Supplementary Materials for

# Multiple Kernel Learning with Random Effects for Predicting Longitudinal Outcomes and Data Integration[1]

**Tianle Chen, Donglin Zeng, Yuanjia Wang**

**Additional Acknowledgements**: ADNI data collection and sharing was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; ; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The authors wish to thank the NIH dbGap data repository (accession number phs000222.v1.p1), and the PREDICT-HD study investigators.

[1]This supplement contains one additional simulation study, and an additional real data example based on PREDICT-HD study.

## A.1 Additional simulation study: single data source

In this simulation setting, we generated the dichotomous outcomes from the following model:

$$Y_{ij} = \text{sign}\{\beta W^*_{ij} + a_{ij} + b_{ij} + \epsilon_{ij}\},$$

where $W^*$ is the radius of the two spheres in the Supplementary Materials Figure A1. First we generated $\boldsymbol{W}_{ij}$, a 3-dimensional vector randomly located either on the outer sphere with a radius equal to 2 (with a small random error) or on the inner sphere with a radius equal to 1 (with a small random error) at each visit for a subject. We used the radius $W^*$ in the score function for generating the binary outcome. The radius changes at each visit (with equal probability to be 1 or 2). A single radial kernel SVM can generate a sphere-shaped boundary and perfectly separate the two groups of $\boldsymbol{W}$'s. $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ are subject-specific random effects. Specifically, $\boldsymbol{a_i}$ is generated from $MVN(\mathbf{0}, \Sigma_a)$, where $\Sigma_a$ is a correlation matrix with compound-symmetric structure ($\rho = 0.5$), and $\boldsymbol{b_i}$ is generated from $MVN(\mathbf{0}, \Sigma_b)$, where $\Sigma_b$ is a correlation matrix with exponential correlation structure, e.g., $\rho_{j,k} = \exp(-\alpha|t_j - t_k|)$ with $\alpha = 1$. Here $\epsilon_{ij}$ are normally distributed random errors of the $i^{th}$ subject at the $j^{th}$ visit. We performed 100 simulation runs and compared various performance indices of the proposed method under a single linear or radial kernel with and without random effects to logistic regression ignoring correlation and generalized mixed effects regression with subject-specific random intercepts. For logistic regression and generalized mixed effects regression, we included all the feature variables, their squared terms and pairwise interactions as predictors. The parameters $\rho$ and $\alpha$ associated with long-term and short-term random effects were determined by small scale cross-validation.

In type A prediction, we predicted outcomes for a new subject based on his/her observed feature variables alone and the trained model. We generated longitudinal data from the single source $W$ plus latent random effects with a sample size of $n = 250$ subjects, each having 4 visits. Two-thirds of the subjects are included in the training set and the rest as the testing set. The results are summarized in the top panel of Table A1 and Figure A3. The performance of

the linear kernel SVM is poor so only the radial kernel SVM results are shown in Figure A3. On average, the two radial kernel SVMs (fixed-effects and mixed-effects) have better accuracy (1-misclassification rate), sensitivity and negative predictive value (NPV). The specificity and positive predictive value (PPV) is slightly lower. Including random effects in the prediction improves accuracy and leads to smaller variability over repeated simulations. A paired $t$-test comparing radial kernel SVM with and without random effects shows a significant decrease in misclassification rate ($p < 0.001$).

In type B prediction, we predicted the future follow-up outcomes for the same subject based on his/her observed features variables and prior visits' outcomes and the trained model. In this case each subject was generated to have 6 visits. The first 3 visits of each subject are used as the model-building set and the rest 3 visits as the testing set. In this case we can compute the fitted random effects for each subject using the designed kernel functions, and the subject-specific outcomes for the last 3 visits can be predicted for each subject incorporating fitted random effects. The results are summarized in the bottom panel of Table A1 and Figure A3. Here we see more improvement for SVM-based approach compared to the generalized mixed effects regression or logistic regression, and again extracting information from the distribution of random effects leads to smaller variability for each of the performance index. A paired $t$-test comparing radial kernel SVM with and without random effects shows a significant decrease in misclassification rate ($p = 0.024$).

## A.2 Application to PREDICT-HD study

In this real data example, we predict Huntington's disease (HD) diagnosis in a large multi-site HD epidemiological study from various sources of clinical interviews and biomarkers and show that the proposed method outperforms single kernel approaches and multiple kernel approaches without accounting for subject-specific correlations in terms of both predicting future subjects and predicting future outcomes on the same subject.

We applied the developed method to PREDICT-HD (Paulsen et al., 2008), a multi-center

observational study on Huntington's disease (HD). HD is an autosomal dominant disease caused by an expansion of CAG trinucleotide repeats in ITI5 gene on chromosome 4 (Huntington's Disease Collaborative Research Group, 1993). Majority of subjects with an expansion of CAG repeats in IT15 gene (CAG repeats $\geq 36$) on one allele will develop HD if not censored by death (Kieburtz and Huntington Study Group, 1996). The diagnosis of HD is based on the diagnostic confidence level (DCL), a measure ranging from 0 to 4 based on the UHDRS assessment. A DCL of 0 means no abnormalities and 4 means motor abnormalities that are unequivocal signs of HD with 99% confidence. Subjects with a DCL of 2 or higher can be considered as showing motor abnormalities that may be signs of HD with more than 50% confidence. There are 941 CAG-expanded participants in the data set who have complete data for analysis. The median age is 40 years old and the range is from 18 to 75. 195 participants have a DCL of 2 or higher at the baseline and totally 126 subjects reached a DCL of 4 during the study.

The goal of PREDICT-HD analysis is to distinguish among those who showed noticeable motor signs of HD from those who did not. The analysis sample included 449 participants who had 4 or more visits, and the outcome of interest is whether a subject had a DCL$\geq$ 2 versus DCL$<$ 2 at each visit. The data sources include demographic data (age, gender and education level), genetic marker (CAG repeat length), motor and functional measures (total motor score and total functional capacity), cognitive function measures (stroop color, digital and word, and symbol digit modalities tests) and psychiatric assessment scores obtained through FRSBE (Frontal Systems Behavior Scale), SCL90 (Global Severity Index, Positive Symptom Total and Distress Index) and UHDRS (Unified Huntington's disease rating scale). For the multiple fused kernel SVM, we used 5 separate kernels for the feature variables: a linear kernel for age at visit since age appears to be an important biomarker for HD (Chen et al. 2003), a radial kernel for all the continuous clinical measures and cognitive scores, another radial kernel for their interaction with age, a linear kernel for genetic marker and other demographic variables, and another linear kernel for their interaction with age. We considered several different combinations of kernel types and suggested the one with the smallest misclassification rate by small scale cross-validation.

We first assessed the type A prediction by treating one third of subjects as new subjects and predicting their outcomes at several visits based on the model trained from the rest of two-thirds subjects. For the standard approaches (logistic regression and generalized mixed effects regression), all the feature variables and their interactions with age were used as predictors in the model. We only reported results from logistic regression without random effects since generalized mixed effects model failed to converge potentially due to a large number of feature variables included in the model. We compared the performance of five methods: a logistic regression, a fixed-effects single radial kernel SVM, a fixed-effects multiple radial kernel SVM, a fixed-effects multiple fused kernel SVM, and a mixed-effects multiple fused kernel SVM. The performance of methods using multiple kernels are much better than using the single kernel in all the measures except for specificity (due to very low sensitivity). For example, the accuracy for the single-kernel SVM is only 0.66 and the sensitivity is 0.15, while for all other methods the accuracy is around 0.85 and the sensitivity is between 0.75 and 0.80. In the top panel of Figure A4, we show the performance of four methods. We can see that the kernel-based methods perform better the logistic regression in four out of five fit indices (similar sensitivity). Including random effects into multiple fused kernels improves accuracy, specificity and NPV.

Next, we assessed the type B prediction of future observations on existing subjects. We used the first two visits of each subject as the training set to predict the subject-specific outcomes at the rest of follow-up visits. Since the division of training set and testing set is fixed in this setting, we repeated the process $n$ times, taking one subject out each time. From the bottom panel of Figure A4, we can see that the accuracy is higher when including random effects, and its standard deviation is much smaller, indicating stability of the results. Although the sensitivity and NPV are slightly lower, the specificity and PPV are much higher.

## References

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinu-

cleotide repeat that is expanded and unstable on huntingtons disease chromosomes. *Cell* **72,** 971–983.

Kieburtz, K. and Huntington Study Group (1996). The unified huntington's disease rating scale: reliability and consistency. *Movement Disorders* **11,** 136–142.

Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L. J., Duff, K., Kayson, E., Biglan, K., Shoulson, I., Oakes, D., and Hayden, M. (2008). Detection of huntington's disease decades before diagnosis: the predict-hd study. *Journal of Neurology, Neurosurgery & Psychiatry* **79,** 874–880.

Table A1: Simulation: single data source.

| | type A prediction | | | | |
|---|---|---|---|---|---|
| | misclassification | sensitivity | specificity | PPV | NPV |
| logistic regression | 0.0609 (0.0163) | 0.9339 (0.0254) | 0.9443 (0.0206) | 0.9433 (0.0208) | 0.9352 (0.0242) |
| generalized mixed effects regression | 0.0613 (0.0156) | 0.9316 (0.0249) | 0.9458 (0.0202) | 0.9446 (0.0203) | 0.9332 (0.0237) |
| single linear kernel (fixed-effects) | 0.4482 (0.0331) | 0.3950 (0.0652) | 0.7085 (0.0645) | 0.5768 (0.0612) | 0.5398 (0.0354) |
| single linear kernel (mixed-effects) | 0.4867 (0.0431) | 0.4960 (0.1678) | 0.5310 (0.2136) | 0.5337 (0.0889) | 0.4985 (0.0706) |
| single radial kernel (fixed-effects) | 0.0591 (0.0163) | 0.9450 (0.0228) | 0.9370 (0.0206) | 0.9373 (0.0208) | 0.9445 (0.0230) |
| single radial kernel (mixed-effects) | 0.0572 (0.0155) | 0.9469 (0.0219) | 0.9387 (0.0208) | 0.9391 (0.0207) | 0.9465 (0.0220) |
| | type B prediction | | | | |
| | misclassification | Sensitivity | Specificity | PPV | NPV |
| logistic regression | 0.0590 (0.0092) | 0.9334 (0.0141) | 0.9485 (0.0122) | 0.9476 (0.0126) | 0.9347 (0.0130) |
| generalized mixed effects regression | 0.0618 (0.0097) | 0.9322 (0.0136) | 0.9440 (0.0149) | 0.9434 (0.0145) | 0.9332 (0.0123) |
| single linear kernel (fixed-effects) | 0.4497 (0.0223) | 0.4093 (0.0397) | 0.6911 (0.0549) | 0.5720 (0.0371) | 0.5388 (0.0251) |
| single linear kernel (mixed-effects) | 0.4676 (0.0237) | 0.5400 (0.0912) | 0.5236 (0.0821) | 0.5312 (0.0310) | 0.5342 (0.0302) |
| single radial kernel (fixed-effects) | 0.0553 (0.0091) | 0.9482 (0.0120) | 0.9414 (0.0122) | 0.9416 (0.0131) | 0.9478 (0.0122) |
| single radial kernel (mixed-effects) | 0.0547 (0.0088) | 0.9487 (0.0121) | 0.9420 (0.0118) | 0.9423 (0.0122) | 0.9484 (0.0121) |

Figure A1: A typical set of simulated data (3-dimensional vector $\boldsymbol{W}$).

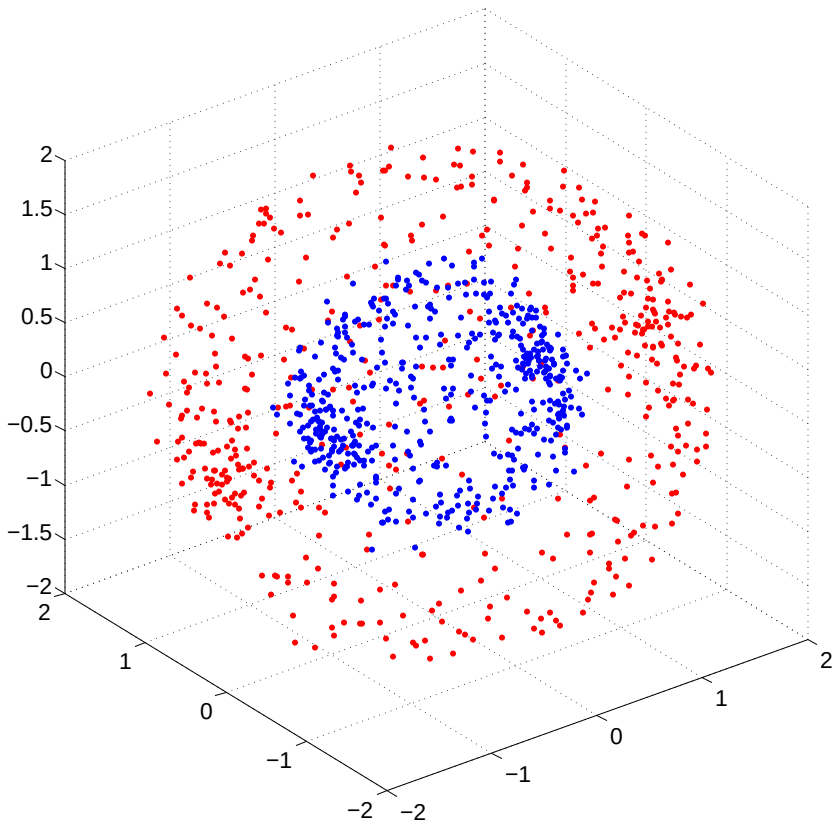

Figure A2: A typical set of simulated data (2-dimensional vector $\boldsymbol{X}$). Top panel: nonlinear boundary in original space. Bottom panel: linear boundary (separating plane) in new 3-dimensional space.

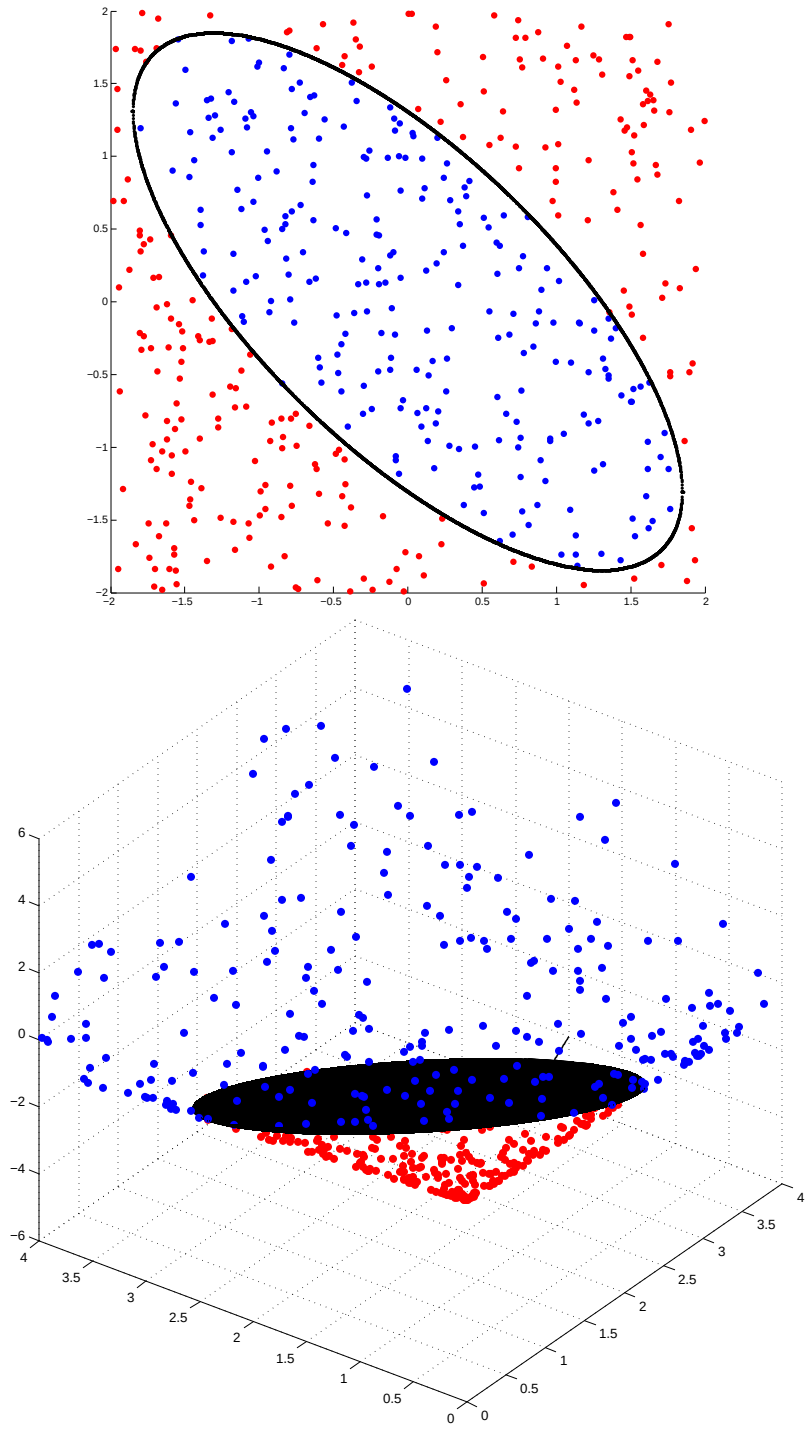

Figure A3: Simulation: single data source. Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2-generalized mixed effects regression, 3-single radial kernel SVM (fixed-effects), 4-single radial kernel SVM (mixed-effects). Bottom panel presents type B prediction of outcomes at future visits on the same subjects (labels same as top the panel).

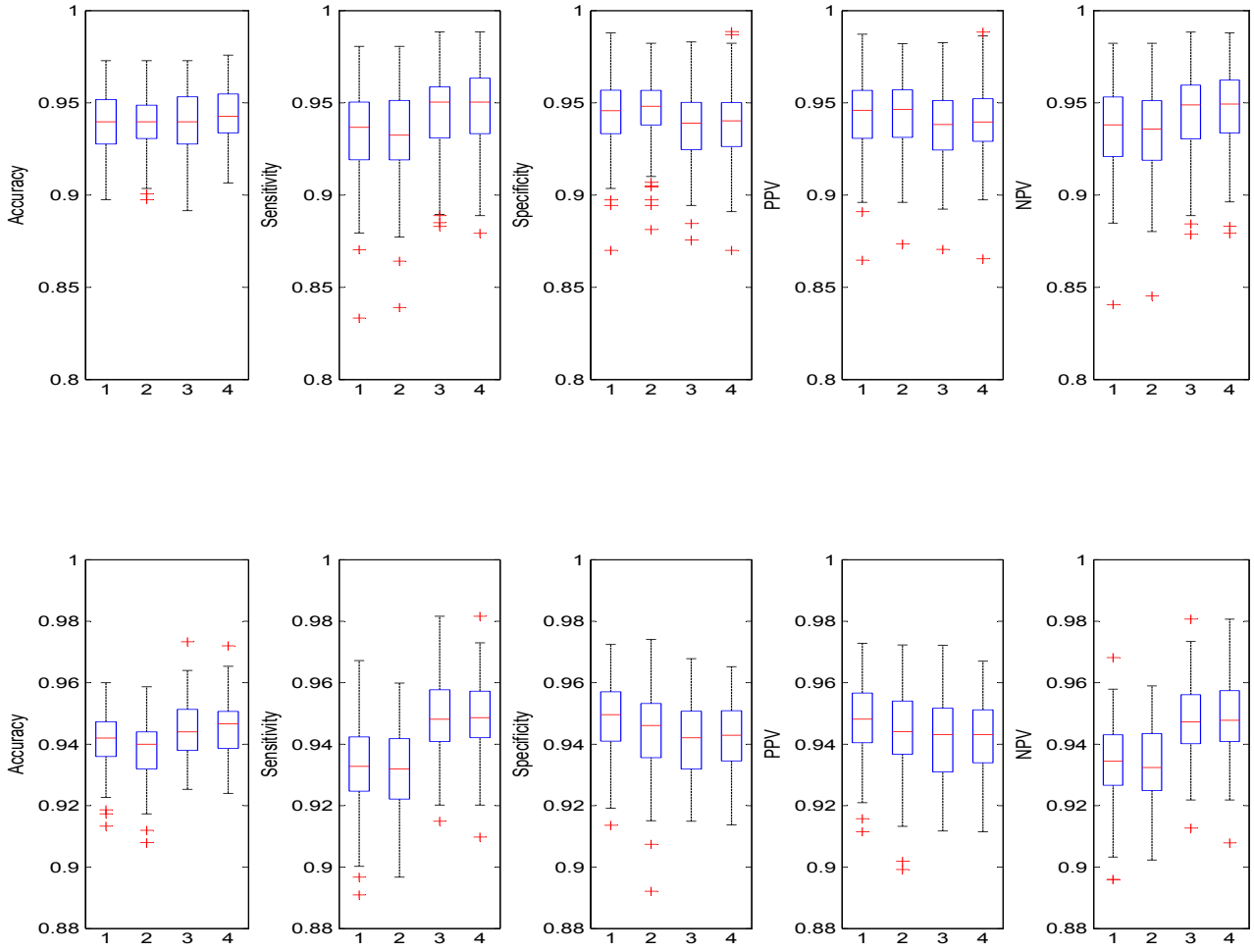

Figure A4: PREDICT-HD study. Top panel presents type A prediction of new subjects (left to right): 1-logistic regression, 2- multiple radial kernel SVM (fixed-effects), 3-multiple fused kernel SVM (fixed-effects), 4-multiple fused kernel SVM (mixed-effects). Bottom panel presents type B prediction of outcomes at future visits on the same subjects (left to right): 1-multiple fused kernel SVM (fixed-effects), 2-multiple fused kernel SVM (mixed-effects).