A Bayesian Approach to Predicting Occupational Transitions

Lilia Chang[1⋆], Lisa Simon[2], Karthik Rajkumar[3], and Susan Athey[2]

[1] Institute of Computational and Mathematical Engineering, Stanford University, lilia.chang@stanford.edu
[2] Stanford Graduate School of Business, lksimon@stanford.edu, athey@stanford.edu
[3] Economics Department, Stanford University, krajkumar@stanford.edu

Abstract. There is large uncertainty regarding the future of the labor market when considering the development of technologies that may displace low-wage workers. Towards our goal of being able to make counterfactual recommendations to people so they may maximize their expected, long-term financial returns, we must first understand the transition probabilities of persons between occupations. An ideal model should incorporate the observed attributes of the start and end-states of a person. But it is also reasonable to assume that there are many unobserved factors that may determine a person's transition probabilities. For this end we adapt a Bayesian factorization model presented by Athey, et al. (2018) that captures the latent factors of observed "start" and "end" states for a person. The Travel-Time Factorization Model (TTFM) uses variational inference to estimate the large posterior and stochastic gradient descent. It is able to achieve accuracy measures that improve upon those of competing models due to the personalization of the estimated covariates (F1 score of 0.241 compared to 0.05 and 0.118 of simpler models).

1 Introduction

The future of the labor market is uncertain for many, particularly for low-wage workers in sectors vulnerable to disruptive technologies. Workers displaced by AI technologies, for example, will be forced to transition to new jobs. For this reason we are interested in understanding what factors make job transitions more likely. Doing so will allow us to map out the possible career paths for a person with certain characteristics and predict how choices such as enrolling in a career development program, for example, might affect their path and future outcomes.

Calculating the transition probabilities between states (defined by an occupation and education level pair) is itself a non-trivial task: one ought to take into account the various factors that may determine whether a person transitions between jobs, and such factors are often unknown or difficult to define. As a first step towards the analysis we hope to do, we take a Bayesian approach by adapting the techniques used in Athey, et. al.'s paper on estimating consumer preferences over restaurants [2]. We compare the performance of the TTFM with traditional multinomial logit models and a conditional logit model. Our results show that personalization of covariates with such a heterogeneous input and output space is necessary and achievable for this problem setting.

We first discuss the related work in the domain of labor market analysis and how the broader literature on discrete-choice behavior models may be applied to improve on this domain-specific problem. We then introduce our dataset and the necessary transformations we took. We conclude by presenting our results and discussing what future work still needs to be done.

2 Related work

In the broader discrete choice literature, it is common to include latent variables on individual persons' preferences or on person-item mixed effects. The prior techniques used in the estimation of occupational transitions have mainly used probabilities estimated by multinomial logit models in a Markov chain that assume time-homogeneous transitions, or have dealt with the large state-space by looking simply at the statuses of "unemployment" and "employment"[4] [6]. These models rely on estimating covariates of observed attributes and do not include latent factors, nor do they have a notion of utility for persons. Including latent factors on start-states, end-states, and on time is then a large improvement in the domain of analyzing occupational transitions.

The TTFM model proposed by Athey, et al. serves as our main reference for this paper [2]. The TTFM uses a discrete choice framework to model each person's choice for an item (in particular, a restaurant) to infer the covariates for the person's utility functions from their choices. The TTFM allows for efficient estimation of many latent variables, allowing for "personalization" of the model for persons/restaurants. The results of the 2018 paper show how robust the TTFM is compared to standard logit models in prediction. For other works related to estimating many latent variables for in discrete choice, see also Donnelly et al. (2019) [5] and Ruiz et al. (2017) [7].

⋆ The only student in CS 229. All other authors are affiliated faculty and postdoc at the GSB and PhD student at the economics department.

3 Data

We are using the Current Population Survey (CPS) dataset [1], a national U.S. labor force survey, supplemented by Autor and Dorn's dataset of U.S. occupation indices [3]. The CPS has a panel of 54k respondents per year and is representative of the nation's population. We have included data from 1991-2018, where each example consists of a person and their characteristics (see §3.1), and their occupation in one year and then the next. The data unfortunately does not keep persons in the dataset after observing them for two years, and so we only observe a given person's occupation for two consecutive years. We use Autor and Dorn's indices as attributes of the occupations, also listed in §3.1.

3.1 Transformations

An ideal model will capture the heterogeneity of the participants by estimating a unique set of covariates for each person. Because our dataset only consists of one observed transition per survey participant we do not have enough observations per person to be able to generalize what their unique characteristics are. Instead we pool similar survey participants together and take the means over their attributes to form "personas." We define a persona by occupation and number of years of education. So if person A and person B both share the same occupation and number of years of education, their individual attributes both contribute to the those of the appropriate persona.

Thus the attributes of a persona, defined by the tuple (`occupation, # years of education`) are the following:

- `age`: mean age of persons in persona group,
- `sex`: mean sex of persons in group (0 for `male`, 1 for `female`),
- `experience`: mean number of years of being in the workforce,
- `race_x`: percent of persons of race `x` in persona group, with `x` ∈ `white, black, asian, hispanic, native american, mixed, other`.
- `industry_x`: percent of persons in industry `x` in persona group. There are 10 industry categories.
- `task_abstract`: index of level of "abstract" work required in occupation according to Autor and Dorn.
- `task_manual`: index of level of "manual" work required in occupation according to Autor and Dorn.
- `task_routine`: index of level of "routine" work required in occupation according to Autor and Dorn.

4 Models

We model a persona's choice for persona in the next year, with the observed transitions taken from the transitions of individuals in the dataset. We assume that a persona selects the persona that maximizes their utility, where the utility of persona i choosing persona j at time t is U_{ijt}. This utility is defined differently for each of the following models.

For all the models, the transition probability from i to j is given by

$$P(y_{it} = j) = \frac{\exp\{U_{ijt}\}}{\sum_k \exp\{U_{ikt}\}}. \tag{1}$$

We operate in a sparse environment where there are many more potential choices an individual could make regarding their destination states than they actually do in the data. To improve convergence and accuracy in the model, we limit the number of choices each individual can actually make to only those destination personas that we observe actual transitions to in the data. This also makes semantically infeasible transitions, say going from a doctor to a lawyer in one year, irrelevant. For each of the 4,000 states, there are on average 236 states they may transition to in the next time period.

4.1 Bayesian factorization (TTFM)

The model for utility that TTFM assumes is the following:

$$U_{ijt} = \underbrace{\theta_i^\top \beta_j}_{\text{Latent-Latent intercept}} + \underbrace{W_i^\top \rho_j}_{\text{start-state observables}} + \underbrace{\sigma_i^\top X_j}_{\text{end-state observables}} + \underbrace{\mu_j^\top \delta_t}_{\text{time-varying effect}} + \underbrace{\epsilon_{ijt}}_{\text{noise}} \tag{2}$$

where W_i are the observed traits of persona i and X_j the observed traits of persona j. We describe the estimated variables in detail below.

- *Start-state covariates.* Each end-state persona j has covariates on start-state attributes that are inferred from the data and represented as ρ_j. Then the product $W_i^\top \rho_j$ describes how "aligned" the traits of the start-state persona i are with the attributes of persona j. The dimension of ρ_j is determined by the number of the start-state's attributes.
- *End-state covariates.* Similar to end-state variables but for the persona as a starting state. Note that a persona i will have ρ_j not equal to σ_i.
- *Latent intercept.* Each start-state persona i has an intercept-vector with dimension k (hyperparameter), as does each end-state persona j. The inner-product $\theta_i^\top \beta_j$ then captures the "agreeability" of the latent variables of start-state i and end-state j.
- *Time-varying effects.* Taking into account time-varying effects allows us to model how utilities of transitions vary with time. $\mu_j^\top \delta_t$ captures the variation of the utility for end-persona j in year t.
- *Noise terms.* We place a Gumbel prior over the error ϵ_{ijt} which leads to a softmax model.

4.2 Conditional Logit

The conditional logit model (CLM) is a classic model of choice behavior in econometrics. The setting is one where individuals make a discrete choice, i.e. they choose between a discrete number (larger than 2) of distinct choices. In this model, covariates are allowed to be individual-specific, choice-specific or even an interaction of user and choice.

The utility for our setting for the conditional logit model is then,

$$U_{ijt} = W_i^\top \rho_j + \sigma_i^\top X_j + \epsilon_{ijt}. \tag{3}$$

Compared to the utility used in TTFM, the CLM utility lacks a latent-latent intercept and time-varying effect. Then the chief difference between TTFM and the conditional logit is the level of personalization: TTFM is able to personalize the model to a much greater extent because it includes coefficients that are specific to the starting state and the ending state, as well as extract, latent, unobserved factors for i and j. It does this while remaining computationally feasible because of the latent variable structure that captures the most essential aspects of this personalization in a reduced-dimension space. A benefit of this is that the model is able to fit flexible curves to the data without the need to explicitly specify it. The conditional logit, on the other hand, needs careful thought into the exact functional form because of the limits on how many features one can include in it. This lack of flexibility can potentially be beneficial, however, as the conditional logit is less likely to overfit to the data and it also well grounded in discrete choice theory. It is therefore a good benchmark to have.

4.3 Multinomial Logit

For comparison we consider a classical multinomial logit model (MNL) that is even more restrictive than the CLM or TTFM: there is no incorporation of end-state attributes X_j. Simply we have,

$$U_{ijt} = W_i^\top \rho_j + \epsilon_{ijt}. \tag{4}$$

It is expected that with no incorporation of end-state attributes that the MNL will perform even more poorly than the CLM.

4.4 Estimation

For all models, we may set a prior over the learned variables $(\theta_i, \beta_j, \rho_j, \sigma_i, \mu_j, \delta_t)$ where they exist. It is expected that this should make any of the models more generalizable and thus more robust. There are a number of ways we can choose the prior: we may set them to be standard Gaussians, or we could let the observed attributes affect the mean and variance of the respective latent variables (relevant to just ρ_j, σ_i). This should allow further "personalization" of the model to the different personas, although it may introduce over-fitting.

Further, we estimate the posterior over the latent terms using variational inference as Athey, et. al. do. Variational inference approximates the posterior with a more simple distribution, choosing the one that is "closest" according to Kullback-Leibler (KL) divergence. It may be shown that minimizing KL divergence is equivalent to maximizing the evidence lower bound (ELBO),

$$\mathcal{L} = \mathbb{E}_{q(\mathcal{H})}\big[\log p(y, \mathcal{H}) - \log q(\mathcal{H})\big] \tag{5}$$

with y being the observed data and $\mathcal{L} \leq \log p(y)$. That is, intead of finding the exact posterior over the latent variables we use the "close-enough" distribution $q(\mathcal{H})$.

Following Athey, et. al. we use Gaussian variational factors for all latent variables and so we maximize the ELBO $\mathcal{L}$ with respect to the mean and variance parameters of these Gaussians. To overcome the difficulty of finding the expectations see that $\nabla\mathcal{L}$ can be expressed as an expectation. TTFM then uses Monte Carlo estimators of the gradient, specfically the reparametrization gradient, at each optimization step. To consider the large size of our dataset, we use batched stochastic gradient descent.

5 Experiments and Results

The dataset has been split at random into 70% train, 10% validation, and 20% test. We try, with various hyperparameters, the following experiments for TTFM and CLM: setting no priors on the latent variables (equivalently, high variance priors) and setting standard Gaussian priors on the latent variables. The results below for TTFM are with the best-found hyperparameters for the dimension of θ_i, β_j being 30 and the dimension of δ_t being 10. Thus the TTFM has more than twice the number of latent variables as the CLM. Considering our total dataset size, we use batch-size of 5000 and train for 3000 iterations with time-decaying learning rate, initialized at 0.005.

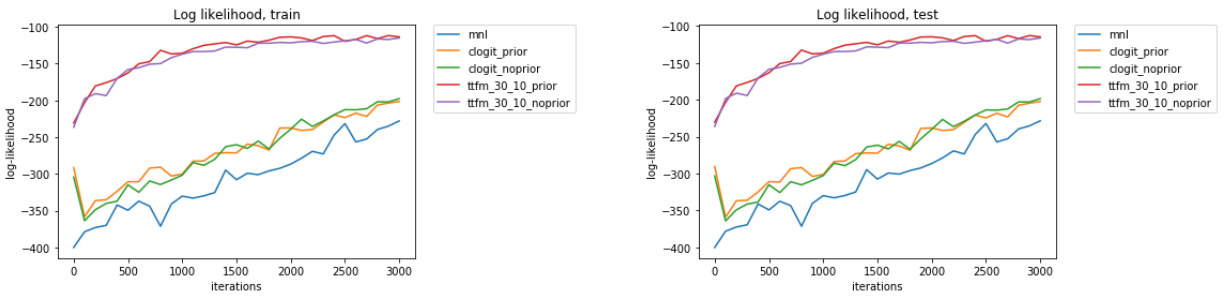
Table 1: Results of different model performances

Model	Prior Dist.		Log-likelihood	Accuracy	Precision	Recall	F1
TTFM	none	Train	−115.1	0.318	0.384	0.163	0.229
		Test	−116.3	0.316	0.357	0.182	0.241
	N(0,1)	Train	−113.6	0.313	0.376	0.156	0.220
		Test	−114.8	0.312	0.356	0.175	0.235
CLM	none	Train	−197.6	0.168	0.213	0.080	0.117
		Test	−198.5	0.168	0.184	0.084	0.115
	N(0,1)	Train	−201.5	0.141	0.212	0.079	0.115
		Test	−202.3	0.140	0.191	0.086	0.118
MNL	–	Train	−227.9	0.073	0.138	0.028	0.046
		Test	−228.4	0.073	0.114	0.032	0.050

The log-likelihood column is the mean log-likelihood over all the samples. Accuracy is defined as the fraction of correctly classified instances. Precision is the fraction of instances where we correctly classify the end-state as persona j where the algorithm declares j, and then the mean over all all personas. Recall is the fraction of instances where we correctly classify persona j out of all cases where the true choice was j (and then the mean over all personas, again).

In contrast to our expectations, having no prior over the latent variables slightly improves the model as shown in the above result, more obviously for CLM than for TTFM. This may indicate that we should not assume that the covariates across different personas be from the same distribution at all. Assuming no prior then allows the model to better "personalize" the covariates.

The plots in Figure 1 show that TTFM more quickly converges than the other models. Additionally, its F1 score is vastly improved as shown in Figure 2.

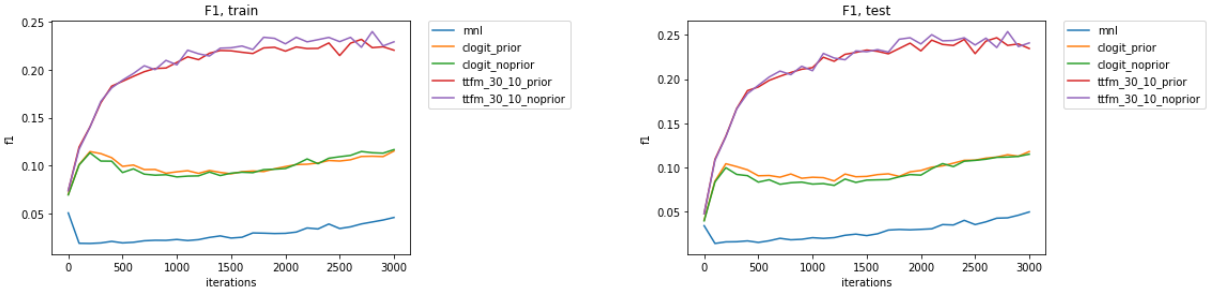


(a) (b)

Fig. 1: Log-likelihoods of various models for train and test.

We list the observed attributes that had greatest mean-valued covariates below for the TTFM model with no priors in Table 2, as well as those with greatest mean-variance.

That is, according to Table 2, the age of the start persona contributed to 2.8% of the utility compared to all other parameters (including latent). Covariates for observed attributes contributed to 61% of the total variance of the utility on average.



(a) (b)

Fig. 2: F1 scores for various models for train and test.

Table 2: Observed attributes with most significant mean-values and mean-variances (top 5).

Covariate on	Predictor	Mean contribution	Predictor	Variance contribution
start_persona	age	0.028	race_mixed	0.023
	task_abstract	0.028	race_other	0.022
	experience	0.027	ind_mining	0.020
	race_white	0.027	ind_agriculture	0.019
	task_routine	0.026	race_native_amer	0.018
end_persona	sex	0.004	race_mixed	0.023
	ind_personal_service	0.004	race_other	0.023
	task_manual	0.003	race_native_amer	0.020
	experience	0.003	ind_mining	0.018
	task_routine	0.002	ind_hawaiian_pacific	0.018

6 Discussion and future work

To improve the TTFM we will incorporate individual-level characteristics into the utility. Recall that the current utility model for TTFM defines the observed characteristics of a persona as the mean characteristics of persons ever in that persona-state. This was to solve the issue of not having enough observations per individual to be able to generalize what their covariates should be across time. However we can incorporate the individual persons attributes if we include an analogous term to Athey, et al.'s "distance" term into the utility model. Athey, et al. took the distance to be the geographic distance between a user i and restaurant j at time t. This essentially allows for a time-specific, time-varying observable trait to be included. With an analogous term our model would then be

$$U_{ijt} := U_{ijt} + \gamma_j^\top \eta_t \|Z_{is} - X_j\| \tag{6}$$

with Z_{is} being the observed traits of an individual s whose start state is persona i. Recall that X_j are the observed attributes of persona j. Then we estimate additional latent vectors γ_j and η_t for end-persona j and time t. While we could use the Euclidean norm for $\|Z_{is} - X_j\|$, another sensical choice would be the Mahalanobis distance, which weights the elements by their covariances. That is,

$$\|Z_{is} - X_j\|_{\text{Mahalanobis}} = \sqrt{(Z_{is} - X_j)^\top \Sigma^{-1} (Z_{is} - X_j)}. \tag{7}$$

Additionally, recall that the models with standard Gaussian priors did not perform better than the models with no priors (i.e. high-variance priors). We reasoned that this may be due to lack of "personalization" when assuming identical priors across personas. To account for this we could allow the observed attributes of each persona affect the prior distribution of that persona.

After improving upon our model we hope to use it to experiment with counterfactual questions, such as what would a person k's job prospects look like if they were to obtain an additional year of experience in the workforce. This experiment may be done by keeping the learned latent variables for personas the same, changing $Z_{is,\text{experience}} := Z_{is,\text{experience}} + 1$, and using the latent variables for the appropriate year to recalculate utilities.

7 Conclusion

We improve upon the existing, sparse literature on occupational transition probabilities by adapting a more rich model for our problem setting. The vast difference in performance between the TTFM and the more simple logit models show that incorporating start-state and end-state specific latent variables makes this prediction problem feasible. Future work lies in the improvement of this model and then in its use for counterfactual experiments.

8 Contributions

The main group members are Lilia Chang, Lisa Simon, and Karthik Rajkumar. Prof. Susan Athey is overseeing the project and lending guidance. Lisa and Lilia have contributed to the data cleaning and transformations. Lisa and Karthik were in charge of implementing the conditional logit models with Stata. Stata's models were unable to converge and so we moved to adapting the TTFM model to replicate the CLM. Lilia was in charge of the necessary data transformations for TTFM and of running the models presented in this paper. Acknowledgements go to Ayush Kanodia for his guidance in the implementation of the model.

9 Code

You may find the code relevant to this project in the zip file linked here. While the most recent version of TTFM is unable to be shared by the lab, a public version is available here.

References

1. Current population survey variables, documentation available at https://cps.ipums.org/cps-action/variables/group
2. Athey, S., Blei, D., Donnelly, R., Ruiz, F., Schmidt, T.: Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. American Economic Association Papers and Proceedings **108** (01 2018). https://doi.org/10.1257/pandp.20181031
3. Autor, D.H., Dorn, D.: The growth of low-skill service jobs and the polarization of the us labor market. American Economic Review **103**(5), 1553–97 (August 2013). https://doi.org/10.1257/aer.103.5.1553, http://www.aeaweb.org/articles?id=10.1257/aer.103.5.1553
4. Constant, A.F., Zimmermann, K.F.: Self-employment against employment or unemployment: Markov transitions across the business cycle. Eurasian Business Review **4**(1), 51–87 (2014). https://doi.org/10.1007/s40821-014-0005-x, https://doi.org/10.1007/s40821-014-0005-x
5. Donnelly, R., Ruiz, F.J.R., Blei, D.M., Athey, S.: Counterfactual inference for consumer choice across many product categories. CoRR **abs/1906.02635** (2019), http://arxiv.org/abs/1906.02635
6. Fabrizi, E., Mussida, C.: The determinants of labour market transitions. Giornale degli Economisti e Annali di Economia **68 (Anno 122)**(2), 233–265 (2009), http://www.jstor.org/stable/41954996
7. Ruiz, F.J.R., Athey, S., Blei, D.M.: Shopper: A probabilistic model of consumer choice with substitutes and complements (2017)