# Supplementary Material to “Nonparametric variable importance using an augmented neural network with multi-task learning”

**Jean Feng** [*1] **Brian D. Williamson** [*1] **Marco Carone** [1 2] **Noah Simon** [1]

## A. Proof of Lemma 1

*Proof.* For any $\mathcal{S}$, define the augmented conditional function $g_{P_0}(x,m)$ given explicitly by

$$g_{P_0}(x,m) \coloneqq \mu_{P_0}(x)\mathbb{1}\{m=0\} + \sum_{s\in\mathcal{S}} \mu_{P_0,s}(x)\mathbb{1}\{m=e_s\}. \tag{1}$$

Let $\mathcal{E} = \{0\} \cup \{e_s : s \in \mathcal{S}\}$, and let $\tilde{g}_{P_0}(x,m)$ be any continuous function defined over the domain $K \times [-1,2]^p$ that shares the same values as $g_{P_0}(x,m)$ over all $K \times \mathcal{E}$. Using the result of Leshno et al. (1993), there exists a sequence of neural networks $\{f_j\}_{j=1}^\infty \in \mathcal{F}$ with parameters $\{\theta_j\}_{j=1}^\infty \in \Theta$ such that

$$\lim_{j\to\infty} \left\| f_j(x,m;\theta_j) - \tilde{g}_{P_0}(x,m) \right\|_{L^\infty(K\times[-1,2]^p)} = 0.$$

Our desired result follows from the fact that

$$\begin{aligned} &\left\| f_j(x,m;\theta_j) - \tilde{g}_{P_0}(x,m) \right\|_{L^\infty(K\times[-1,2]^p)} \\ &\geq \max_{s\subseteq\mathcal{S}} \left\| f_j(x,e_s;\theta_j) - \mu_{P_0,s}(x) \right\|_{L^\infty(K)} . \end{aligned}$$

□

## B. Experiments on simulated data

Table B.1 displays the neural network structures that we cross-validated over for the non-additive six-variable example in Section 5.1.

For the eight-variable simulation in Section 5.2, we also compare how well we can estimate the conditional means when we use 0 vs the standard normal distribution for the missing inputs $W_s$ in (8). Figure B.1 shows that using random noise results in a much lower multi-task loss (8) over simply using zero. (The minimum loss in this setting is 1, due to the variance of the outcome.) These results were generated using 15 replicates for each training set size.

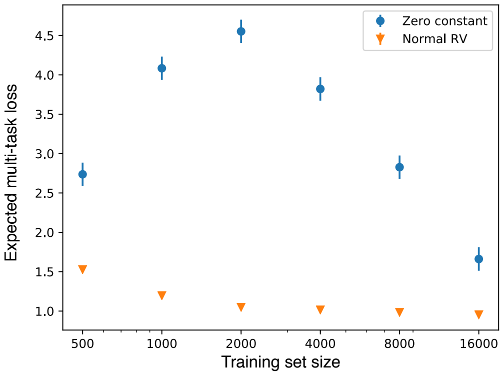

*Figure B.1.* The multi-task loss (8) for the simulation specified by (11) when fitting MTL augmented networks with $W_s \equiv 0$ vs. $W_s \sim N(0,1)$. The points and error bars represent the mean multi-task loss and its 95% confidence interval; the errors bars do not show for the normally distributed inputs since the CI is very narrow.

Additionally, Table B.1 indicates that the time to train a single network is on the same order of magnitude between the multiple networks approach and the augmented network with multi-task learning (MTL) approach. The multiple networks approach may be parallelized, yielding a procedure that estimates all required conditional means on the same time-scale as the augmented network with MTL approach. However, using the multiple networks approach results in a marked increase in time: the network structures for different groups of covariates may be quite different (as seen in Table B.1), and significant user time must be spent finding a set of network structures to cross-validate over. This increase in pre-fitting user time is far larger than the user time spent finding network structures to cross-validate over in the augmented MTL approach.

[*] Equal contribution [1] Department of Biostatistics, University of Washington, Seattle, Washington, USA [2] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. Correspondence to: Jean Feng <jeanfeng@uw.edu>, Brian Williamson <brianw26@uw.edu>.

*Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

| | NN structures to cross-validate over | Time to train single network (sec) | |
|---|---|---|---|
| | | Smallest network | Largest network |
| Multiple networks | Full: 6,40,20,1;6,40,40,1;6,20,20,20,1;6,40,20,20,1 | 82.5 | 106.8 |
| | Reduced $\{x_1, x_2\}$: 4,5,5,1;4,10,5,1 | 57.5 | 60.5 |
| | Reduced $\{x_3, x_4\}$: 4,5,5,1;4,10,5,1 | 57.5 | 60.5 |
| | Reduced $\{x_5, x_6\}$: 4,20,20,20,1;4,40,20,20,1;4,40,40,20,1 | 82.1 | 109.8 |
| Augmented MTL network | 12,40,40,20,1;12,40,40,40,1;12,80,40,40,1 | 133.2 | 161.0 |

*Table B.1.* Network structures used for multiple networks vs the augmented MTL network in the non-additive six-variable example when there are 16000 training observations. Training time for a single network, for the smallest and largest network structures in the cross-validation set, are given in the rightmost columns.

## C. Predicting Mortality of ICU Patients

Here we describe our analysis of the data from the PhysioNet/CinC Challenge 2012 (Silva et al., 2012) in more detail.

We computed summary features based on those proposed in a neural-network submission to the challenge (Xia et al., 2012) and those used to calculate SAPS I and SAPS II scores (Le et al., 1984; Le Gall et al., 1993). Xia et al. (2012) chose to use 18 of the 37 original variables and compute from them a total of 27 features, such as mean, min/max, and the last measurement; their model was then fit on these 27 computed features. We included these 27 computed features as well as the minimum, maximum, and mean (from fitting linear regression) from the time series of the 18 original variables if they were not already included. In addition, we (1) added five variables that are used in SAPS I and SAPS II but were not in this set of 18 original variables and (2) included all general descriptors measured at admission. This procedure resulted in a total of 55 computed and original features in our model (Table C.2).

We estimate the importance of 25 variable groups which fall into two categories: “medical test groups” contain summary features for variables measured by the same medical test and “individual variable groups” contain summary features from the same variable. Here, we discuss our results for the individual variable groups; medical test groups are discussed in the main manuscript.

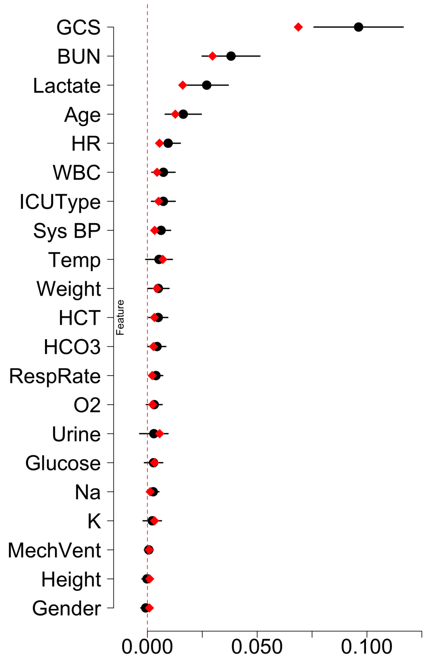

*Figure C.2.* Variable importance estimates for tests in the ICU data (naive = red diamonds; corrected = black circles). Confidence intervals for the true importance, based on the corrected estimator only, are displayed as black bars.

The individual variable groups that we consider are given in the second column of Table C.2. The groups corresponding to GCS, systolic blood pressure (Sys BP), temperature, lactate, heart rate, and urine are all the same as the medical test groups for these variables analyzed in the main manuscript. The individual variables in the metabolic panel medical test group are the summary features of bicarbonate, BUN, sodium, potassium, and glucose. The complete blood count medical test (CBC) group consists of the summary features for white blood cells and hematocrit. The respiration medical test group consists of the summary features for respiration rate, mechanical ventilation, fraction of inspired oxygen, and partial pressure of oxygen. The general descriptors group consists of age, sex, height, weight, and ICU admission type.

We tuned the network structure via an 80/20 training/validation split, and chose layer sizes 110,4,3,2,1 with relu activation functions for the hidden nodes and a sigmoid function for the output. The final variable importance estimates are based on models fit on all the data.

We estimate that among the individual variable groups, the Glasgow Coma Score test has the highest variable importance score by far (Figure C.2). This makes sense as the

| Variable (Meta)-Group | Variable | Summary (computed or original) |
|---|---|---|
| GCS | GCS | last, weighted mean, max, min, slope |
| Metabolic panel | HCO3 | min, max, last, weighted mean |
| | BUN | min, max, last, weighted mean |
| | Na | min, max, weighted mean |
| | K | min, max, weighted mean |
| | Glucose | min, max, weighted mean |
| SysABP | SysABP | min, max, last, weighted mean |
| CBC | WBC | min, max, last, weighted mean |
| | HCT | min, max, weighted mean |
| Temp | Temp | min, max, last, weighted mean |
| Lactate | Lactate | min, max, last, weighted mean |
| HR | HR | min, max, weighted mean |
| Respiration | RespRate | min, max, weighted mean |
| | MechVent | max |
| | FiO2, PaO2 | ratio of means |
| Urine | Urine | sum (based on SAPS II urine item) |
| General Desc. | Gender | measured at admission |
| | Height | measured at admission |
| | Weight | measured at admission |
| | Age | measured at admission |
| | ICU admission type | measured at admission |

*Table C.2.* Features included for analysis of the PhysioNet/CinC Challenge 2012. CBC = complete blood count test. Weighted mean = fit linear regression of response vs. time and get the estimate at the mean measurement time. Slope = fit linear regression of response vs. time and get slope. Last = last measurement. Impossible values were dropped (zero or lower for many of these variables).

Glasgow Coma Score scores the consciousness of a patient and the GCS score can contribute the most number of points to the SAPS II score. Figure C.2 shows that the primary driver of the importance of the metabolic test is blood urea nitrogen (BUN), which assesses kidney function.

## References

Le, JRG, Loirat, P, Alperovitch, A, Glaser, P, Granthil, C, Mathieu, D, Mercier, P, Thomas, R, and Villers, D. A simplified acute physiology score for icu patients. *Critical care medicine*, 12(11):975–977, 1984.

Le Gall, J-R, Lemeshow, S, and Saulnier, F. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24): 2957–2963, 1993.

Leshno, M, Lin, VY, Pinkus, A, and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Silva, I, Moody, G, Scott, DJ, Celi, LA, and Mark, RG. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC), 2012*, pp. 245–248. IEEE, 2012.

Xia, H, Daley, BJ, Petrie, A, and Zhao, X. A neural network model for mortality prediction in icu. In *Computing in Cardiology (CinC), 2012*, pp. 261–264. IEEE, 2012.