A Conditional Gradient Approach for Nonparametric Estimation of Mixing Distributions

Mixture models have been studied extensively in numerous fields and estimating the mixing distribution is a fundamental problem of interest. A common application of mixtures in operations is for the problem of accurately predicting the customer demand for different product offerings, which can help firms make key decisions such as the optimal assortment to offer to customers, determine which of the existing products to discontinue from the market because of low demand/popularity or enforce price changes to attract more customers to different products. Given historical sales transactions and inventory data—which provides product availability information—the goal is to fit a mixture of choice models to the data, where the choice model specifies how customers make choices from a set of offered products. The fitted mixture model is then used to make predictions and inform decisions. In particular, mixtures of logit models have found great practical success and have been shown to approximate a wide class of mixtures.

Given its significance, this paper considers the problem of estimating the mixing distribution in mixtures of logit models, given sales transactions and product availability data. Our goal is to propose a general-purpose technique that is applicable to a broad set of contexts (with little to no customization), is numerically stable, and can scale to large datasets.

Most existing techniques for estimating the mixture of logit model broadly fall into two categories. The first assumes a pre-specified parametric form (such as the normal or log-normal) for the mixing distribution and then estimates the parameters (typically) via maximum likelihood estimation [3]. However, these techniques suffer from potential misspecification issues because of imposing a priori parametric forms for the mixing distribution. In order to overcome this issue, nonparametric estimation techniques have been proposed which search for the best fitting mixture distribution from a large class of mixing distributions. For instance, [2] considers the class of discrete mixture of normal distributions as well as the class of discrete distributions with a “large” support size, both of which can be shown to approximate any general class of continuous distributions. The best fitting distribution is determined via maximum likelihood estimation. However, the estimation problem becomes difficult to solve and encounters numerical issues as the number of parameters increases.

To address the above limitations, we propose a novel nonparametric method to estimate the mixture of logit models. Our approach finds the best fitting mixing distribution to the data, where the fit is measured through a convex loss function, amongst the class of all possible mixing distributions. Our approach broadly consists of the following two steps: (1) we formulate the estimation problem as a constrained convex program and (2) we apply the conditional gradient (aka Frank-Wolfe) algorithm to solve this convex program, showing that it iteratively generates the support of the mixing distribution as well as the mixing weights.
In particular, we use the insight that instead of optimizing over the mixing distribution, the estimation problem can be solved by directly optimizing over the predicted choice probabilities under the mixture—subject to the constraint that they are consistent with some underlying mixing distribution. This results in a constrained convex program with the variables as the predicted choice probabilities. However, solving this convex program is not sufficient since our original goal was to recover the mixing distribution.

Our key contribution is to apply the conditional gradient algorithm to solve the above convex program which simultaneously performs both tasks of optimizing over the predicted choice probabilities and recovering the underlying mixing distribution consistent with those probabilities. When applied to our setting, the subproblem obtained in each iteration of the conditional gradient algorithm is non-convex and can be hard to solve in general. However, it is sufficient to determine an improving solution in each iteration to ensure convergence (since we are solving a convex program), and we show that standard off-the-shelf solvers are able to obtain improving solutions in our numerical studies. In addition, there has been extensive work in the machine learning community on designing and improving conditional gradient algorithms which offers two key advantages to our estimation technique: (a) precise convergence guarantees and (b) scalability to large-scale and high dimensional settings. Lastly, another nice feature of our methodology is the plug-and-play nature of the estimation approach—the algorithm can be applied with little to no customization for a broad class of loss functions including the popular log-likelihood and squared losses.

**Summary of key results.** Our work makes the following contributions:

1. **Novel mixture estimation methodology.** Our estimator is (a) general-purpose: can be applied with little to no customization for a broad class of loss functions; (b) fast: order of magnitude faster than the benchmark expectation-maximization (EM) algorithm; and (c) nonparametric: makes no assumption on the mixing distribution and estimates customer types in the population in a data-driven fashion.

2. **Analytical results.** Our theoretical analysis is two-fold:

   (a) We provide convergence guarantees for our conditional gradient based estimator, for both the log-likelihood and squared loss functions. Specifically, we establish sub-linear convergence—$O(1/k)$ where $k$ is number of iterations—towards the optimal loss objective.

   (b) We analyze the structure of the estimated mixing distribution and show that our method recovers customer types that assign zero probability to a subset of the products, bearing close resemblance to the concept of consideration sets in existing literature where customers are assumed to first consider a subset of products and then choose from this subset. We refer to such types as boundary types.
provide a characterization that reveals insights into their choice behavior as well as allowing for consistent out-of-sample predictions.

3. Empirical results. We conducted three numerical studies to validate our methodology:

(a) On the SUSHI Preference Dataset, consisting of preference orderings given by respondents for different sushi varieties, we show that our method achieves superior in-sample fit compared to fitting a latent class MNL (LC-MNL) model [1] using the EM algorithm, for both the log-likelihood (24% better) and squared loss (58% better), with $16\times$ speedup in the estimation time. Our approach also achieves better predictive accuracy than EM, with an average 27% and 16% reduction in the RMSE (root mean square error) and MAPE (mean absolute percentage error) metrics for predicting market shares on new assortments.

We then use the estimated mixture model to solve the assortment optimization decision, i.e. determining the subset of products to offer to the population that maximizes revenue, and show that our method can extract $\sim 23\%$ more revenue from the population. This is significant because we are fitting the exact same model (mixture of logit) and only changing the estimation technique, and shows the amount of money that a sub-optimal method can leave on the table.

(b) On real-world sales transaction data from the IRI Academic Dataset, we show that our approach achieves an average of 3.9% and 2.9% improvement respectively in the in-sample and out-of-sample log-likelihood loss compared to the EM benchmark. Similarly, we achieve an average 5.7% and 4.2% reduction in the in-sample and out-of-sample squared loss respectively. In particular, our method outperformed EM-based estimation in all 5 product categories that we considered.

(c) Using synthetic data, we test the robustness of our estimator to different ground-truth mixing distributions and show that it consistently recovers a good approximation to the underlying distribution, which is significantly better than a benchmark method imposing a parametric assumption on the mixing distribution.

References

