How many faces can be recognized? Performance extrapolation for multi-class classification

Charles Y. Zheng
Department of Statistics
Stanford University
Stanford, CA 94305
snarles@stanford.edu

Rakesh Achanta
Department of Statistics
Stanford University
Stanford, CA 94305
rakesha@stanford.edu

Yuval Benjamini
Department of Statistics
Hebrew University
Jerusalem, Israel
yuval.benjamini@mail.huji.ac.il

Abstract

The difficulty of multi-class classification generally increases with the number of classes. Using data from a subset of the classes, can we predict how well a classifier will scale with an increased number of classes? Under the assumption that the classes are sampled exchangeably, and under the assumption that the classifier is generative (e.g. QDA or Naive Bayes), we show that the expected accuracy when the classifier is trained on k classes is the $k-1$st moment of a *conditional accuracy distribution*, which can be estimated from data. We discuss estimation approaches based on pseudolikelihood, unbiased estimation, and high-dimensional asymptotics. These methods can be extended to a larger class of *asymptotically generative* classifiers, which include k-nearest neighbors, one-vs-one and one-vs-all classifiers. We compare these methods in simulations and real data.

1 Introduction

In multi-class classification, one observes pairs (z, y) where $y \in \mathcal{Y} \subset \mathbb{R}^p$ are feature vectors, and z are unknown labels, which lie in a countable label set $\mathcal{Z}$. The goal is to construct a classification rule for predicting the label of a new data point; generally, the classification rule $h : \mathcal{Y} \to \mathcal{Z}$ is learned from previously observed data points. In many applications of multi-class classification, such as face recognition or image recognition, the space of potential labels is practically infinite. However, one considers the classification problem on a finite subset of the labels $\mathcal{Z}_1 \subset \mathcal{Z}$: for instance, classifying the faces of 100 selected individuals from the population. At a later time, one might consider a larger (but still finite) classification problem on $\mathcal{Z}_2 \subset \mathcal{Z}$ with $\mathcal{Z}_2 \supset \mathcal{Z}_1$. In general, consider an infinite sequence of classification problems on subsets $\mathcal{Z}_1 \subset \cdots \subset \mathcal{Z}_t \subset \cdots$. Let S_i represent the training data available for the ith classification problem, and let $h_i : \mathcal{Y} \to \mathcal{Z}_i$ be the learned classification rule. Define the accuracy for the ith problem as

$$\text{acc}_i = \Pr[h_i(y) = z | z \in \mathcal{Z}_i].$$

Using data from only $\mathcal{Z}_k$, can one predict the accuracy achieved on the larger label set $\mathcal{Z}_K$, with $K > k$? This is the problem of *prediction extrapolation*.

A practical instance of prediction extrapolation occurs in neuroimaging studies, Kay et al. (2008) obtain fMRI brain scans which record how a single subject's visual cortex responds to natural images.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The label set $\mathcal{Z}$ corresponds to the space of all grayscale photographs of natural images, and the set $\mathcal{Z}_1$ is a subset of 1750 photographs used in the experiment. Kay et al. construct a classifier based on a combination of regularized multiple-response regression and Naive Bayes: they achieve over 0.75 accuracy on the subset of 1750 photographs, which by itself is already a convincing demonstration of the richness of the information contained in the fMRI scan. However, it would also be of interest to know what accuracy could be achieved on a larger set of photographs. Kay et al. calculated (based on exponential extrapolation) that it would take on the order of $10^{9.5}$ photographs before the accuracy of the model drops below 0.10! Directly validating this estimate would take immense resources, so it would be useful to develop the theory needed to understand how to compute such extrapolations in a principled way.

The most important assumption for our theory is that of *exchangeable sampling*. The labels in $\mathcal{Z}_i$ are assumed to be an exchangeable sample from $\mathcal{Z}$. The exchangeability further implies that the marginal distributions of $z \in \mathcal{Z}$ are equiprobable within every subset $\mathcal{Z}_i$. Without the condition of exchangeability (or a similar assumption), it is impossible on construct non-trivial bounds on the accuracy achieved on the new classes $\mathcal{Z}_K \setminus \mathcal{Z}_k$ based only on knowledge of $\mathcal{Z}_k$, since $\mathcal{Z}_k$ could consist entirely of well-separated classes while the new classes $\mathcal{Z}_K \setminus \mathcal{Z}_k$ consist entirely of highly inseparable classes, or vice-versa. The condition of exchangeability ensures that the separability of random subsets of $\mathcal{Z}$ can be inferred by looking at the empirical distributions in $\mathcal{Z}_k$, and therefore that some estimate of the achievable accuracy on $\mathcal{Z}_K$ can be obtained.

Unfortunately, the assumption of exchangeability is clearly violated in a majority of instances of multi-class classification. Many multi-class classification problems have a hierarchical structure, where the initial label set $\mathcal{Z}_1$ corresponds to a coarse-grained partition of the instances, and an expanded label set $\mathcal{Z}_2$ corresponds to a refinement of the partition induced by $\mathcal{Z}_1$: for instance, $\mathcal{Z}_1$ consists of the categories $\{\text{animal}, \text{vegetable}, \text{mineral}\}$, while $\mathcal{Z}_2$ consists of subcategories $\{\text{mammal}, \text{bird}, \text{insect}, \text{reptile}, \text{fungus}, \text{tree}, \text{flower}, \text{rock}, \text{metal}\}$. Not only is $\mathcal{Z}_2$ not a superset of $\mathcal{Z}_1$, but the marginal distributions within $\mathcal{Z}_2$ are necessarily more concentrated than the marginals of $\mathcal{Z}_1$. Many non-hierarchical classification problems are also excluded by the requirement of exchangeability. Consider the problem of annotating spoken words: the set $\mathcal{Z}_1$ might consist of data from the 100 most common words, while the set $\mathcal{Z}_2$ consists of data from the 1000 most common words. Exchangeability is violated because the words $z \in \mathcal{Z}$ are not equiprobable, but rather follow a long-tail law. It would be interesting to extend our theory to the hierarchical setting, or to handle non-hierarchical settings with non-uniform prior class probabilities, but we leave the subject for future work.

In addition to the assumption of exchangeability, we restrict the set of classifiers considered. We focus initially on *generative classifiers*, which are classifiers which work by training a model separately on each class. This convenient property allows us to characterize the accuracy of the classifier by selectively conditioning on one class at a time, which then reveals an equivalence between the expected accuracies of $\mathcal{Z}_k$ to moments of a common distribution. This moment equivalence result allows standard approaches in statistics, such as U-statistics and nonparametric pseudolikelilood, to be directly applied to the extrapolation problem. In non-generative classifiers, the classification rule has a joint dependence on the entire set of classes, and cannot be analyzed by conditioning on individual classes. However, in a particular limit (where the number of classes grows to infinity), we note that some non-generative multi-class classifiers can be *approximated* by a generative classifier, which therefore allows our framework to be extended to the class of *asymptotic generative* classifiers. In this paper, we show that particular variants of k-nearest neighbors, one-vs-one (OVO) and one-vs-all (OVA) classifiers are asymptotically generative. There are other classifiers, such as multinomial logistic regression, which we conjecture to be asymptotically generative, but at the same time we suspect that classifiers capable of representation learning, such as deep neural networks, are not asymptotically generative. Unlike generative classifiers, representation-learning classifiers can improve the model learned for a single class by using data from other classes. One can construct examples of representation-learning classifiers with counter-intuitive behaviors, such as a non-monotonic expected accuracy in the number of classes. Intuitively, we expect our theory to apply poorly to such representation-learning classifiers, which we confirm in data examples. We speculate that achieving prediction extrapolation for a representation-learning classifier would require a theory tailored to the dynamics of that particular classifier.

In section 2 we formalize the concepts of *classifier* and *prediction extrapolation*, and define generative and asymptotic generative classifiers. In section 3 we develop the theory of prediction extrapolation for generative classifiers, culminating in the definition of the conditional accuracy distribution and the equivalence between the moments of the conditional accuracy distribution and the expected accuracy. In section 4 we present three classes of methods for prediction extrapolation: moment-based, pseudolikelihood-based, and a method derived from the high-dimensional asymptotic classification theory of Anon (2016.) Section 5 presents simulated and real data examples, and section 6 concludes.

2 Prediction extrapolation

2.1 Multi-class classification

A natural approach for implementing a recognition system is through *supervised learning*: the data is input into a supervised learning algorithm in order to *learn* a classification rule $f(y)$. Examples of multi-class classifiers include k-nearest neighbors, multinomial logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), decision trees, and random forests, as well as the two 'divide and conquer' approaches, one-vs-one (OVO) and one-vs-all (OVA) (Friedman et al, 2008.) A unified definition of a classifier is as follows: a k-class classifier $\mathcal{C}$ is a function which takes $k+2$ arguments: the first k arguments are *probability distributions* $F_1, \dots, F_k$, the $(k+1)$st argument is a vector of prior probabilities $\vec{\pi}$, and the $(k+2)$nd argument is a *query* y [1]. The level sets $\{y : \mathcal{C}(F_1, \dots, F_k, y) = k\}$ are called *decision regions.* We say the classifier is *continuous* if and only if the *decision regions* of $\mathcal{C}$ are continuous in the first k arguments with respect to the topology of weak convergence[2]. Most of the commonly used classifiers satisfy this definition of continuity: an exception is k-nearest neighbors with fixed k, but then again, k-nearest neighbors with $\lim_{n\to\infty} k/n \in (0,1)$ is continuous. *Training a classifier* refers to defining a classification rule

$$f(y) = \mathcal{C}(\hat{F}_1, \dots, \hat{F}_k, \vec{\pi}, y)$$

where $\hat{F}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{y^{(i)},j}$ is the empirical distribution, and where $\vec{\pi}$ is commonly set to either be the uniform distribution on k elements, or the empirical proportions of the classes, but can be adjusted in order to favor certain classes over others.

Whenever new data is added, the classifier is retrained. Let $\hat{F}_{i,t}$ denote the empirical distribution of the measurements of species i at time t, and $\hat{F}_t$ denote the empirical distribution of all measurements at time t.

2.2 Exchangeability

A key assumption for our theory is that the marginal distributions of the classes $F_1, F_2, \dots$ are *exchangeable*: equivalently, that there exists a measure $\mathbb{F}$ on probability distributions in $\mathcal{Y}$, and $F_i \sim i.i.d.\mathbb{F}$ for $i = 1, \dots$. We call $\mathbb{F}$ a meta-distribution.

[[Discuss the signficance of this assumption. Without it, extrapolation is impossible.]]

2.3 Asymptotic Generative

Define a generative classifier $\mathcal{C}$ as satisfying

$$\mathcal{C}(F_1, \dots, F_k, \vec{\pi}, y) = \text{argmax}_i \mathcal{Q}(F_i, \vec{\pi}_i, y).$$

where $\mathcal{Q}$ is *scoring rule*: a real-valued function which takes three arguments: a distribution F on $\mathcal{Y}$, a prior probability π, and a query measurement $y \in \mathcal{Y}$. For notational convenience, we assume that ties occur with probability zero: that is, $\mathbb{F}$ and $\mathcal{Q}$ jointly satisfy the *tie-breaking* property: for any $\pi \in [0,1]$, letting $\hat{F}$ be an empirical distribution of finitely many points drawn from $F \sim \mathbb{F}$ respectively, $Y, Y' \overset{iid}{\sim} F$,

$$\Pr[\mathcal{Q}(\hat{F}, \pi, Y) = \mathcal{Q}(\hat{F}, \pi, Y')] = 0. \tag{1}$$

Quadratic discriminant analysis and Naive Bayes are two examples of generative classifiers. For QDA, the scoring rule is given by

$$\mathcal{Q}_{QDA}(F, \pi, y) = -(y - \mu(F))^T \Sigma(F)^{-1}(y - \mu(F)) - \log\det(\Sigma(F)) + 2\log\pi$$

[1] We treat randomized classifiers (such as random forests) as a probability distribution over deterministic classifiers.

[2] We say that a randomized classifier is continuous if and only if it is continuous with probability one.

where $\mu(F) = \int y dF(y)$ and $\Sigma(F) = \int (y - \mu(F))(y - \mu(F))^T dF(y)$. In Naive Bayes, the scoring rule is

$$\mathcal{Q}_{NB}(\hat{F}, \pi, y) = \log \pi + \sum_{i=1}^n \log \hat{f}_i(y_i)$$

where $\hat{f}_i$ is a density estimate for the ith component of F. Generative classifiers have the property that *information is not shared between classes*: one consequence is that such classifiers can be easily trained in parallel. More importantly, unsupervised data cannot be used to improve the classification rule. A larger class of classifiers shares this key property: these are *asymptotically generative* classifiers.

Definition 2.1. (Asymptotically generative) Suppose that for time stages $t = 1, 2, \ldots$, $k_t = O(t)$ and $n_{i,t} = O(t)$, with $F_1, F_2, \ldots$ drawn i.i.d. from $\mathbb{F}$, and $y^{(i),1}, \ldots$ drawn i.i.d. from F_i for each $i = 1, 2, \ldots$. A recognition system (characterized by mappings f_t) is considered *asymptotically generative* if and only there exists a scoring rule $\mathcal{Q}$ and probability vector $\tilde{\pi}$ such that defining

$$\tilde{f}_t(y) = \text{argmax}_{i=1}^{k_t} \mathcal{Q}(\hat{F}_{i,t}, \tilde{\pi}_i, y)$$

we have

$$\lim_{t \to \infty} \frac{1}{k_t} \sum_{i=1}^{k_t} \Pr[f_t(Y) = \tilde{f}_t(Y) | Y \sim p(y|x^{(i)})] \to 1.$$

We will show that recognition systems based on certain implementations of k-nearest neighbors, LDA, one-vs-one, or one-vs-all classifiers satisfy this definition of separability.

Definition 2.2.(i) Define a binary classifier $\mathcal{B}$ as a binary-valued mapping with four arguments: distributions F_0, F_1, prior probability π_0 and a query y. A one-vs-one (OVO) recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \sum_{j \neq i} I\left(\mathcal{B}(\hat{F}_{i,t}, \hat{F}_{j,t}, \frac{n_{i,t}}{n_{i,t} + n_{j,t}}, y) = 0\right),$$

resolving ties arbitrarily.

(ii) Define a binary scoring rule $\mathcal{D}$ as a real-valued mapping with four arguments: distributions F_0, F_1, prior probability π_0, and a query y. A one-vs-all (OVA) recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \mathcal{D}\left(\hat{F}_{i,t}, \sum_{j \neq i} \frac{n_{j,t}}{n_t - n_{i,t}} \hat{F}_{j,t}, \frac{n_{i,t}}{n_t}, y\right).$$

(iii) Let d be a distance metric on $\mathcal{Y}$. Let $D_t(y)$ denote the induced distribution of $d(Y, y)$ when $Y \sim \hat{F}_t$, and let $d_{\alpha,t}$ denote the α-quantile of $D_t(y)$. A kNN recognition system with neighborhood size α is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \Pr[d(y, Y) < d_{\alpha,t} | Y \sim \hat{F}_{i,t}].$$

(iv) Assume WLOG that $y_1 = 1$ for all $y \in \mathcal{Y}$, and let B^t be a $p \times k_t$ matrix which minimizes the log-likelihood

$$\sum_{j=1}^{k_t} n_{j,t} \mathbf{E}_{\hat{F}_j}\left[\langle Y, B_j^t \rangle - \log\left[\sum_{\ell=1}^{k_t} \exp[\langle Y, B_\ell^t \rangle]\right]\right].$$

A multinomial logistic regression recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \langle y, B_i^t \rangle.$$

Theorem 2.1.(more like conjecture; not actually proved yet!!) *(i) an OVO recognition system equipped with a continuous binary classifier is separable; (ii) an OVA recognition system equipped with a continuous binary scoring rule is separable; (iii) a kNN recognition system with fixed neighborhood size $\alpha \in (0, 1)$ is separable; (iv) a multinomial logistic regression recognition system is separable.*

3 Prediction Extrapolation

3.1 Problem formulation

Recall the notation from section 2.1. Assume that $F_1, F_2, \dots$ are sampled i.i.d. from the meta-distribution $\mathbb{F}$, and $y^{(i),1}, \dots, y^{(i),r}$ from F_i for $i = 1, 2, \dots$. Take $k_t = t$ and $n_{i,t} = r$.

Unlike in section 2.1., only the first $r_1 < r$ measurements in each species will be used to construct the classifier: redefine

$$\hat{F}_{i,t} = \frac{1}{r_1}\sum_{j=1}^{r_1} \delta_{y^{(i),j}}.$$

Since $\hat{F}_{i,t}$ no longer depends on t, we will write it as $\hat{F}_i$. The remaining $r_2 = r - r_1$ measurements of each species constitute the *test set*, used to evaluate the performance of the classifier.

The generalization accuracy at time t is defined

$$\text{acc}^{(t)} = \frac{1}{k}\sum_{i=1}^{k} \Pr[f_t(y) = i | y \sim p(y|x^{(i)})].$$

The extrapolation problem is the problem of predicting $\text{acc}^{(K)}$ using only information known at time $k < K$, namely, $\{y^{(i),j}\}_{i=1,j=1}^{k,r}$.

3.2 Conditional accuracy

Consider estimating the expected accuracy at time t,

$$p_t \overset{def}{=} \mathbf{E}[\text{acc}^{(t)}].$$

Assume that the classifier is based on a scoring rule $\mathcal{Q}$. Further assume that $\mathcal{Q}$ has a trivial dependence on the prior probability parameter: $\mathcal{Q}(F, a, y) = \mathcal{Q}(F, b, y)$ for all F, y, and $a, b \in [0, 1]$. This assumption is more mild than it appears, since most classifiers indeed have a trivial dependence on $\vec{\pi}$ in the case when $\vec{\pi}$ is set to the uniform distribution.

Define the *conditional accuracy* function $u(F, y)$ which maps a distribution F on $\mathcal{Y}$ and a *test* observation y to a real number in $[0, 1]$. The conditional accuracy gives the probability that for independently drawn F and F' from $\mathbb{F}$, letting $\hat{F}'$ be the empirical distribution of r_1 measurements drawn from F', that the scoring function $\mathcal{Q}(F, 0, y)$ will give a higher score to y than the scoring function $\mathcal{Q}(\hat{F}', 0, y)$:

$$u(F, y) = \Pr[\mathcal{Q}(F, 0, y) > \mathcal{Q}(\hat{F}', 0, y)].$$

Define the *conditional accuracy* distribution μ as the law of $u(\hat{F}, Y)$ when $F \sim \mathbb{F}$, and $\hat{F}, Y$ are both obtained from F. The significance of the conditional accuracy distribution is that the expected generalization error p_t can be written in terms of its moments.

Theorem 3.1. *Let U be defined as the random variable*

$$U = u(F, Y)$$

for X, Y drawn from $p(x, y) = p(x)p(y|x)$, and $\hat{F}(X) = \frac{1}{r_1}\sum_{j=1}^{r_1} \delta Y^j$ with $Y^i \overset{iid}{\sim} p(y|X)$ Then $p_k = \boldsymbol{E}[U^{k-1}]$.

Proof. Write $q^{(i)}(y) = \mathcal{Q}(\hat{F}_i, 0, y)$, and let $Y^{(i),*} \sim p(y|X^{(i)}$ for $i = 1, \ldots, k$. Note that by using conditioning and conditional independence, p_k can be written

$$
\begin{aligned}
p_k &= \mathbf{E}\left[\frac{1}{k}\sum_{i=1}^{k} \Pr[q^{(i)}(Y^{(i),*}) > \max_{j \neq i} q^{(j)}(Y^{(i),*})]\right] \\
&= \mathbf{E}\left[\Pr[q^{(1)}(Y^{(1),*}) > \max_{j \neq 1} q^{(j)}(Y^{(1),*})]\right] \\
&= \mathbf{E}[\Pr[q^{(1)}(Y^{(1),*}) > \max_{j \neq 1} q^{(j)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]] \\
&= \mathbf{E}[\Pr[\cap_{j>1} q^{(1)}(Y^{(1),*}) > q^{(j)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]] \\
&= \mathbf{E}[\prod_{j>1} \Pr[q^{(1)}(Y^{(1),*}) > q^{(j)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]] \\
&= \mathbf{E}[\Pr[q^{(1)}(Y^{(1),*}) > q^{(2)}(Y^{(1),*})|Y^{(1),*}, \hat{F}_1]^{k-1}] \\
&= \mathbf{E}[u(\hat{F}_1, Y^{(1),*})^{k-1}] = \mathbf{E}[U^{k-1}].
\end{aligned}
$$

□

Theorem 3.1 tells us that the problem of extrapolation can be approached by attempting to estimate the conditional accuracy distribution. The $(t-1)$th moment of U gives us p_t, which will in turn be a good estimate of $\text{acc}^{(t)}$.

3.3 Properties of the conditional accuracy distribution

The conditional error distribution ν is determined by $\mathbb{F}$ and $\mathcal{Q}$. What can we say about the the conditional accuracy distribution without making any assumptions on either $\mathbb{F}$ or $\mathcal{Q}$? The answer is: not much–for an arbitrary probability measure ν' on $[0, 1]$, one can construct $\mathbb{F}$ and $\mathcal{Q}$ such that $\nu = \nu'$.

Theorem 3.2. *Let U be defined as in Theorem 2.1, and let ν denote the law of U. Then, for any probability distribution ν' on $[0, 1]$, one can construct a meta-distribution $\mathbb{F}$ and a scoring rule $\mathcal{Q}$ such that $\nu = \nu'$.*

In practice, however, the scoring rule $\mathcal{Q}$ must approximate a monotonic function of the conditional density $f = \frac{dF}{dy}$ in order to yield an effective classifier.

It is therefore notable that in the case that F has a density with respect to Lesbegue measure, and where $\mathbb{F}$ has no atoms, taking an *optimal* scoring rule, with the property that $\mathcal{Q}(\hat{F}, y) = g(f(y))$ for monotonic g, the distribution of U has a monotonically increasing density.

Theorem 3.3. *Let U be defined as in Theorem 3.1, and let ν denote the law of U. Suppose F has a density $f(y)$ with respect to Lebesgue measure on $\mathcal{Y}$ with probability one, $\mathbb{F}$ has no atoms, and ($\mathbb{F}$, $\mathcal{Q}$) jointly satisfy the property of monotonicity*

$$f(y) > f(y') \textit{ implies } \mathcal{Q}(\hat{F}, 0, y) > \mathcal{Q}(\hat{F}, 0, y')$$

and the property of tie-breaking (1) *with probability one. Then μ has a density $\eta(u)$ on* $[0, 1]$ *which is monotonic in u.*

4 Nonparametric Estimation

Let us assume that U has a density $\eta(u)$. While $U = u(\hat{F}, 0, Y)$ cannot be directly observed, we can estimate $u(\hat{F}_i, 0, y^{(i),r_1+j})$ for any $1 \leq i \leq k$, $1 \leq j \leq r_2$ from the data.

Theorem 4.1. *For given $p(x, y)$ and scoring rule $\mathcal{Q}$, assume that U as defined in Theorem 3.1 has a density $\eta(u)$ and that $\mathcal{Q}$ satisfies the tie-breaking property* (1)*. Define*

$$V_{i,j} = \sum_{i=1}^{k} I(q^{(i)}(y^{(i),j}) > q^{(j)}(y^{(i),j})).$$

Then

$$V_{i,j} \sim \textit{Binomial}(k, u(\hat{F}_i, y^{(i),j})).$$

At a high level, we have a hierarchical model where U is drawn from a density $\eta(u)$ on $[0, 1]$ and then $V_{i,j} \sim \text{Binomial}(k, U)$; therefore the marginal distribution of $V_{i,j}$ can be written

$$\Pr[V_{i,j} = \ell] = \binom{k}{\ell} \int_0^1 u^\ell (1-u)^{k-\ell} \eta(u) du.$$

However, the observed $\{V_{i,j}\}$ do *not* comprise an i.i.d. sample.

We discuss the following three approaches for estimating $p_t = \mathbf{E}[U^{t-1}]$ based on $V_{i,j}$. The first is *unbiased estimation* based on binomial U-statistics, which is discussed in Section 4.1. The second is the *psuedolikelihood* approach. In problems where the marginal distributions are known, but the dependence structure between variables is unknown, the *psuedolikelihood* is defined as the product of the marginal distributions. For certain problems in time series analysis and spatial statistics, the maximum psuedolikelihood estimator (MPLE) is proved to be consistent (CITE). We discuss psuedolikelihood-based approaches in Sections 4.2 and 4.3.

4.1 Unbiased estimation

If $V \sim \text{Binomial}(k, \eta)$, then an unbiased estimator $f_t(V)$ of $\eta^(t-1)$ exists if and only if $0 \leq t \leq k$.

The theory of U-statistics provides the minimal variance unbiased estimator for $\eta^(t-1)$:

$$\eta^t = \mathbf{E}\left[\frac{\binom{V}{t}}{\binom{k}{t}}\right].$$

This result can be immediately applied to yield an unbiased estimator of p_t, when $t \leq k$:

$$\hat{p}_t^{UN} = \mathbf{E}\left[\frac{1}{kr_2} \sum_{i=1}^{k} \sum_{j=1}^{r_2} \frac{\binom{V_{i,j}}{t}}{\binom{k}{t}}\right]. \tag{2}$$

The problem of *extrapolation* concerns the case $t > k$, in which the expression (2) is undefined. Still, the estimator (2) is worthy of study, since it has close to optimal performance for the case $t \leq k$.

4.2 Maximum pseudo-likelihood

The psuedolikelihood is defined as

$$\ell_t(\eta) = \sum_{i=1}^{k} \sum_{j=1}^{r_1} \log\left(\int u^{V_{i,j}} (1-u)^{k-V_{i,j}} \eta(u) du\right), \tag{3}$$

and a maximum psuedolikelihood estimator (MPLE) is defined as any density $\hat{\eta}$ such that

$$\ell(\hat{\eta}_{MPLE}) = \sup_\eta \ell_t(\eta).$$

The motivation for $\hat{\eta}_{MPLE}$ is that it consistently estimates η in the limit where $k \to \infty$.

Theorem 4.2. *For given $\mathbb{F}$ and scoring rule $\mathcal{Q}$, assume that U as defined in Theorem 3.1 has a density $\eta(u)$ and that $\mathcal{Q}$ satisfies the tie-breaking property* (1)*, and also that $r_2 \geq 1$. For $t = 1, 2, \ldots,$, let $\hat{\eta}_t$ be any MPLE for ℓ_t. As $k_t \to \infty$, $\hat{\eta}_t$ weakly converges to η.*

However, in finite samples, $\hat{\eta}_{MPLE}$ is not uniquely defined, and if we define the plug-in estimator

$$\hat{p}_t^{MPLE} = \int u^{t-1} \hat{\eta}_{MPLE}(u) du,$$

$\hat{p}_t^{MPLE}$ can vary over a large range, depending on which $\hat{\eta} \in \text{argmax}_\eta \ell_t(\eta)$ is selected. These shortcomings motivate the adoption of additional constraints on the estimator $\hat{\eta}$.

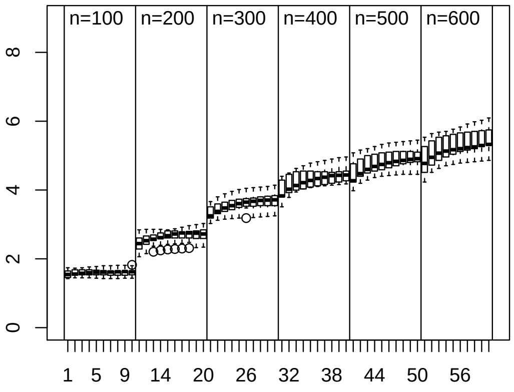

Figure 1: Extrapolation classification performance for CIFAR data. (This simulation needs to be fixed later.) PMLE: maximum psuedolikelihood. MCPMLE: Moment-constrained max psuedolikelihood. Info: Zheng and Benjamini's info-theoretic method. Unbiased: U-statistic (cannot be used to extrapolate.)

4.3 Constrained pseudo-likelihood

Theorem 3.2. motivates the *monotonicity constraint* that $\frac{d\hat{\eta}}{du} > 0$, hence we define $\hat{\eta}_{INC}$ as a solution to

$$\text{maximize } \ell_t(\eta) \text{ subject to } \frac{d\hat{\eta}}{du} > 0.$$

An alternative strategy is to directly attack the variability is $\hat{p}_t$ due to non-uniqueness of $\hat{\eta}$. Therefore, we define $\hat{\eta}_{MC}$ (where MC stands for moment-constrained) as

$$\text{maximize } \ell_t(\eta) \text{ subject to } \int u^{k-1}\eta(u)du = \hat{p}_k^{UN}.$$

Thirdly, we can combine both the moment constraint and the monotonicity constraint, yielding $\hat{\eta}_{COM}$, which is obtained by solving

$$\text{maximize } \ell_t(\eta) \text{ subject to } \int u^{k-1}\eta(u)du = \hat{p}_k^{UN} \text{ and } \frac{d\hat{\eta}}{du} > 0.$$

Unfortunately, none of the three density estimators are uniquely defined. An easy way to see this is to transform the parameterization of $\eta(u)$, defining

$$\eta(u) = \int_0^u \xi(u)du;$$

the monotonicity constraint is equivalent to the condition that $\xi > 0$, and the moment condition translates into a linear equality constraint on ξ.

5 Results

6 Discussion

Acknowledgments

CZ is supported by an NSF graduate research fellowship.

References

[X] Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, 56(2), 400-410.

[X] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2008.