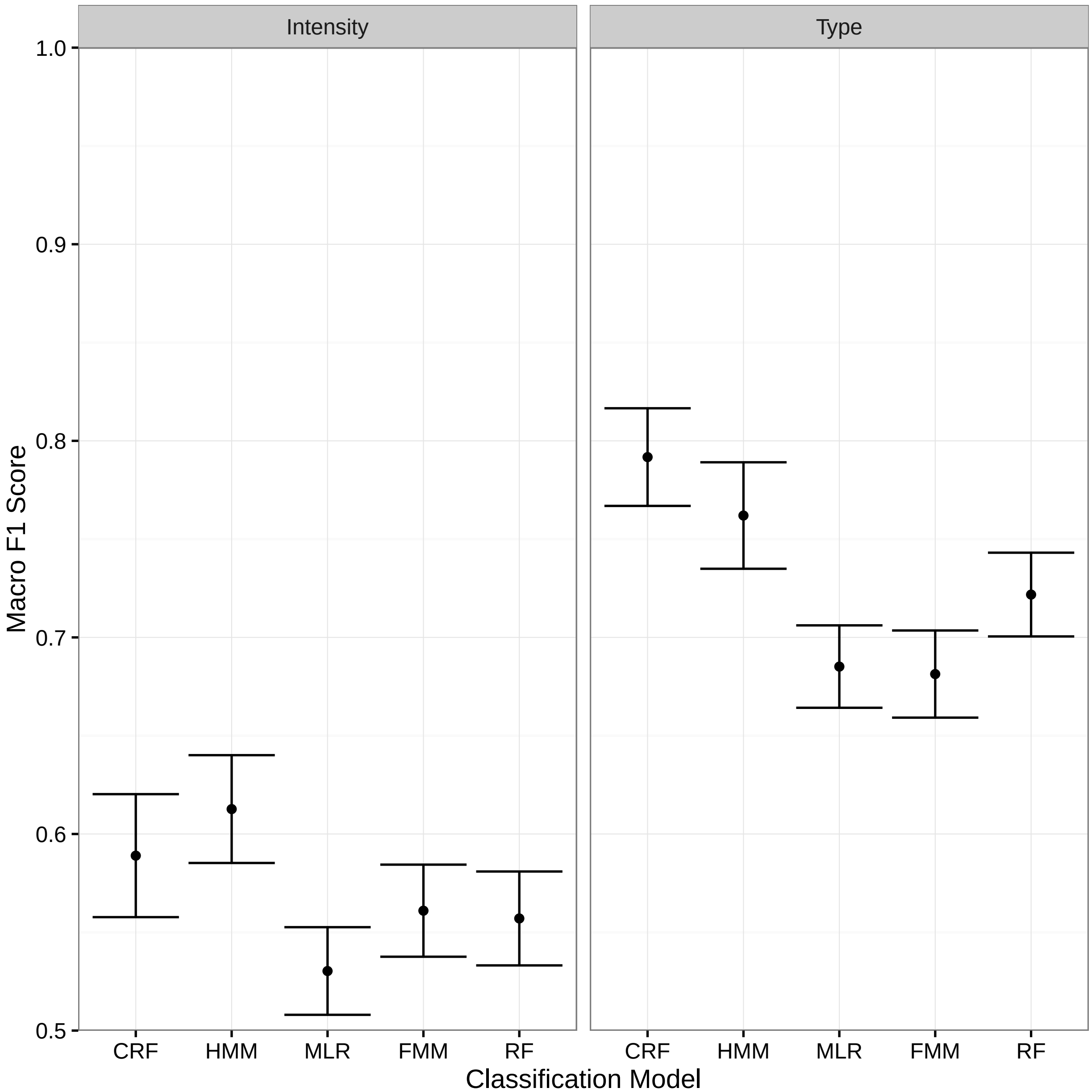
The confidence intervals displayed in Figure 3 of the main manuscript and Supplemental Figure 2, as well as the hypothesis test results discussed throughout the text, were obtained using linear mixed effects models with the following specification:

$$\begin{aligned} y_{r,l,d,c,s} &= \mu_{r,l,d,c} + \alpha_{d|s} + \varepsilon_{r,l,d,c,s} \\ \alpha_{d|s} &\sim N(0, \sigma_s^2) \\ \varepsilon_{r,l,d,c,s} &\sim N(0, \xi_{r,l,d,c}^2) \end{aligned}$$

In this notation $y_{r,l,d,c,s}$ is a measure of classifier quality (either proportion correct or macro F_1 score) for one instance, indexed by r denoting the response (activity type or activity intensity), l denoting the accelerometer location (ankle or wrist), d denoting the data set (Mannini, Sasaki Free Living, or Sasaki Lab), c denoting the classifier (**CRF**, **HMM**, **MLR**, **FMM**, **RF**), and s denoting the subject within each study. The $\alpha_{d|s}$ term is a random effect for each subject; the notation $d|s$ emphasizes that we treat the subjects in different data sets separately for the purpose of this model, even though the subjects in the Sasaki Free Living data set also participated in the Sasaki Lab data collection. The error term, $\varepsilon_{r,l,d,c,s}$, has a separate variance for each combination of response, location, data set, and classifier. We fit a separate model for each measure of classifier quality using the `nlme` package [Pinheiro et al., 2017] in `R` [R Core Team, 2016]. For each measure of classifier quality, we conducted all hypothesis tests simultaneously with construction of the confidence intervals in Figure 3 of the manuscript and Supplemental Figure 2 in this document using the `multcomp` package [Hothorn et al., 2008] for `R`.



Supplemental Figure 1: Box plots showing the macro F_1 score combining precision and recall across all three classes in the simulation study. A separate box plot is displayed for each combination of the complexity level of the feature emission distributions, the Bayes error rate, and the classification method. Each point corresponds to a combination of distribution complexity, Bayes error rate, classification method, and simulation index.



Supplemental Figure 2: 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%.

Response	Location	Data Set	CRF	HMM	MLR	FMM	RF
Intensity	Ankle	Mannini	0.480	0.574	0.468	**0.579**	0.555
Intensity	Ankle	Sasaki Free Living	**0.538**	0.522	0.526	0.505	0.483
Intensity	Ankle	Sasaki Lab	**0.789**	0.737	0.680	0.654	0.675
Intensity	Wrist	Mannini	0.611	**0.690**	0.496	0.630	0.608
Intensity	Wrist	Sasaki Free Living	0.419	**0.451**	0.413	0.421	0.417
Intensity	Wrist	Sasaki Lab	0.696	**0.703**	0.599	0.577	0.604
Type	Ankle	Mannini	**0.978**	0.978	0.921	0.916	0.941
Type	Ankle	Sasaki Free Living	**0.590**	0.547	0.523	0.516	0.541
Type	Ankle	Sasaki Lab	**0.938**	0.882	0.783	0.762	0.808
Type	Wrist	Mannini	**0.872**	0.867	0.737	0.786	0.829
Type	Wrist	Sasaki Free Living	0.424	0.459	0.424	0.434	**0.485**
Type	Wrist	Sasaki Lab	**0.949**	0.839	0.723	0.674	0.727

Table 1: Estimated mean macro F_1 score 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.

References

Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

Andrea Mannini, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. Activity recognition using a single accelerometer placed at the wrist or ankle. *Medicine and science in sports and exercise*, 2013. doi: 10.1249/MSS.0b013e31829736d6.

Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2017. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-131.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

Jeffer Sasaki, Amanda Hickey, John Staudenmayer, Dinesh John, Jane Kent, and Patty S. Freedson. Performance of activity classification algorithms in free-living older adults. *Medicine and Science in Sports and Exercise*, 2016. To appear.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.