OUT-OF-DOMAIN TRANSFER PERFORMANCE OF RANDOM FORESTS VS. TRADITIONAL ECONOMIC MODELS (A SKETCH)

ANDRÉS ARADILLAS FERNÁNDEZ

"A little learning is a dangerous thing; Drink deep, or taste not the Pierian spring." - Alexander Pope.

1. Premise

Recent work in the intersection of econometrics and machine learning has resulted in questions surrounding the relative cost of re-estimation of parameters ("transfer performance"). This paper proposal is inspired by Andrews et al. (2022), which defines various notions of transfer performance and then compares traditional prediction methods (e.g., estimation of parameters in an economic model) versus pure "blackbox" methods. The empirical findings of the paper are that these blackbox machine learning methods have far worse transfer performance than economic models. We aim to (roughly) replicate this paper and see if we attain the same results. However, we will define my own measure of transfer performance loosely based on those defined in Andrews et al. (2022). Furthermore, the setup/prediction problem will be similar to that of Andrews et al. (2022), but the details will be different, and the definition of "domains" will be slightly different. Nevertheless, the "spirit" of a domain will remain the same.

The hope of such a work would be two-fold. First of all, the authors of Andrews et al. (2022) make a claim that out-of-domain transfer performance tends to be worse for blackbox machine learning methods when compared to traditional economic models (e.g., a parametrization of expected utility). While the original work has some empirical results adding towards this claim, further results would give more credence to this claim. Another goal of this work would be to see how much the implementation of a particular transfer performance measure affects the findings. While the measures we present here are certainly inspired and follow the same logic of those presented in Andrews et al. (2022), they are nevertheless mathematically different. Furthermore, the third measure presented in Section 3, a so-called "similarity-weighted transfer performance measure" could potentially reduce the gap in out-of-domain performance of machine learning models versus classical economic models. If the problem with blackbox methods lies in overfitting, then similarly-shaped domains should also reap most of the rewards that overfitting brings in terms of prediction error reduction, and so it could be possible that machine learning models actually outperform traditional estimation methods with respect to this transfer measure.

Further extensions of this work could look at completely different sorts of prediction problems. Perhaps even prediction problems that would be related to using LASSO and Ridge regression versus vanilla Ordinary-Least-Squares (OLS) to estimate and predict some data. However, this particular prediction problem is of interest because results are already presented in Andrews et al. (2022), and so replication of the problem, especially with regards to a novel measure presented in Section 3 (T_{sim}), is also of interest.

2. The Prediction Problem

Here, the prediction problem will consist of predicting the certainty equivalent of lotteries. Each lottery L_i consists of three parts $L_i = (z_1, z_2, p)$ where z_1 is earned with probability p and z_2 is earned with probability $1-p$. The data will be sectioned off into 20 "domains," with each domain d having a certain distribution $P_{\theta_d}(\eta)$ unknown to the econometrician. θ_d here parametrizes the distribution for domain d, which will be assumed to be governed by a CRRA utility function $\sigma_\eta(z_1, z_2, p)$ to be described in Section 4.1. The details of $P_{\theta_d}(\eta)$ will be expanded upon in Section 5.1.

For each individual $j \in \{1, ..., n_d\}$ within a domain d, they are endowed with a true risk aversion parameter $\eta_{j,d}$ governing their choice, and the associated certainty equivalent $y_{j,d}(L_i)$ for a lottery L_i is determined from $y_{j,d}(L_i) = \sigma_{\eta_{j,d}}(L_i) + u_{j,d,i}$ where $u_{j,d,i}$ is some i.i.d. noise drawn from an entirely independent distribution. The precise distribution of the noise will be clarified in Section 5.

While it is not essential to the estimation methods, for simplicity, all randomly simulated individuals will be exposed to the same set of lotteries across domains. However, these simulations could be redone with each set of lotteries $\mathbb{L}_d$ being unique to a particular domain. This reflects the sort of empirical lottery choice data used in Andrews et al. (2022), where most of the different domains consist of data generated by presenting different sets of lotteries to participants. However, there is not much reason to suspect that this would significantly alter any results regarding transfer performance differences between machine learning and traditional prediction methods.

3. Transfer Performance Measure

Consider the problem of out-of-sample performance. Suppose we pick some fixed domain $d_0 \in \{1, ..., 20\}$ that contains the data with which we will fit some model $\hat{m}(L_i; \hat{\theta}_{(d_0)})$ (where $\hat{m}(\cdot)$ denotes an arbitrary model with estimated parameter $\hat{\theta}_{(d_0)}$ fitted on domain d_0). This could be as simple as estimating an expected utility function with a single parameter or fitting some blackbox machine learning model (e.g., Random Forest, Neural Network, etc.). Of special interest will be how this fitted model performs in the various other domains.

One measure that seems reasonable would be an equally-weighted average of the square loss. This can be called $T_{avg}(\hat{\mu})$:

$$T_{avg}(\hat{\mu}) \equiv \frac{1}{19} \sum_{\substack{d \in \{1,\dots,20\}, \\ d \neq d_0}} \frac{1}{n_d} \sum_{j=1}^{n_d} \sum_{L_i \in \mathbb{L}} \left(y_{j,d}(L_i) - \hat{m}(L_i; \hat{\theta}_{(d_0)}) \right)^2$$

Another notion that is worth considering would be "worst-case" out-of-sample performance, for which a candidate measure could be:

$$T_{wc}(\hat{\mu}) \equiv \max_{\substack{d \in \{1,\dots,20\} \\ d \neq d_0}} \frac{1}{n_d} \sum_{j=1}^{n_d} \sum_{L_i \in \mathbb{L}} \left(y_{j,d}(L_i) - \hat{m}(L_i; \hat{\theta}_{(d_0)}) \right)^2$$

Both of these candidate measures act under the assumption that a mistake in any domain is "equally as bad." However, we might have some signal of how "similar" two domains are to one another (e.g., some distance metric between the underlying η_d values for two domains). Consider having a normalized distance metric $s(d_1, d_2)$ that calculates the normalized similarity score of domains d_1 and d_2 (in relation to their underlying parameters η_{d_1} and η_{d_2}). Then we have the measure

$$T_{sim}(\hat{\mu}) = \frac{1}{19} \sum_{\substack{d \in \{1,\dots,20\} \\ d \neq d_0}} \frac{s(d, d_0)}{n_d} \sum_{j=1}^{n_d} \sum_{L_i \in \mathbb{L}} \left(y_{j,d} - \hat{\mu}(L_i; \hat{\theta}_{d_0}) \right)^2$$

For instance, we may have $s(d, d_0) \propto (||\eta_{d_1} - \eta_{d_2}||^2)^{-1}$. This particular similarity metric seems reasonable, as it will make small differences have a larger weight, placing greater emphasis on the squared prediction performance of domains similar to d_0. Another sort of metric could also be introduced where a similar parabolic decay occurs for domains less similar to d_0 until after a certain threshold, domains fall out of consideration.

All three of these out-of-sample performance measures will be used to compare the models of interest. We believe that of particular interest, however, will be the third measure. We think this not only because it introduces an extension to ideas of out-of-domain transfer performance presented in Andrews et al. (2022), but also because it may actually give favorable results to blackbox machine learning models that may suffer from overfitting and overtraining on data, as domain robustness is not as important when prediction error performance is weighted towards stochastically-similar domains.

4. Estimation Models of Interest

In the interest of novelty, we will attempt to provide results for the following estimation methods: Expected Utility and Random Forest (RF).

4.1. **Expected Utility.**

For simplicity, we will adopt the same CRRA utility function as in Andrews et al. (2022), which is defined as, for $\eta \geq 0, \eta \neq 1$:

$$v_\eta(z) = \begin{cases} \frac{z^{1-\eta}-1}{1-\eta} & \text{if } z \geq 0 \\ -\frac{(-z)^{1-\eta}-1}{1-\eta} & \text{if } z < 0 \end{cases}$$

and for $\eta = 1$, we have

$$v_\eta(z) = \begin{cases} ln(z) & \text{if } z > 0 \\ -ln(-z) & \text{if } z < 0 \end{cases}$$

For a given lottery (z_1, z_2, p), the predicted expected utility can then be described as

$$\sigma_\eta(z_1, z_2, p) = v_\eta^{-1}(p \cdot v_\eta(z_1) + (1-p) \cdot v_\eta(z_2))$$

The $\hat{\eta}$ obtained will be a maximum likelihood estimate given data. It is $\sigma_\eta(z_1, z_2, p)$ that will give the certainty equivalent for a given lottery $L_i = (z_1, z_2, p)$.

4.2. **Random Forest.** Put simply, a random forest consists of embedding various decision trees within a classification model. Andrews et al. (2022) utilize a random forest within their analysis of lottery choice. Therefore, we too will use a random forest to give a tractable comparison between the out-of-domain transfer performance (under our measures) of machine learning models versus more traditional prediction methods. As is discussed in Andrews et al. (2022), there are at least two cross-validation methods available to us with regards to the size parameter for random forests. However, these methods generally provide an edge only when multiple training domains are being used. Therefore, since we will restrict training to a single domain d_0, cross-validation does not make sense to utilize in this situation.

One key feature of random forests that is especially important is that oftentimes, despite the fact that they consist of multiple decision trees embedded within one another, they can be difficult to make sense of in context. That is to say, they have a "blackbox" to them where the random forest may be conducting some sort of overfitting on features that may prove to be disastrous when considering out-of-domain data. Segal (2004) is an empirical example where some degree of overfitting. However, as both Andrews et al. (2022) and Hastie et al. (2009) discuss, overfitting is not of particularly great concern in random forests, as they tend to be quite good at being robust to irrelevant features. Nevertheless Andrews et al. (2022) do find unfavorable out-of-domain prediction results for random forests, and so the hope with this replication is to find similar results, albeit with our slightly different transfer performance measures.

5. Monte Carlo Results

5.1. **Simulated Data.** To simulate the data, we will need to impose additional distributional assumptions on this example. For the CRRA utility function described in Section 4.1 to work, we will need for $\eta \geq 0$. Therefore, we impose that $\eta_{j,d} \overset{i.i.d}{\sim} \Gamma(\theta_d, 1)$. This ensures that $\eta_{j,d}$ is nonnegative, while also being continuously distributed, allowing for "similar" and "dissimilar" (under $s(\cdot)$) domains. For simplicity, the Gamma distribution has been restricted to a single parameter.

For tractability, a particular similarity metric $s(\cdot)$ must be picked. It makes sense for us to weight similar domains more, and so define $s(\cdot)$ as follows:

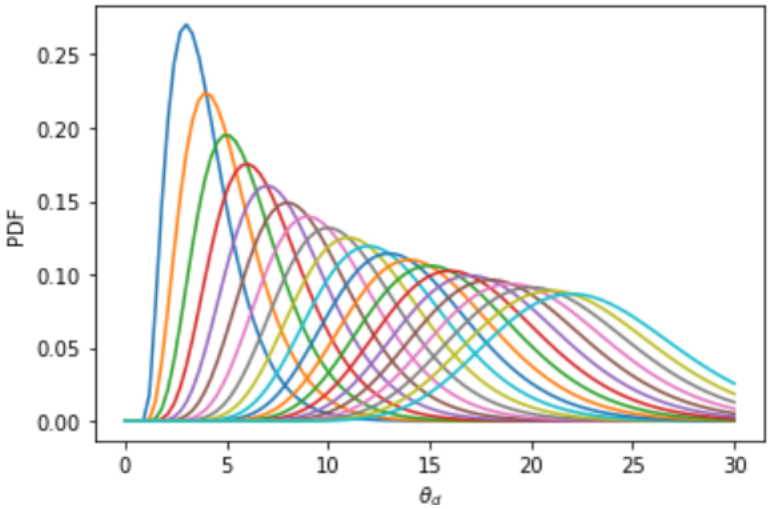
$$s(d, d_0) \equiv \frac{(||\theta_d - \theta_{d_0}||^2)^{-1}}{\sum_{\substack{d \in \{1,...,20\} \\ d \neq d_0}} (||\theta_d - \theta_{d_0}||^2)^{-1}}$$

We then generate data in the following way:

(1) Fix a collection of lotteries $\mathbb{L}$.
(2) For each individual j within each domain d, draw $\eta_{j,d} \overset{i.i.d.}{\sim} \Gamma(\theta_d, 1)$ and keep it fixed.
(3) Generate $y_{j,d}(L_i) = \sigma_{\eta_{j,d}}(L_i) + u_{j,d,i}$ where $u_{j,d,i} \overset{i.i.d.}{\sim} N(0, 0.25)$

Then, to tally up all of the three transfer performance across domains, we can cycle through all of $\{d_1, ..., d_{20}\}$ as our choice of d_0, creating graphics similar to Figure 3 in Andrews et al. (2022).

Consider the following graph showing the 20 different domains that each individual's risk aversion parameter $\eta_{j,d}$ is drawn from:



To generate this graph, the value of θ_d ranged from 3 all the way up to 23. The idea, then, is that for a given θ_d within this range, some number n_d of participants will be simulated by drawing i.i.d. $\eta_{j,d}$, $j = 1, ..., n_d$, where the risk aversion plus some completely independent normal noise slightly alters the certainty equivalent predicted by $\sigma_{\eta_{j,d}}(L_i)$ for a given lottery $L_i \in \mathbb{L}$.

References

Andrews, I., Fudenberg, D., Lei, L., Liang, A., and Wu, C. (2022). The transfer performance of economic models. *arXiv preprint arXiv:2202.04796*.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J. (2009). Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, pages 587–604.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.