Validation of probabilistic classifiers

Tim Hillel
Michel Bierlaire
Mohammed Elshafie
Ying Jin

EPFL, University of Cambridge

May 2018

STRC

18th Swiss Transport Research Conference
Monte Verità / Ascona, May 16 – 18, 2018

EPFL, University of Cambridge

Validation of probabilistic classifiers

Tim Hillel, Michel Bierlaire
ENAC IIC TRANSP-OR
EPFL
CH-1015 Lausanne
phone: +41-21-693 24 08
fax: +41-21-693 80 60
{tim.hillel,michel.bierlaire}
@epfl.ch

Tim Hillel, Mohammed Elshafie
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ
phone: +44-1223-332 600
fax: +44-1223-332 662
{th389,me254}@cam.ac.uk

Ying Jin
Department of Architecture
University of Cambridge
1 Scroope Terrace,
Cambridge CB2 1PX
phone: +44-1223-332 950
fax: +44-1223-332 960
ying.jin@aha.cam.ac.uk

May 2018

Abstract

Non-parametric probabilistic classification models are increasingly being investigated as an alternative to Discrete Choice Models (DCMs), e.g. for predicting mode choice. There exist many strategies within the literature for model selection between DCMs, either through the testing of a null hypothesis, e.g. likelihood ratio, Wald, Lagrange Multiplier tests, or through the comparison of information criteria, e.g. Bayesian and Aikaike information criteria. However, these tests are only valid for parametric models, and cannot be applied to non-parametric classifiers.

Typically, the performance of Machine Learning classifiers is validated by computing a performance metric on out-of-sample test data, either through cross validation or hold-out testing. Whilst bootstrapping can be used to investigate whether differences between test scores are stable under resampling, there are few studies within the literature investigating whether these differences are significant for non-parametric models.

To address this, in this paper we introduce three statistical tests which can be applied to both parametric and non-parametric probabilistic classification models. The first test considers the analytical distribution of the expected likelihood of a model given the true model. The second test uses similar anaylsis to determine the distribution of the Kullback-Leibler divergence between two models. The final test considers the convex combination of two classifiers under comparison. These tests allow ML classifiers to be compared directly, including with DCMs.

Keywords

Discrete choice models, machine learning, significance testing

1 Introduction

Probabilistic classification models, which predict a probability distribution over a set of classes from a given input, are found in many applications, including behavioural modelling, disease detection, image recognition, document analysis, and fraud and spam classification (Hastie *et al.,* 2008). Discrete choice models (DCMs) (Ben-Akiva *et al.,* 1985*,* Ben-Akiva and Bierlaire*,* 2003) are a class of parametric probabilistic classification models, which have seen extensive use in the field of transportation in particular. With recent advances in the abundance, scale, and depth of data available in many applications, including transportation, non-parametric machine learning models are increasingly being investigated as an alternative to parametric models.

Selection of a suitable model is an essential task for any application of probabilistic classifiers. There exist many strategies within the literature for model selection for parametric models. This includes both hypothesis tests e.g. likelihood ratio, Wald, Lagrange Multiplier tests, as well as the comparison of information criteria, e.g. Bayesian and Aikaike information criteria (Ben-Akiva *et al.,* 1985*,* Akaike*,* 1998). However, as these tests rely on the analysis of model parameters, they cannot be applied to non-parametric classifiers.

In this paper we present three statistical tests which can be applied to both parametric and non-parametric probabilistic classification models. First we introduce the theoretical background for probabilistic classification and model validation. In the next section, we present an overview of the three statistical tests. Next, we introduce an experimental methodology to validate the proposed tests on probabilistic classification models. Finally, we present the results from the initial applications.

2 Theoretical background

Consider a set C containing N_P elements, called the *population*. C is supposed to be sufficiently large that it is not feasible to enumerate its elements explicitly. We also consider a partition composed of J subsets C_i, $i = 1, \ldots, J$, which we call *classes*. We have

$$C = \cup_{i=1}^{J} C_i, \tag{1}$$

and

$$C_i \cap C_j = \emptyset, \forall i \neq j. \tag{2}$$

Each element $n \in C$ is associated with a vector of K features $x_n \in \mathbb{Z}^K$. We assume the features to be discrete to simplify the following development. However, it is straightforward to extend it with data sets containing both discrete and continuous variables. A *probabilistic classifier* is a model which maps the vector of features x_n into a probability distribution on the classes:

$$P : \mathbb{Z}^K \to [0, 1]^J. \tag{3}$$

We use the notation $P(i|x_n)$ to represent the probability that element n belongs to class C_i, as provided by the classifier. We have

$$P(i|x_n) \geq 0, i = 1, \ldots, J \text{ and } \sum_{i=1}^{J} P(i|x_n) = 1, \forall x_n \in \mathbb{Z}^K. \tag{4}$$

For example, consider travelers who choose between the car and public transportation to commute to work, so that there are $J = 2$ classes. The set C is the population of travelers in a specific city on a specific day, partitioned into those who travel by car (C_1), and those who travel by public transportation (C_2). The choice of traveler n in population C can be explained by features such as the travel time, the travel cost, the weather conditions, the purpose of the trip, etc. Note that qualitative and categorical variables can always be modeled numerically. These features form the vector x_n. The choice model provides the probability $P(1|x_n)$ that traveler n chooses to travel by car, and $P(2|x_n)$ the probability that they travel by public transportation.

2.1 Validation set

We have at our disposal a *validation* set containing N elements, $1 \leq N \leq N_P$, which can be enumerated. The validation set is separate from the data used to train or fit the classifier. Each n in the validation set is associated with

1. a vector of features x_n, and,
2. a set $y_n \in \{0, 1\}^J$ of class indicators, such that

$$\sum_{i=1}^{J} y_{in} = 1. \tag{5}$$

We denote the validation set $V = (x_n, y_n)_{n=1}^N = (x_V, y_V)$.

The class indicators y_n are themselves drawn from an unknown probability distribution, which

we call the *true model*. For example, within the context of choosing between the car and public transportation for a work commute, an individual selects the mode from an unknown distribution which is dependent on the observed features x_n, as well as unobserved features not captured in the validation set. Each element n has been independently sampled from the population C with probability

$$\Pr(x_n, y_n) = \Pr(y_n|x_n)\Pr(x_n) = \prod_{i=1}^{J} P^*(i|x_n)^{y_{in}} \Pr(x_n), \tag{6}$$

where P^* is the true model which has been involved in the data generation process, and $\Pr(x_n)$ is the probability to find the vector x_n of features in the population. It can be convenient to denote i_n the index of the class associated with element n. Because of (5), it is defined as

$$i_n = \sum_{i=i}^{J} iy_{in}. \tag{7}$$

In this case, we can write (6) as

$$\Pr(x_n, y_n) = P^*(i_n|x_n)\Pr(x_n). \tag{8}$$

2.2 Validation process

Suppose that we have a collection of M different classifiers P^m, $m = 1, \ldots, M$. We want to use the validation set to evaluate the performance of each classifier. Consider classifier P^m. For each element n in the validation set a measure of fit $d_n^m(x_n, y_n)$ or $d_n^m(x_n, i_n)$ which measures how well or how poorly the classifier P^m is able to predict the true class membership $P^*(i|x_n)$ when using x_n as an input.

The validation set contains only realisations y_n of the true class membership model. For element n, the probability that a model P^m correctly predicts y_n is defined as

$$\prod_{i=1}^{J} P^m(i|x_n)^{y_{in}}, \tag{9}$$

or, equivalently

$$P^m(i_n|x_n). \tag{10}$$

An aggregate measure of fit measures the overall performance of classifier m on the dataset. The quantity

$$L^m = \sum_{n=1}^{N} P^m(i_n|x_n). \tag{11}$$

is the expected number of "true positives", that is the expected number of times that the classifier P^m correctly predicts the observed class. The quantity

$$\prod_{n=1}^{N} P^m(i_n|x_n), \tag{12}$$

is the likelihood of the validation set for classifier P^m, that is the probability that the classifier correctly predict all observed classes. In practice, it is more convenient to consider the natural logarithm of the likelihood, and use the quantity

$$\mathcal{L}^m = \sum_{n=1}^{N} \ln P^m(i_n|x_n). \tag{13}$$

Therefore, $d_n^m(x_n, y_n)$ can be defined as

$$d_n^m(x_n, i_n) = \ln P^m(i_n|x_n), \tag{14}$$

or

$$d_n^m(x_n, i_n) = P^m(i_n|x_n). \tag{15}$$

The rest of the discussion is based on (14). A similar derivation can be obtained for (15).

The quantity $d_n^m(x_n, i_n)$ is a random variable. Each realisation corresponds to a different element in the validation set. The expected value of d_n^m is defined as

$$\mathrm{E}[d_n^m] = \sum_{x} \sum_{i=1}^{J} \ln P^m(i|x) P^*(i|x) \Pr(x_n), \tag{16}$$

where the first sum scans all the possible vectors of features in the population, therefore accounting for the variability of d_n^m due to the sampling of the feature vectors in the validation set, and the second sum accounts for the variability of d_n^m due to generating the class indicators from an unknown distribution P^*.

In practice, it is infeasible to calculate the first sum, and even to obtain a good approximation of

$\Pr(x)$. It is therefore convenient to consider the conditional mean for a given element n:

$$\mathrm{E}[d_n^m|x_n] = \sum_{i=1}^{J} \ln P^m(i|x_n) P^*(i|x_n). \tag{17}$$

This is equivalent to the negative cross-entropy loss of the true model P^* with the classifier P^m: $-H_n(P^*, P^m)$.

As the elements of the validation set have been drawn independently, the total conditional mean is therefore

$$\mathrm{E}[\mathcal{L}^m|x_V] = \sum_{n=1}^{N} \mathrm{E}[d_n^m|x_n] = \sum_{n=1}^{N} \sum_{i=1}^{J} \ln P^m(i|x_n) P^*(i|x_n). \tag{18}$$

The conditional variance is

$$\begin{aligned}
\mathrm{Var}[d_n^m|x_n] &= \mathrm{E}[(d_n^m)^2|x_n] - (\mathrm{E}[d_n^m|x_n])^2 \\
&= \sum_{i=1}^{J} (\ln P^m(i|x_n))^2 \, P^*(i|x_n) - \left(\sum_{i=1}^{J} \ln P^m(i|x_n) P^*(i|x_n) \right)^2 \\
&= \sum_{i=1}^{J} (\ln P^m(i|x_n))^2 \, P^*(i|x_n) \, (1 - P^*(i|x_n)) \\
&\quad - 2 \sum_{i<j} \ln P^m(i|x_n) P^*(i|x_n) \ln P^m(j|x_n) P^*(j|x_n).
\end{aligned} \tag{19}$$

The total conditional variance is

$$\begin{aligned}
\mathrm{Var}[\mathcal{L}^m|x_V] &= \sum_{n=1}^{N} \left(\sum_{i=1}^{J} (\ln P^m(i|x_n))^2 \, P^*(i|x_n) - \left(\sum_{i=1}^{J} \ln P^m(i|x_n) P^*(i|x_n) \right)^2 \right) \\
&= \sum_{n=1}^{N} \left(\sum_{i=1}^{J} (\ln P^m(i|x_n))^2 P^*(i|x_n)(1 - P^*(i|x_n)) \right. \\
&\quad \left. -2 \sum_{i<j} \ln P^m(i|x_n) P^*(i|x_n) \ln P^m(j|x_n) P^*(j|x_n) \right).
\end{aligned} \tag{20}$$

This analysis accounts for the stochasticity of the validation set due to the fact that the true model P^*, may generate different outcomes y_n for the same value of x_n. The variability due to x is ignored, but can be numerically estimated using bootstrapping.

(14) can be generated from (17) and (19) by substituting in $P^* = y_n$, i.e. by assuming P^* is a discrete probability distribution, with the indicated class having probability 1, and all

other classes having probability 0. This results in a zero variance value of $\ln P^m(i_n|x_n)$. This is equivalent to the negative cross-entropy loss of the class labels y_n with the classifier P^m: $-H_n(y_n, P^m)$.

2.3 Simple example

Consider a simple example with $J = 2$ classes, where the true model is defined as

$$P^*(1|x) = p^* \text{ and } P^*(2|x) = 1 - p^*, \forall x. \tag{21}$$

The model to be validated is defined similarly as

$$P^m(1|x) = p^m \text{ and } P^m(2|x) = 1 - p^m, \forall x. \tag{22}$$

Therefore, (17) is

$$p^* \ln p^m + (1 - p^*) \ln(1 - p^m), \tag{23}$$

and (19) is

$$\Big((\ln p^m)^2 + (\ln(1 - p^m))^2 - 2 \ln(p^m) \ln(1 - p^m)\Big) p^*(1 - p^*). \tag{24}$$

For instance, Figure 1 represents the mean as a function of p^m when $p^* = 0.2$. A similar representation for $p^* = 0.5$ is represented in Figure 2, and in Figure 3 for $p^* = 0.01$.

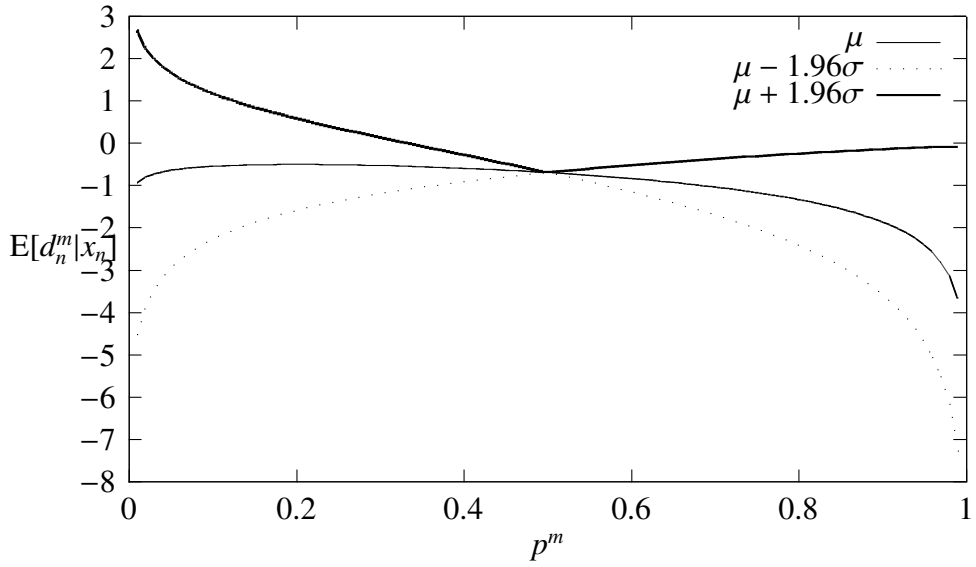

Figure 1: $p^* = 0.2$

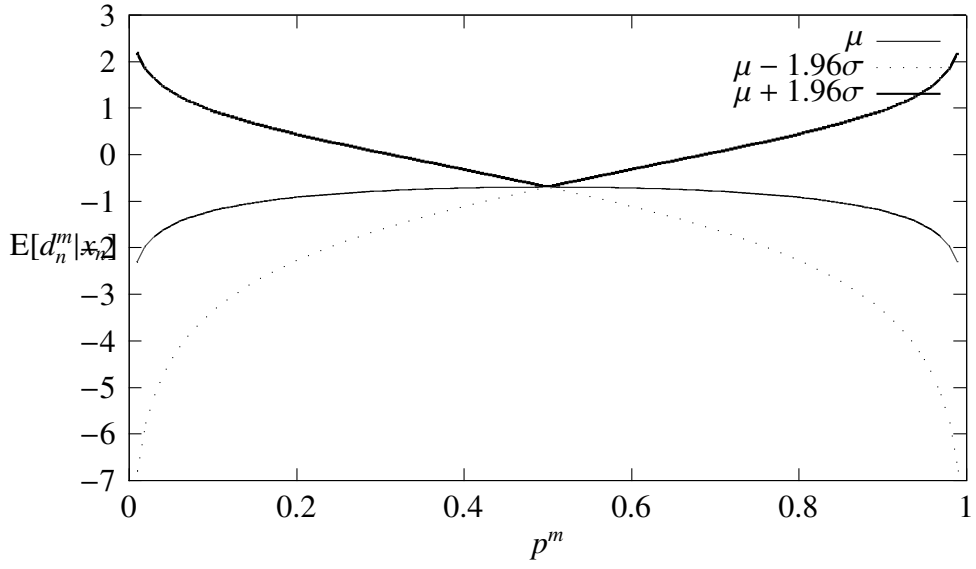

Figure 2: $p^* = 0.5$

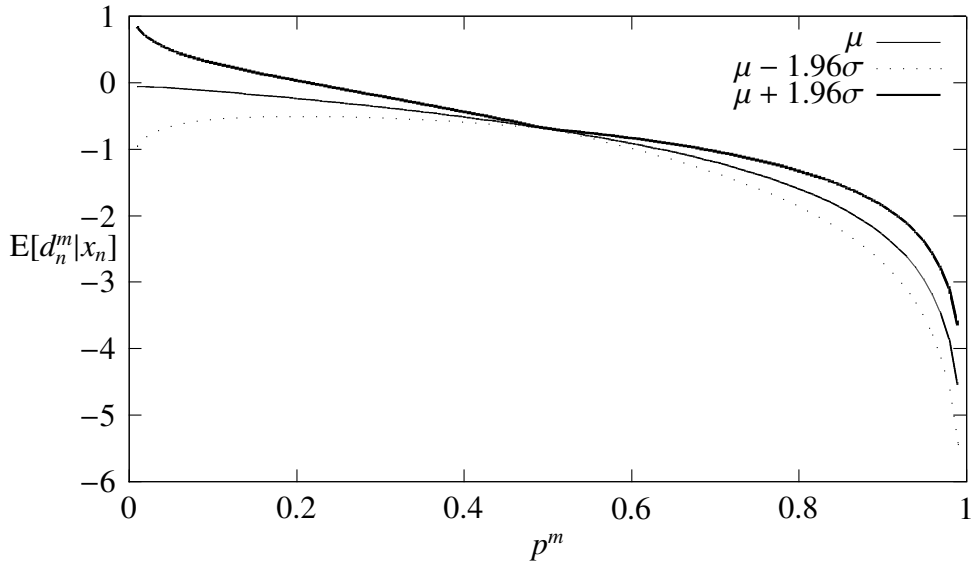

Figure 3: $p^* = 0.01$

2.4 Kullback-Leibler divergence

The relative divergence between one probability distribution and another can be measured using the Kullback-Leibler (KL) divergence, or relative entropy (Kullback and Leibler, 1951). For two classifiers P^r and P^m, the KL-divergence for each element in the validation set is given by

$$d_{KL}(P^r||P^m) = \sum_{i=1}^{J} \ln\left(\frac{P^r(i|x_n)}{P^m(i|x_n)}\right) P^r(i|x_n) \tag{25}$$

The KL-divergence is a positive real value which is 0 when the two probability distributions P^r and P^m are equivalent. Note that the KL-divergence is non-symmetric, and one model (P^r) is the reference model against which the other is compared.

(25) is analogous to (17), when substituting in $P^m = \frac{P^r}{P^m}$ and $P^* = P^r$. As such, we can substitute these values into (18) and (20) to obtain the mean and variance of the total KL-divergence $\mathcal{D}(P^r||P^m)$:

$$\mathrm{E}[\mathcal{D}(P^r||P^m)] = \sum_{n=1}^{N} \mathrm{E}[d_{KL}(P^r||P^m)] = \sum_{n=1}^{N}\sum_{i=1}^{J} \ln\left(\frac{P^r(i|x_n)}{P^m(i|x_n)}\right) P^r(i|x_n) \tag{26}$$

$$\begin{aligned}
\mathrm{Var}[\mathcal{D}(P^r||P^m)] &= \sum_{n=1}^{N}\left(\sum_{i=1}^{J}\left(\ln\left(\frac{P^r(i|x_n)}{P^m(i|x_n)}\right)\right)^2 P^r(i|x_n) - \left(\sum_{i=1}^{J}\ln\left(\frac{P^r(i|x_n)}{P^m(i|x_n)}\right)P^r(i|x_n)\right)^2\right) \\
&= \sum_{n=1}^{N}\left(\sum_{i=1}^{J}\left(\ln\left(\frac{P^r(i|x_n)}{P^m(i|x_n)}\right)\right)^2 P^r(i|x_n)(1-P^r(i|x_n))\right. \\
&\quad \left. -2\sum_{i<j}\ln P^m(i|x_n)P^r(i|x_n)\ln\left(\frac{P^r(i|x_n)}{P^m(i|x_n)}\right)P^r(j|x_n)\right).
\end{aligned} \tag{27}$$

3 Statistical tests

3.1 Single true-model test

Suppose now that we consider the classifier P^m. We want to test the hypothesis H_0 that it is the true model that has generated the data:

$$H_0 : P^m = P^* \tag{28}$$

Under this assumption, the statistic

$$\mathcal{L}^m = \sum_{n=1}^{N} \ln P^m(i_n|x_n) \tag{29}$$

is normally distributed with mean

$$\mu_m = \sum_{n=1}^{N} \sum_{i=1}^{J} P^m(i|x_n) \ln P^m(i|x_n) \tag{30}$$

and variance

$$\begin{aligned} \sigma_m^2 = & \sum_{n=1}^{N} \sum_{i=1}^{J} (\ln P^m(i|x_n))^2 P^m(i|x_n)(1 - P^m(i|x_n)) \\ & -2 \sum_{i<j} \ln P^m(i|x_n) P^m(i|x_n) \ln P^m(j|x_n) P^m(j|x_n). \end{aligned} \tag{31}$$

Equivalently, the statistic

$$\frac{\mathcal{L}^m - \mu_m}{\sigma_m} \sim N(0, 1). \tag{32}$$

Consider again the example from Section 2.3. Suppose that the true model corresponds to $p^* = 0.2$. Figure 4 represents the value of the statistic (32) for different values of p^m when the sample contains exactly the proportion of observed classes corresponding to the true model. As such, the log likelihood (29) of the validation set is given by

$$\mathcal{L}^m = N(p^* \ln P^m + (1 - p^*) \ln(1 - P^m)). \tag{33}$$

In this specific case, it is difficult to reject the hypothesis that P^m is the true classifier, even when it is actually very different from the true one.

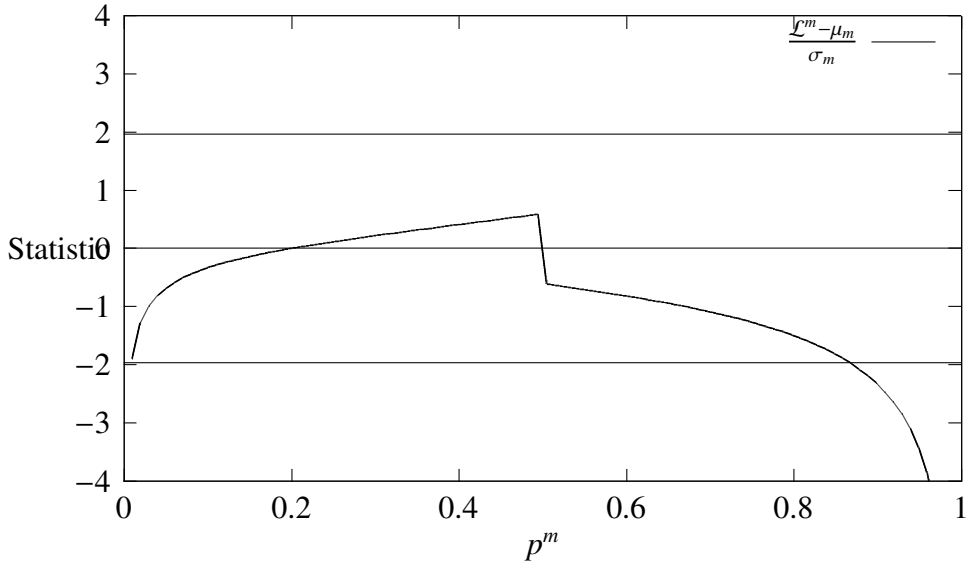

Figure 4: $p^* = 0.2$

3.2 Comparing two classifiers

3.2.1 Kullback-Leibler divergence test

We consider two classifiers P^m and P^r. We want to test the hypothesis H_0 that the candidate model P^m is equivalent to the reference model P^r

$$H_0 : P^m = P^r \tag{34}$$

For this assumption to hold, the total KL-divergence $\mathcal{D}(P^r||P^m)$ must be 0. As in (32)

$$\frac{\mathcal{D}(P^r||P^m) - \mu_{r||m}}{\sigma_{r||m}} \sim N(0,1) \tag{35}$$

with mean $\mu_{r||m}$ and variance $\sigma^2_{r||m}$ given in (26) and (27) respectively. As such, we can test the equivalent assumption

$$H_0 : \mathcal{D}(P^r||P^m) = 0 \tag{36}$$

with a t-test. Whilst this test determines if the two models are equivalent, it gives no indication of respective model performance, as it is independent of the class labels y_n. If two models are significantly different under this test, $\mathcal{L}$ can then be used to differentiate performance.

3.2.2 Convex combination of classifier test

Again, we consider two classifiers P^m and P^r. We need to decide which one is performing the best on the validation set V. To do so, we define a third classifier as the convex combination of the two classifiers:

$$P(i|x_n;\lambda) = \lambda P^r(i|x_n) + (1-\lambda)P^m(i|x_n), \text{ where } 0 < \lambda < 1. \tag{37}$$

λ can be estimated using maximum likelihood estimation with the validation set. Maximum likelihood estimation allows an estimate of the variance to be obtained from the Hessian. This allows for hypothesis testing against expected values of λ using a t-test. If P^r is the true model the true value of λ is 1. If P^m is the true model, the true value of λ is 0. For intermediate values of λ, the convex combination of the two classifiers is superior to each individual classifier in terms of log-likelihood fit.

The log likelihood of the validation set is

$$\begin{aligned}\mathcal{L}(\lambda) &= \sum_{n=1}^{N}\sum_{i=1}^{J} y_{in} \ln P(i|x_n;\lambda) \\ &= \sum_{n=1}^{N}\sum_{i=1}^{J} \ln(\lambda P^m(i_n|x_n) + (1-\lambda)P^r(i_n|x_n)).\end{aligned} \tag{38}$$

4 Experimental methodology

Parametric DCM models are used in order to assess the suitability of the proposed tests. Using parametric models allows the tests to be compared to traditional parametric tests. The models are trained on stated preference data from the SwissMetro dataset (Bierlaire *et al.,* 2001). The dataset is divided into a train and test dataset using a 70:30 split. The split is stratified by mode choice, and is grouped by individual, so that the test data is independent from train data. Each model is trained on the train set and then tested on the test set.

First the simple multinomial logit model from Bierlaire *et al.* (2001) is replicated on the train data (base model). The utility specification of this model is given in Table 1. New class indicators $\hat{y}_n$ are generated from the model's predicted probability distributions, for both the train and test data. As such, the true model P^* is known for the remaining models.

Table 1: Utility function of simple multinomial logit model

Variable		Alternative		
		SM	Car	SBB
ASC	Constant	SM	Car	-
TT	Travel time	B-Time	B-Time	B-Time
Cost	Travel cost	B-Cost	B-Cost	B-Cost
Freq	Frequency	B-Freq	-	B-Freq
GA	Annual season	B-GA	-	B-GA
Age	Age in classes	-	-	B-Age
Luggage	Pieces of luggage	-	B-Luggage	-
Seats	Airline seating	B-Seats	-	-

The same model specification is then used to fit a corresponding model on the simulated class indicators $\hat{y}$. Three further models are defined by removing an insignificant parameter, a significant parameter, and multiple significant parameters respectively. All models are tested using the single true-model test. Additionally, all models trained and tested on the simulated class indicators $\hat{y}_n$ are compared to each other using the KL-divergence test and the convex combination of classifiers test. A 5% confidence interval is used for all statistical tests.

In total 5 models are trained and tested:

1. **Base**: Base MNL using original class indicators y_n.
2. **True**: Base MNL using simulated class indicators $\hat{y}_n$.
3. **No-luggage**: Remove *B-Luggage* (insignificant parameter) only from utility specification, using simulated class indicators $\hat{y}_n$.
4. **No-age**: Remove *B-Age* (significant parameter) only from utility specification, using simulated class indicators $\hat{y}_n$.
5. **Time-cost**: Remove all parameters from utility specification, except *ASC*s, *B-Time* and *B-Cost*, using simulated class indicators $\hat{y}_n$.

5 Results

The parameter values for the MNL model estimated on the train set are given in Table 2. All parameters are significant except *B-Luggage*.

Table 2: MNL parameter estimates

Parameter	Estimate	z
ASC Car	1.382	8.599
ASC SM	1.548	10.14
B-Cost	-0.9082	-14.49
B-Time	-1.174	-17.38
B-Freq	-0.0045	-3.924
B-GA	1.242	5.448
B-Age	0.3168	7.675
B-Luggage	-0.1141	**-1.841**
B-Seats	-0.5283	-4.994
Final log-likelihood	-3712.07	
N	4734	

Table 3 shows the results for the single true-model test for each model, alongside the train and test log-likelihood. H_0 is held for both the base model and true model. H_0 is rejected for all other models.

Table 3: Train log-likelihood, test log-likelihood, and results for single true-model test for each model

Model	Base	True	No-luggage	No-age	Time-cost
Train LL (N=4374)	-3712.07	-3772.96	-3772.69	-3781.53	-3811.95
Test LL (N=2034)	-1545.31	-1552.75	-1558.13	-1567.46	-1576.81
μ_m	-1593.11	-1593.11	-1619.01	-1625.42	-1636.49
σ_m	26.39	26.39	26.20	26.16	26.23
z	**1.811**	**1.529**	2.324	2.216	2.275
p	**0.0701**	**0.1262**	0.0201	0.0267	0.0229

Table 4 shows the results for the Kullback-Leibler divergence test for each model. The true model and no-luggage model are shown not to be significantly different from each other. All other pairs of models are shown to be statistically different, including the true model and no-age model. This is consistent with the parameter value significances shown in Table 2.

Figure 5 and Table 5 show the results of the convex combination of classifiers test for each

Table 4: Results for Kullback-Leibler divergence test for each model. Each reference model defines a row (bold) and each candidate model defines a column.

μ				
	True	No-luggage	No-age	Time-cost
True	-	3.437	15.92	29.41
No-luggage	3.583	-	11.55	16.39
No-age	16.04	11.44	-	12.72
Time-cost	31.23	16.76	13.17	-
σ				
	True	No-luggage	No-age	Time-cost
True	-	2.568	5.620	7.431
No-luggage	2.734	-	4.829	5.657
No-age	5.681	4.760	-	4.960
Time-cost	8.135	5.852	5.226	-
p				
	True	No-luggage	No-age	Time-cost
True	-	**0.1809**	0.0046	0.0001
No-luggage	**0.1900**	-	0.0167	0.0038
No-age	0.0048	0.0162	-	0.0103
Time-cost	0.0001	0.0042	0.0117	-

model pair combination. In all cases, H_0 holds that λ does not differ significantly from one. This is consistent with the ranking of test log-likelihood scores shown in Table 3.

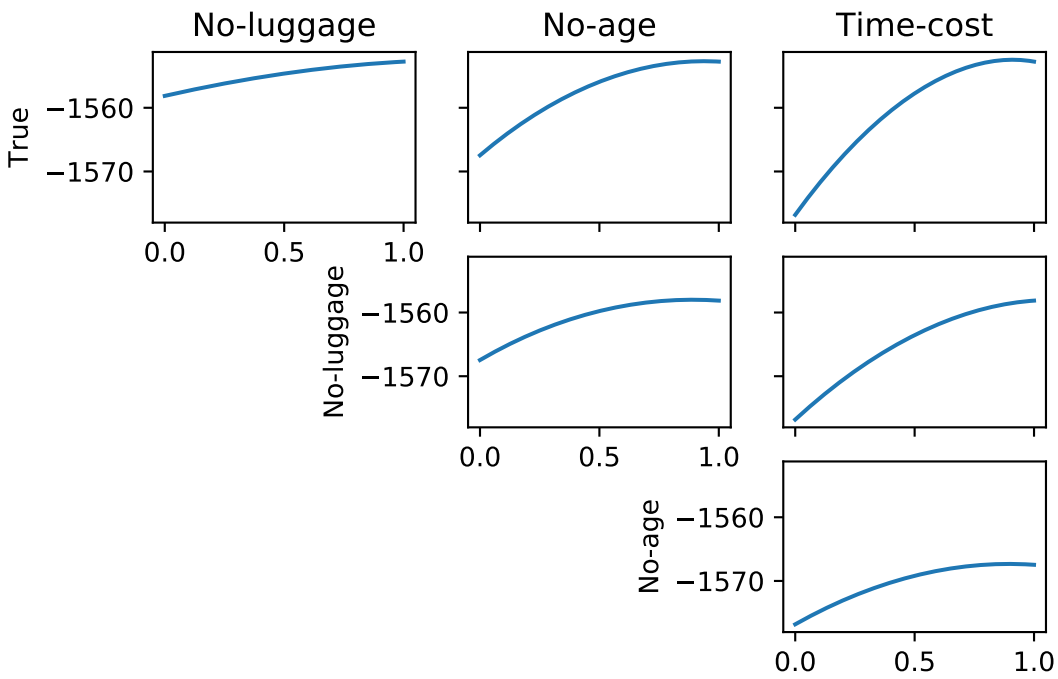

Figure 5: Log-likelihood plots of each convex combination of classifiers for λ

Table 5: Results for convex combination of classifiers test for each model pair combination. In each case $H_0 : \lambda = 1$.

μ			
	No-Luggage	No-age	Time-cost
True	1	0.937885	0.91034
No-luggage	-	0.890577	1
No-age	-	-	0.894997
σ			
	No-Luggage	No-age	Time-cost
True	0.367417	0.169064	0.118942
No-luggage	-	0.207129	0.167707
No-age	-	-	0.202137
p			
	No-Luggage	No-age	Time-cost
True	0.5	0.356658	0.225481
No-luggage	-	0.29865	0.5
No-age	-	-	0.301719

6 Conclusions

In this paper we present three statistical tests for model selection which are applicable to both parametric and non-parametric models. The *single true-model test* tests the analytical distribution of the expected likelihood of a model given the true model. The *Kullback-Leibler divergence test* uses similar anaylsis to determine the distribution of the Kullback-Leibler divergence between two models. Finally, the *convex combination of classifiers test* considers the log-likelihood of the convex combination of two classifiers under comparison. Through an applying the tests to parametric DCMs trained on the SwissMetro dataset we show the tests appear to be consistent with parametric statistical tests.

Planned further work includes further validation of the tests, including on applications of non-parametric models trained on larger datasets. Additionally, we plan to develop the methodology to allow for the analysis for cross-validation results.

7 References

Akaike, H. (1998) Information Theory and an Extension of the Maximum Likelihood Principle, in *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, 199–213, Springer, New York, NY, ISBN 978-1-4612-7248-9 978-1-4612-1694-0.

Ben-Akiva, M. and M. Bierlaire (2003) Discrete Choice Models with Applications to Departure Time and Route Choice, in *Handbook of Transportation Science*, International Series in Operations Research & Management Science, 7–37, Springer, Boston, MA, ISBN 978-1-4020-7246-8 978-0-306-48058-4.

Ben-Akiva, M. E., S. R. Lerman and S. R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, ISBN 978-0-262-02217-0.

Bierlaire, M., K. Axhausen and G. Abay (2001) The acceptance of modal innovation: The case of Swissmetro, paper presented at the *Swiss Transport Research Conference*.

Hastie, T., J. Friedman and R. Tibshirani (2008) *The Elements of Statistical Learning*, 2 edn., Springer Series in Statistics, Springer, ISBN 978-1-4899-0519-2 978-0-387-21606-5.

Kullback, S. and R. A. Leibler (1951) On Information and Sufficiency, *The Annals of Mathematical Statistics*, **22** (1) 79–86.