

Learning High-Dimensional Concave Utility Functions for Discrete Choice Models

Yuxue Qi[†] Min Xu^{*} John Lafferty[†]

[†]*Department of Statistics, University of Chicago, Chicago IL 60637*

^{*}*Machine Learning Department, Carnegie Mellon University, Pittsburgh PA 15213*

Abstract: The discrete choice model explains and predicts the behavior of a consumer who chooses one item to purchase from among a set of alternatives. The consumer is assumed to assign a utility to each item based on its features, and to choose an item with probability proportional to its utility. We consider the high-dimensional setting where the number of item features is large. It is important in this setting to identify and remove features that are irrelevant to the utility function. Not only is feature selection important to skirt the curse of dimensionality for statistical estimation, it is of interest to understand the factors on which people base their spending decisions. We show that, when the utility function has the diminishing returns property and is thus concave, an additive approximation is safe, in that in the population setting it does not erroneously remove relevant features. We propose a two stage feature screening method based on this result, evaluate our method on both simulation data and a novel job/housing survey dataset, and show that it is practical and effective.

1. Introduction

Many human behaviors can be modeled as a consumer selecting one item to purchase from among a set of alternatives. Examples include buying a product on Amazon, choosing the bus or car for commuting [Ortuzar and Willumsen \(1994\)](#), deciding where to buy a house [Nechyba and Strauss \(1998\)](#), and even choosing where to commit a crime [Bernasco and Block \(2009\)](#). The discrete choice model (DCM) originated in econometrics [McFadden \(1973\)](#) as a general method to model such finite choice problems. The DCM measures the attractiveness of item i to consumer n by a utility function $f(\mathbf{x}_i, \mathbf{s}_n)$ where $\mathbf{x}_i, \mathbf{s}_n$ are feature vectors of the item and the consumer, respectively. The consumer is more likely to pick item i over the alternatives if the utility $f(\mathbf{x}_i, \mathbf{s}_n)$ is higher. The utility function in the DCM is estimated from a dataset of purchases; each purchase consists of a consumer, a set of items, and the consumer's choice from that set. The AI and machine learning communities have in recent years rediscovered the DCM as a form of *preference learning* [Chu and Ghahramani \(2005\)](#); [Fürrnkranz and Hüllermeier \(2010\)](#).

Because it has become easier to extract and store information digitally, the number of features in a modern dataset is often large, possibly larger than the number of samples. Variable selection becomes important, where an estimation technique must select and use

only a small set of relevant variables to avoid the well known curse of dimensionality. Variable selection among the item features \mathbf{x}_i is especially important in the DCM, as people tend to make decisions based on a few important cues or factors [Shah and Oppenheimer \(2008\)](#). Good variable selection methods give insight into how consumers make choices.

We assume that the utility function $V(\mathbf{x}_i, \mathbf{s}_n)$ is decomposed as $f(\mathbf{x}_i) + h(\mathbf{s}_n)$ and focus on the estimation of $f(\mathbf{x}_i)$. We suppose $f(\mathbf{x}_i)$ obeys certain shape-constraints, mainly concavity. We do not assume that f is additive, but we show that for the purpose of screening out irrelevant variables, it is safe to approximate the possibly non-additive f with an additive concave model, followed by a sequence of decoupled convex models to catch non-concave residuals.

We prove that this procedure, in the population setting, is *faithful* in that it will not erroneously mark a relevant variable as irrelevant. The assumptions we make on the underlying density are mild, and do not restrict correlations between the variables. This is in contrast to linear models where, if the true function is non-linear, one must make stronger covariance structure assumptions in order to provide the same guarantee. While estimation of a low-dimensional concave utility function for the DCM is studied by Matzkin using parametric distributional assumptions [Matzkin \(1991\)](#), we are unaware of previous results on variable selection in the DCM in the high dimensional nonparametric setting that we study in this paper.

The utility function is often assumed to be linear [McFadden et al. \(1978\)](#); [Nechyba and Strauss \(1998\)](#) but a concavity assumption is less restrictive. In many economics applications, the concavity assumption is popular and natural because of the law of diminishing returns. For example, Nechyba and Strauss [Nechyba and Strauss \(1998\)](#) represent the attractiveness of a community in the DCM with features such as per-pupil school spending. The law of diminishing returns in this case states that once a school spends enough per pupil, further spending will be less effective and thus effect a smaller increase in a household's utility.

Though our estimation method is a nonparametric generalization of the linear model, it requires no additional tuning parameters, such as the smoothing bandwidth that makes local polynomial methods difficult. Concavity (and other similar shape-constraints) thus offers an attractive computational compromise between a parametric and fully nonparametric model. We formulate a convex optimization in the infinite dimensional constraint space of concave functions, which reduces to a finite dimensional space of piecewise linear functions.

2. Discrete Choice Model

In discrete choice model, each consumer n chooses one item out of a set \mathcal{A}_n of alternatives based on the utility-maximization principle: each item $i \in \mathcal{A}_n$ has a utility U_{ni} and the consumer chooses item i if $U_{ni} \geq U_{nj}$ for all $j \in \mathcal{A}_n$ (ties broken arbitrarily). The utility U_{ni} is unobservable but is assumed to equal a function of some observable features of the

items and of the consumer plus noise

$$U_{ni} = V_{ni} + \epsilon_{ni} = V(\mathbf{x}_i, \mathbf{s}_n) + \epsilon_{ni}.$$

The vector \mathbf{x}_i denotes features of item i , \mathbf{s}_n denotes features of consumer n , and ϵ_{ni} denotes the noise term. The probability of consumer n choosing item i depends on the assumptions on the distribution of the noise vector $\epsilon_n = (\epsilon_{n1}, \dots, \epsilon_{n|\mathcal{A}_n|})$.

$$P_{ni} = \mathbb{P}(U_{ni} > U_{nj}, \forall j \neq i) = \mathbb{P}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}, \forall j \neq i)$$

For example, $\epsilon_n \sim_{iid}$ Gaussian yields the probit model, $\epsilon \sim_{iid}$ extreme value yields the logit model. In this paper, we consider the logit model, also known as the Bradley-Terry model. The probability consumer n chooses item i under the logit model has the expression:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j \in \mathcal{A}_n} \exp(V_{nj})} = \frac{\exp(V(\mathbf{x}_i, \mathbf{s}_n))}{\sum_{j \in \mathcal{A}_n} \exp(V(\mathbf{x}_j, \mathbf{s}_n))}$$

We follow the standard assumption that the representative utility function $V(\mathbf{x}_i, \mathbf{s}_n)$ is decomposed additively as $f(\mathbf{x}_i) + h(\mathbf{s}_n)$ and focus on the estimation of $f(\mathbf{x}_i)$, similar to Chu and Ghahramani [Chu and Ghahramani \(2005\)](#). We assume h can be modeled well and the results in this paper hold regardless of how one chooses to model $h(\mathbf{s}_n)$.

We assume that $f(\mathbf{x}_i)$ is concave, which is strictly more general than the usual linear assumption and is justified by the principle of diminishing returns present in many economics applications.¹ Since discrete variables are common in practice and concavity applies only to functions of continuous variables, we add a linear term $\alpha^T \mathbf{z}_i$ for the discrete features \mathbf{z}_i :

$$V(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_n) = f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i + h(\mathbf{s}_n)$$

We can then model the consumer choices by

$$\mathbb{P}(\text{consumer } n \text{ chooses } i \mid \mathcal{A}_n) = \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i + h(\mathbf{s}_n))}{\sum_{j \in \mathcal{A}_n} \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j + h(\mathbf{s}_n))}$$

The unknown f , α , h can be estimated from a dataset of purchases. We represent each purchase by a vector $\mathbf{y}_n = (y_{ni})_{i \in \mathcal{A}_n}$; $y_{ni} = 1$ iff consumer n chooses item i . For notational simplicity, we assume that each consumer makes exactly one purchase. It is straightforward to extend the model to cases where each consumer makes multiple purchases.

Given N purchases, the likelihood under the logit DCM is $\frac{1}{N} \sum_{i=1}^N \ell(\mathbf{y}_n \mid \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n)$ where

$$\ell(\mathbf{y}_n \mid \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) = \sum_{i \in \mathcal{A}_n} y_{ni} (f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i + h(\mathbf{s}_n)) - \log \left(\sum_{i \in \mathcal{A}_n} \exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i + h(\mathbf{s}_n)) \right)$$

¹our results readily apply to the estimation of $h(\mathbf{s}_n)$ in the cases where h can be assumed concave.

3. Additive Faithfulness

Let d_1 and d_2 denote the number of features in \mathbf{x}_i and \mathbf{z}_i respectively. In the high-dimensional setting where d_1 and d_2 are large, it is necessary to select a small subset $S_1 \subset \{1, \dots, d_1\}$ and $S_2 \subset \{1, \dots, d_2\}$ such that $f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i \approx f(\mathbf{x}_{S_1,i}) + \alpha_{S_2}^T \mathbf{z}_{S_2,i}$ where $\mathbf{x}_{S_1,i}$ is the restriction of the vector \mathbf{x}_i to coordinates in S_1 .

The lasso effectively tackles high-dimensional problems by adding an ℓ_1 penalty on the linear coefficients to the likelihood maximization. The natural extension for high-dimensional concave function estimation is to add to the likelihood a group ℓ_1 penalty on the *subgradients* of f :

$$\begin{aligned} \underset{\mathbf{f}, \beta, \gamma, \mathbf{h}}{\text{minimize}} \quad & - \sum_{n=1}^N \ell(\mathbf{y}_n | \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda_1 \sum_{k=1}^{d_1} \|\beta_{k\cdot}\|_\infty + \lambda_2 \|\alpha\|_1 \\ \text{subject to} \quad & f_j \leq f_i + \beta_i^T (x_j - x_i), \text{ for all } i, j \end{aligned}$$

where f_i is the estimated function value $f(\mathbf{x}_i)$ and the vector $\beta_i \in \mathbb{R}^{d_1}$ is the subgradient at \mathbf{x}_i .

This convex optimization problem has $O(Mp)$ variables and $O(M^2)$ constraints, leading to a potentially cumbersome and computationally inefficient method. In the following section, we will use additive functions to approximate $f(\cdot)$ and argue that concave functions are additively faithful with respect to variable selection, under very mild conditions. The resulting variable selection framework is much more computationally efficient.

3.1. Additive Faithfulness of Concave Functions

The *additive approximation* to a multivariate function f is a sum of one-dimensional functions f_k such that $\sum_{k=1}^d f_k(\mathbf{x}_k)$ approximates $f(\mathbf{x})$. In general, if the true model is non-additive, an additive approximation may introduce false negatives and cause potential misspecification problems. However, we show that concave functions have a unique property: as long as the true function we approximate is concave and monotone on the boundary, we can safely mark as irrelevant any variable that is zeroed out by the optimization algorithm. In other words, it is *faithful* in terms of variable selection under an additive approximation. Before giving our main result, which makes this precise, we begin with a lemma that characterizes the components of the optimal additive approximation.

For notational simplicity, we suppose that $|\mathcal{A}_n| = m$ for all n . We assume that each purchase, which comprises $(\{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^m, \mathbf{s}_n)$, is iid drawn from some distribution F with density p . For this section, we use k to index features, i, j to index items, and n to index consumers.

Lemma 3.1. *Let F be a distribution on $[0, 1]^{md_1} \times \{0, 1\}^{md_2}$ with a mixed density p . Let $f : [0, 1]^{d_1} \rightarrow \mathbb{R}$ be an integrable true function of the items. Define the following for any*

fixed γ, h :

$$\{f_k^*\}_{k=1}^{d_1} = \arg \min_{\{f_k\}_{k=1}^{d_1}} \mathbb{E} \left[-\sum_{i=1}^m Y_i V(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_n) + \log \left(\sum_{j=1}^m \exp(V(\mathbf{x}_j, \mathbf{z}_j, \mathbf{s}_n)) \right) \right] \quad (3.1)$$

$$\text{where} \quad V(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_n) = \sum_{k=1}^{d_1} f_k(x_{ki}) + \gamma^T \mathbf{z}_i + h(\mathbf{s}_n)$$

$$Y_i | \mathbf{x}, \mathbf{z}, \mathbf{s}_n \sim \text{Bernoulli} \left(\frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i + h(\mathbf{s}_n))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j + h(\mathbf{s}_n))} \right).$$

Then f_k^* satisfies

$$\mathbb{E} \left[\frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i, \mathbf{s}_n))}{\sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j, \mathbf{s}_n))} - \frac{\exp(f(x_{ki}, \mathbf{x}_{-k,i}) + \alpha^T \mathbf{z}_i + h(\mathbf{s}_n))}{\sum_{j=1}^m \exp(f(x_{kj}, \mathbf{x}_{-k,j}) + \alpha^T \mathbf{z}_j + h(\mathbf{s}_n))} \middle| x_{ki} \right] = 0,$$

where $\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i, \mathbf{s}_n) = \sum_{k' \neq k} f_{k'}(x_{k'i}) + \gamma^T \mathbf{z}_i + h(\mathbf{s}_n)$, $\varphi(\mathbf{x}_i, \mathbf{s}_n) = \sum_{k=1}^d f_k(x_{ki}) + h(\mathbf{s}_n)$. Furthermore, this solution is unique.

Lemma 3.1 easily follows from the stationary conditions of the optimal solution. Lemma 3.1 states the intuitive fact that the first moment conditional on x_{ki} under the true model must equal that of the optimally fitted f_k^* . We now give our main result and the accompanying assumptions.

Definition 3.1. Let $f : [0, 1]^{d_1} \rightarrow \mathbb{R}$ be an integrable function, f is boundary monotone if for all k and all \mathbf{x}_{-k}

$$\partial_{x_k} f \geq 0 \quad \text{or} \quad \partial_{x_k} f \leq 0 \quad \text{at the boundary } x_k = 0 \text{ and } x_k = 1$$

Definition 3.2. Let p be a mixed density supported on $[0, 1]^{md_1} \times \{0, 1\}^{md_2}$, p satisfies the boundary-points condition, if

$$\frac{\partial}{\partial x_{ji}} p(\mathbf{x}_{-j,i}, \{\mathbf{x}_l\}_{l \neq i}, \{\mathbf{z}_i\}_{i=1}^m | x_{ji}) = 0 \quad \text{at } x_{ji} = 0 \text{ and } x_{ji} = 1, \quad \text{for any } \mathbf{x}_{-j,i}, \{\mathbf{x}_l\}_{l \neq i}, \{\mathbf{z}_i\}_{i=1}^m$$

Theorem 3.1. (Additive Faithfulness) Let p be a positive mixed density supported on $[0, 1]^{md} \times \{0, 1\}^{md'}$ that satisfies the boundary-points property (definition 3.2). Suppose f is concave, boundary-monotone (definition 3.1) and differentiable.

Fix arbitrary γ, h , let $\{f_k^*\}_{k=1}^{d_1}$ be the optimal additive components as defined in equation 3.1. Then $f_k^* = 0$ implies that $\partial_{x_k} f(\mathbf{x}) = 0$, that is, f does not depend on feature k .

Theorem 3.1 is the main theoretical result of this paper. It states that even if the true function f is not additive, the additive approximation yields no false negatives. We defer the proof to section 7 of the appendix.

It is important to note that additive faithfulness does not rely on any restrictions of the correlation structure between the covariates. The only distributional assumption we

make is the mild boundary-point condition (definition 3.2). We allow the density to behave arbitrarily in the interior of the support. In contrast, in linear regression where $\beta^* = \Sigma^{-1}\mathbb{E}[Xf(X)]$, we would need to restrict the covariance to make the same faithfulness guarantee.

The boundary monotone condition (definition 3.1) is reasonable in applications where the concavity assumption is natural. With respect to some features, such as the per-pupil school spending in Nechyba and Strauss [Nechyba and Strauss \(1998\)](#), the utility function is monotone and thus boundary monotone as well. Boundary monotone condition also holds for features of which people want more when there is too little (one boundary point) and less when there is too much (the other boundary point). For instance, people distrust extremely cheap items and refrain from extremely expensive items.

Theorem 3.1 does not give a way to estimate whether $f_k^* = 0$. The next section tackles this problem.

3.2. Concave Additive Model

Since the true function f is concave, it is natural to consider a concave additive model. For notational simplicity, we let γ be arbitrarily fixed and omit $h(s_n)$ in this section.

$$\{\tilde{f}_k^*\}_{k=1}^d = \arg \min_{\tilde{f}_k \in -\mathcal{C}^1} \mathbb{E} \left[- \sum_{i=1}^m Y_i \left(\sum_{k=1}^d \tilde{f}_k(x_{ki}) + \gamma^T \mathbf{z}_i \right) + \log \sum_{j=1}^m \exp \left(\sum_{k=1}^d \tilde{f}_k(x_{kj}) + \gamma^T \mathbf{z}_j \right) \right]$$

where we use $\mathcal{C}^1, -\mathcal{C}^1$ to denote the set of univariate convex and concave functions respectively.

Concave additive components \tilde{f}_k^* are not additively faithful, but we can restore faithfulness by coupling the \tilde{f}_k^* 's with a set of convex functions:

$$g_k^* = \arg \min_{g_k \in \mathcal{C}^1} \mathbb{E} \left[- \sum_{i=1}^m Y_i (g_k(x_{ki}) + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) + \log \left(\sum_{j=1}^m \exp (g_k(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) \right) \right]$$

where $\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i) = \sum_{k' \neq k} f_{k'}(x_{k'i}) + \gamma^T \mathbf{z}_i$.

Theorem 3.2. Suppose $p(\mathbf{x}, \mathbf{z})$ is a mixed positive density on $C \times \{0, 1\}^{md_2}$, where $C \subset \mathbb{R}^{md}$ is a compact set and $p(\mathbf{x}, \mathbf{z})$ satisfies the boundary-points condition. Suppose that $\partial_{x_{ki}} f(\mathbf{x}_i)$, $\partial_{x_{ki}}^2 f(\mathbf{x}_i)$, $\partial_{x_{ki}} p(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j \neq i}, \{\mathbf{z}\}_{i=1}^m | x_{ki})$, and $\partial_{x_{ki}}^2 p(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j \neq i}, \{\mathbf{z}\}_{i=1}^m | x_{ki})$ are all continuous on C . Then $\tilde{f}_k^* = 0$ and $g_k^* = 0$ only if f does not depend on x_k , i.e. $\partial_{x_k} f(\mathbf{x}) = 0$ with probability 1.

We defer the proof of theorem 3.2 to section 8 of the appendix. Theorem 3.2 states that if a covariate is relevant, then at least one of the optimal concave and convex functions that minimizes the negative likelihood should be nonzero. Therefore, if we fit 0 for both the convex and concave component, we can safely zero out the corresponding variable and claim it as irrelevant. Intuitively, the convex \hat{g}_k^* “catches” any non-concave residual that \tilde{f}_k^* could not capture.

4. Estimation Procedure

Theorem 3.2 motivates a two stage procedure for variable selection. In the first stage, we fit a sparse additive concave function under the logistic DCM framework. We then separately fit a convex function on the residuals for each dimension.

Importantly, we do not introduce tuning parameters for smoothing the function. Such smoothing parameters are essential to most nonparametric estimation methods, but are typically very difficult to set. In particular, there is no easy way to optimally adjust smoothing parameters in a traditional additive model, based for example on kernel regression or smoothing splines. This is a key attraction of the shape-constrained approach.

Given sample $\{\mathbf{x}_{ni}, \mathbf{z}_{ni}, \mathbf{s}_n, \mathbf{y}_n\}_{i \in \mathcal{A}_n}^{n=1, \dots, N}$, the following procedure, referred to as AC/DC (additively concave/decoupled convex), is performed.

AC Stage: Compute, jointly

$$\hat{f}_1, \dots, \hat{f}_{d_1}, \hat{\gamma}, \hat{h} = \arg \min_{f_1, \dots, f_{d_1} \in \mathcal{C}^1, \gamma, h} -\frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}_n | \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda_1 \sum_{k=1}^p \|\beta_{k\cdot}\|_{\infty} + \lambda_2 \|\gamma\|_1 \quad (4.1)$$

where

$$\ell(\mathbf{y}_n | \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) = \sum_{i \in \mathcal{A}_n} y_{ni} \hat{V}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_n) - \log \left(\sum_{j \in \mathcal{A}_n} \exp(\hat{V}(\mathbf{x}_j, \mathbf{z}_j, \mathbf{s}_n)) \right)$$

$$\hat{V}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{s}_n) = \sum_k f_k(x_{ki}) + \gamma^T \mathbf{z}_i + h(\mathbf{s}_n)$$

and $\beta_{k\cdot}$ are the corresponding subgradients of $f_k(\cdot)$.

DC Stage: Compute, separately, for each k where $\|\beta_{k\cdot}\|_{\infty} = 0$

$$\hat{g}_k = \arg \min_{g_k \in \mathcal{C}^1} -\frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}_n | \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda_1 \|\tilde{\beta}_{k\cdot}\|_{\infty} + \lambda_2 \|\gamma\|_1 \quad (4.2)$$

where $\tilde{\beta}_{k\cdot}$ are the corresponding subgradients of $g_k(\cdot)$. We then output as the set of continuous relevant variables $\{k : \|\beta_{k\cdot}\|_{\infty} > 0 \text{ or } \|\tilde{\beta}_{k\cdot}\|_{\infty} > 0\}$ and of discrete relevant variables $\{k' : \gamma_{k'} \neq 0\}$

We adopted an ℓ_{∞}/ℓ_1 penalty in 4.1 and the ℓ_{∞} penalty in 4.2 to encourage sparsity. In the AC stage (4.1), any estimation method for h can be used.

4.1. Optimization

We describe the optimization algorithm only for the additive concave logistic regression stage, the second decoupled convex logistic regression stage is a straightforward modification. We observe that a univariate concave function is characterized by non-increasing

subgradients. So we form our optimization problem as

$$\begin{aligned}
& \underset{\mathbf{f}, \beta, \gamma, \mathbf{h}}{\text{minimize}} && - \sum_{n=1}^N \ell(\mathbf{y}_n | \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda_1 \sum_{k=1}^{d_1} \|\beta_k\|_\infty + \lambda_2 \|\gamma\|_1 \\
& \text{subject to} && f_{k(i+1)} = f_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}) \\
& && \sum_{i=1}^M f_{ki} = 0, \quad \beta_{k(i+1)} \leq \beta_{k(i)}, (\forall k, i)
\end{aligned} \tag{4.3}$$

$$\ell(\mathbf{y}_n | \mathbf{X}_{\mathcal{A}_n}, \mathbf{Z}_{\mathcal{A}_n}, \mathbf{s}_n) \equiv \sum_{i \in \mathcal{A}_n} y_{ni} \left(\sum_{k=1}^{d_1} f_k(x_{ki}) + \gamma^T \mathbf{z}_i + h_n \right) - \log \left(\sum_{j \in \mathcal{A}_n} \exp \left(\sum_{k=1}^{d_1} f_k(x_{kj}) + \gamma^T \mathbf{z}_j + h_n \right) \right)$$

where $\{(1), (2), \dots, (M)\}$ is a reordering of $\{1, 2, \dots, M\}$ such that $x_{k(1)} \leq x_{k(2)} \leq \dots \leq x_{k(n)}$. We use the centering constraints $\sum_{i=1}^M f_k(x_{ki}) = 0$ for identifiability.

Motivated by the shooting algorithm for the lasso [Friedman, Hastie and Tibshirani \(2010\)](#), we solve optimization 4.3 with block coordinate descent. In the outer loop, we repeatedly iterate between optimization of (\mathbf{f}, β) and of γ . When estimating \mathbf{f} (or γ), we again in the inner loop iteratively select a dimension k , fix all $\mathbf{f}_{k'}$ (or $\gamma_{k'}$) for $k' \neq k$, and optimize $\{f_k(x_{ki})\}_{i=1, \dots, M}$ (or γ_k). For each inner loop iteration, we apply Newton's method and solve a sequence of quadratic programs. We use the optimization software MOSEK to solve the intermediate QPs in our implementation. In the cases where $h_n = h(\mathbf{s}_n)$ must be estimated as well, we would iterate between $(\mathbf{f}, \beta), \gamma, \mathbf{h}$ in the outer loop and, depending on choice of model for $h(\mathbf{s}_n)$, any appropriate optimization algorithm can be used to optimize \mathbf{h} in a step of the outer loop.

The estimated function can be evaluated on an input item \mathbf{x}_j with the equation $f(\mathbf{x}_j) = \sum_{k=1}^{d_1} f_k(x_{kj}) = \sum_{k=1}^{d_1} \min_i \{f_{ki} + \beta_{ki}(x_{kj} - x_{ki})\}$. For univariate convex function estimation, we modify the linear inequality so that the subgradients are non-decreasing: $\beta_{k(i+1)} \geq \beta_{k(i)}$.

5. Experiment

We evaluate AC/DC on both synthetic data experiments as well as a novel survey dataset. For all of our experiments we do not consider consumer features, i.e., we omit the $h(\mathbf{s}_n)$ term.

5.1. Simulation

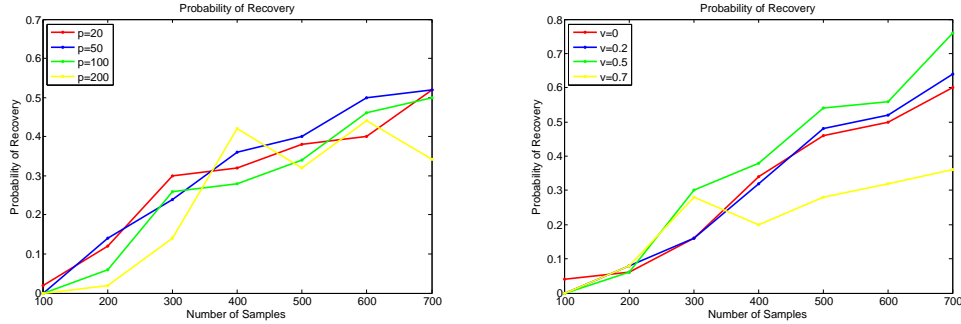
For the M items, we generate continuous feature vectors $\mathbf{x}_1, \dots, \mathbf{x}_M \sim P$ where the distribution $P = c\bar{N}(0, \Sigma) + (1-c)U$ is a mixture between a multivariate Gaussian distribution thresholded to lie in $[-b, b]^{p_1}$ and an uniform distribution supported on $[-b', b']^{p_1}$ where $b' > b$ to fulfill the boundary condition. By ‘‘thresholded’’, we mean that if the Gaussian

sample is greater than b , then we set it equal to b . The discrete feature vectors $\mathbf{z}_1, \dots, \mathbf{z}_M$ are generated similarly, except that we discretize the vectors by setting a continuous value to zero if it is less than 0.

The true utility function is taken to be the sum of a piecewise linear function of the continuous features \mathbf{x}_{S_1} and a linear function of the discrete features \mathbf{z}_{S_2} , where the piecewise linear function is guaranteed to be concave and S_1, S_2 represent the corresponding active feature sets with $|S_1| = |S_2| = 3$. In the simulations, we take $\Sigma_{ij} = \nu^{|i-j|}$ for various ν 's, and pick the set of active features at random to create varying amounts of correlation between relevant and irrelevant variables. In addition, we always set $\lambda_1 = \sqrt{\frac{\log(Np_1)}{N}}$ and $\lambda_2 = 0.3\sqrt{\frac{\log(Np_1)}{N}}$.

In the first simulation, we fix $\nu = 0.3$. We vary $N = 100, 200, \dots, 700$, $p_1 = 20, 50, 100, 200$. For each (N, p_1) , we generate 50 independent datasets and apply AC/DC procedure to infer the function estimates \mathbf{f} , subgradients β , and discrete coefficients γ . We declare correct support recovery, if for $\forall k \in S_1$, $\|\beta_k\|_\infty > 10^{-6}$, $\forall k \notin S_1$, $\|\beta_k\|_\infty < 10^{-6}$ and for $\forall k' \in S_2$, $|\gamma_{k'}| > 10^{-6}$, $\forall k' \notin S_2$, $|\gamma_{k'}| < 10^{-6}$. We show the plot of correct support recovery probability versus different combinations of N and p_1 in figure 1a. As can be seen, the ACDC algorithm achieves higher support recovery rate as sample size increases even when p is large.

In the second simulation, we fix $p_1 = 15$ and investigate the robustness of the ACDC algorithm over different correlation structures. N varies from 100, 200, \dots , 700 and ν varies from 0, 0.2, 0.5, 0.7. As before, we generate 50 data sets and compute the probability of correct support recovery for each combination of N and ν . The results are shown in figure 1b and demonstrate that ACDC can still select relevant variables well for design of moderate correlation.



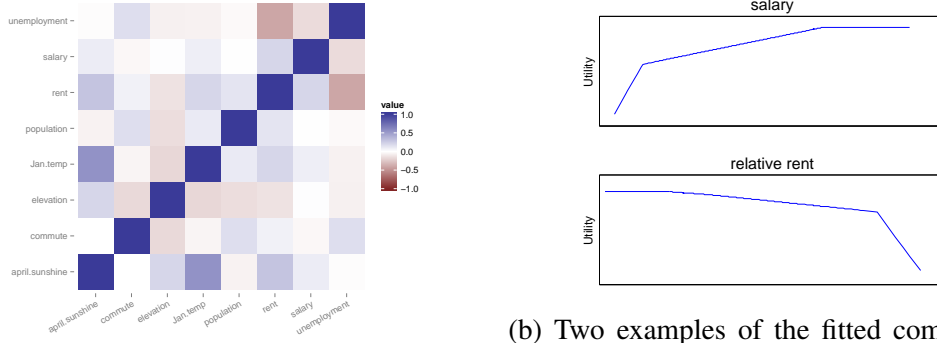
(a) Support recovery results for various p (b) Support recovery results for various ν

Fig 1

5.2. Survey Data

This dataset consists of 530 surveys we gave to the students and staff at our university. Each survey contains three options, each of which is a hypothetical living arrangement

that consists of a job and an apartment in some neighborhood within some city. We ask the respondents to choose the one they most prefer. Each living arrangement is described on the survey by three types of information: personal level, city level, and neighborhood level. The personal level information are *yearly salary*, *monthly rent*, and *commute time*. The city level information are *year round temperature*, *population*, *robbery rate*, and *diabetes rate*. The neighborhood level information are *average income*, *unemployment rate*, and *the percent of college graduates*.



(a) Correlation among some features shown functions. Relative rent is rent divided by average income of the neighborhood.

(b) Two examples of the fitted component functions.

Fig 2

	ACDC	linear
Features in Survey	83.3%	76.9%
Features in Survey (minus % diabetes)	80.8 %	75.7 %

(a) Percentage of features selected that are among the features given on the surveys.

	concave	linear
8 features	0.670	0.680
3 features	0.680	0.686

(b) Negative log-likelihood of models fitted using either the top 8 features or the top 3 features of the feature selection process.

TABLE 1

We created the surveys by gathering information on 68 US cities and a total of 148 zipcode regions (which we call neighborhoods) within those cities. The information is gathered from www.city-data.com. We generate each living arrangement by randomly selecting a city and a zipcode region and then generating a random salary, rent, and commute time based on the average in that zipcode neighborhood. The reader can find examples of the survey as well as more detail about how we made the surveys in section 9 of the appendix.

Feature selection evaluation. In addition to the features shown on the survey, we collected various other features of the cities and zipcode regions we used. These additional features are *July humidity level*, *January snowfall*, *April sunshine rate*, *% households gay/lesbian*, *% households unmarried*, *elevation*, *air quality index*, and *% voted Obama in 2008*. Because these features were not shown on the survey and not known to the respon-

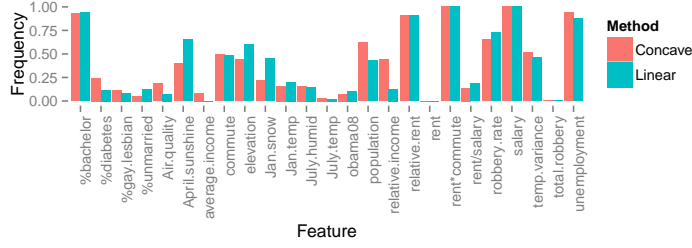


Fig 3: Variable selection frequency for survey data

dents, they are by construction irrelevant to the survey responses. These irrelevant features are, however, correlated with the survey features (Figure 2a).

We evaluate AC/DC by taking random subsamples of the data, performing variable selection, and measuring how often the features shown on the survey are marked as relevant. We compare AC/DC against the sparse standard logistic DCM where we let $f(\mathbf{x}_i) = \beta^T \mathbf{x}_i$ and apply the ℓ_1 penalty on β . The regularization parameters are selected so that on average 9.5 features are selected. From the 400 surveys in the training data, we took 94 random subsamples of 200 surveys and performed both AC/DC and sparse linear model on these subsamples. In addition to the raw features, we also added some interaction terms among the relevant variables.

The results are shown in Figure 3 and Table 1a. AC/DC outperforms the linear model in choosing features that are relevant to the survey. We included *%diabetes* as a feature on the survey though it is unlikely to play a part in a respondent’s decision process. Thus, in the second row of Table 1a, we exclude *%diabetes* as a relevant variable. Not surprisingly, the two features selected with 100% frequency are salary and a rent-times-commute interaction term.

Heldout likelihood evaluation. To ensure that the concavity assumption is reasonable and that we are not overfitting to the training data, we also evaluate the log-likelihood of our estimated model on a heldout dataset of 114 surveys. These surveys use information only from cities that *do not appear in any of the training data surveys*. We use the top 3 or the top 8 features in selection process and refit an additive concave model, unregularized, on the training data, using only those features (likewise with the sparse linear model). For features whose monotonicity in the utility is obvious, we also add a monotone constraint when refitting. Table 1b shows that concave monotone model performs slightly better. Though the improvement is small, the concave monotone model using 3 features achieves the same likelihood as the linear model using 8 features. We show two examples of the fitted functions in Figure 2b. Both salary and the relative rent (rent / average income) exhibit concavity.

6. Discussion

We have developed a smoothing parameter free semiparametric variable selection procedure for learning concave utility functions in the context of the logistic DCM. Our current

focus is on the estimation of concave part of the utility function. However, it would be interesting to further explore the consumer grouping behavior via a multi-task learning procedure, as in [Birlutiu, Groot and Heskes \(2013\)](#). Another interesting direction is to automatically detect whether an additive component is convex or concave.

References

- BERNASCO, W. and BLOCK, R. (2009). Where offenders choose to attack: A discrete choice model of robberies in Chicago. *Criminology* **47** 93–130.
- BIRLUTIU, A., GROOT, P. and HESKES, T. (2013). Efficiently learning the preferences of people. *Machine Learning* **90** 1–28.
- CHU, W. and GHAHRAMANI, Z. (2005). Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning* 137–144. ACM.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33** 1.
- FÜRNKRANZ, J. and HÜLLERMEIER, E. (2010). *Preference learning*. Springer.
- MATZKIN, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society* 1315–1327.
- MCFADDEN, D. (1973). Conditional logit analysis of qualitative choice behavior.
- MCFADDEN, D. and et al. (1978). *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.
- NECHYBA, T. J. and STRAUSS, R. P. (1998). Community choice and local public services: A discrete choice approach. *Regional Science and Urban Economics* **28** 51–73.
- ORTUZAR, J. D. and WILLUMSEN, L. G. (1994). *Modelling transport*.
- SHAH, A. K. and OPPENHEIMER, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological bulletin* **134** 207.

7. Proof of Theorem 3.1

For simplicity, we omit the consumer function $h(\mathbf{s}_n)$ in the proofs.

Proof. (of Theorem 3.1) W.T.S. $f_k^* = c \Rightarrow \frac{\partial}{\partial x_{ki}} f = 0$. Now we assume that for all x_{ki} , (3.1) holds and $f_k^*(x_{ki}) = f_k^*(x_{kj}), \forall j \neq i$. Differentiating with respect to x_{ki} under the integral gives:

$$\begin{aligned} & \sum_{\{\mathbf{z}_i\}_{i=1}^m} \int p'(\mathbf{x}_{-k,i}, \mathbf{z}_i, \{\mathbf{x}_j, \mathbf{z}_j\}_{j \neq i} | x_{ki}) \left[\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(x_{ki}, \mathbf{x}_{-k,i}) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \right] \\ & + p(\mathbf{x}_{-k,i}, \mathbf{z}_i, \{\mathbf{x}_j, \mathbf{z}_j\}_{j \neq i} | x_{ki}) \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i) f'(\mathbf{x}_i) \sum_{j \neq i} \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)}{(\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j))^2} d\mathbf{x}_{-(k,i)} = 0 \end{aligned}$$

If p satisfies the boundary-points condition, then the integral equation can be reduced to:

$$\sum_{\{\mathbf{z}_i\}_{i=1}^m} \int p(\mathbf{x}_{-k,i}, \mathbf{z}_i, \{\mathbf{x}_j, \mathbf{z}_j\}_{j \neq i} | x_{ki}) \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i) f'(\mathbf{x}_i) \sum_{j \neq i} \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)}{(\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j))^2} d\mathbf{x}_{-(k,i)} = 0$$

Recall that f is boundary-monotone, without loss of generality we can assume $f'(x_{ki}^0, \mathbf{x}_{-k,i}) \geq 0$. Also, by the positive density assumption, $p(\mathbf{x}_{-k,i}, \mathbf{z}_i, \{\mathbf{x}_j, \mathbf{z}_j\}_{j \neq i} | x_{ki}) > 0$. So we have $f'(x_{ki}^0, \mathbf{x}_{-k,i}) = 0$ for all \mathbf{x}_{-k} . Likewise, $f'(x_{ki}^1, \mathbf{x}_{-k,i}) = 0$ for all $\mathbf{x}_{-k,i}$.

Because $f(x_{ki}, \mathbf{x}_{-k,i})$ as a function of x_{ki} is concave, it must be that, for all $x_{ki} \in (0, 1)$ and for all $\mathbf{x}_{-k,i}$:

$$0 = f'(x_{ki}^1, \mathbf{x}_{-k,i}) \leq f'(x_{ki}, \mathbf{x}_{-k,i}) \leq f'(x_{ki}^0, \mathbf{x}_{-k,i}) = 0$$

Therefore, f does not depend on x_k . □

8. Proof of Theorem 3.2

Proof. From Theorem 3.1, it suffices to show that $f_k^* = 0$.

Now suppose $\tilde{f}_k^* = g_k^* = 0$. First consider the univariate function $h_k(x_{ki}) = \delta e^{-x_{ki}}$, where $\delta \in \mathbb{R}$. $h_k(x_{ki})$ is convex and decreasing if $\delta > 0$, concave and increasing if $\delta < 0$. Since $\tilde{f}_k^* = g_k^* = 0$, then

$$\begin{aligned} & \arg \min_{\delta \in \mathbb{R}} \left\{ \mathbb{E} \left[- \sum_{i=1}^m Y_i (\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) + \log \left(\sum_{i=1}^m \exp(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) \right) \right] \right\} \\ & = \arg \min_{\delta \in \mathbb{R}} \left\{ \mathbb{E} \left[\log \left(\sum_{i=1}^m \exp(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) \right) - \sum_{i=1}^m p_i (\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) \right] \right\} \\ & = 0 \end{aligned}$$

where

$$p_i = \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)}$$

Recall that the objective function is convex in δ , the stationary condition gives us:

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^m e^{-x_{ki}} \left(\frac{\exp(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\delta e^{-x_{kj}} + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \right) \right] &= 0 \\ \stackrel{\delta^* = 0}{\implies} \mathbb{E} \left[\sum_{i=1}^m e^{-x_{ki}} \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \right) \right] &= 0 \end{aligned}$$

It is not hard to prove that $f_k^*(x_{ki})$ has lower bounded derivatives $f_k^{*'}(x_{ki})$ and $f_k^{*''}(x_{ki})$. Then we can always find an η such that $e^{-x_{ki}} + \eta f_k^*(x_{ki})$ is convex and non-increasing. Therefore, by a similar argument, we have

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^m (e^{-x_{ki}} + \eta f_k^*(x_{ki})) \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \right) \right] &= 0 \\ \implies \mathbb{E} \left[\sum_{i=1}^m f_k^*(x_{ki}) \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \right) \right] &= 0 \\ \implies \mathbb{E} \left[\sum_{i=1}^m f_k^*(x_{ki}) \mathbb{E} \left[\left(\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \right) \middle| x_{ki} \right] \right] &= 0 \end{aligned}$$

Recall that $f_k^*(x_{ki})$ is a unique function that satisfies

$$\mathbb{E} \left[\frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f(\mathbf{x}_i) + \alpha^T \mathbf{z}_i)}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + \alpha^T \mathbf{z}_j)} \middle| x_{ki} \right] = 0.$$

Then we have

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^m f_k^*(x_{ki}) \mathbb{E} \left[\left(\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} \right) \middle| x_{ki} \right] \right] &= 0 \\ \implies \mathbb{E} \left[\sum_{i=1}^m f_k^*(x_{ki}) \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} - \frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}, \mathbf{z}_i))}{\sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} \right) \right] &= 0 \\ \implies \mathbb{E} \left[\sum_{i=1}^m \frac{f_k^*(x_{ki}) \exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) \sum_{j \neq i} \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) (\exp(f_k^*(x_{kj}) - \exp(f_k^*(x_{ki})))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) \sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j))} \right] &= 0 \end{aligned}$$

Note that

$$\begin{aligned}
& \sum_{i=1}^m f_k^*(x_{ki}) \exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i)) \sum_{j \neq i} \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) (\exp(f_k^*(x_{kj}) - \exp(f_k^*(x_{ki}))) \\
&= \sum_{i=1}^m \sum_{j \neq i} f_k^*(x_{ki}) \exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) (\exp(f_k^*(x_{kj}) - \exp(f_k^*(x_{ki}))) \\
&= \sum_{i < j} \exp(\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) \left(f_k^*(x_{ki}) - f_k^*(x_{kj}) \right) \left(\exp(f_k^*(x_{kj}) - \exp(f_k^*(x_{ki}))) \right) \\
&\leq 0 \quad \text{since} \quad \left(f_k^*(x_{ki}) - f_k^*(x_{kj}) \right) \left(\exp(f_k^*(x_{kj}) - \exp(f_k^*(x_{ki}))) \leq 0
\end{aligned}$$

and

$$\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) \sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{z}_j)) > 0.$$

Thus we have

$$\left(f_k^*(x_{ki}) - f_k^*(x_{kj}) \right) \left(\exp(f_k^*(x_{kj}) - \exp(f_k^*(x_{ki}))) = 0, \text{ for all } i \neq j$$

i.e. $f_k^*(x_{ki}) = f_k^*(x_{kj})$, for all $i \neq j$. □

[1]	bridgeport	orlando	columbus	jacksonville	dallas
[6]	charlotte	reno	portland	durham	denver
[11]	jersey_city	paradise	spokane	rockford	chesapeake
[16]	chicago	cambridge	austin	seattle	raleigh
[21]	allentown	berkeley	philadelphia	pittsburgh	boston
[26]	san_diego	las_vegas	lynn	atlanta	richmond
[31]	cincinnati	warren	madison	houston	san_antonio
[36]	miami	fremont	nyc	albany	la
[41]	newark	vancouver	sf	detroit	aurora
[46]	stamford	ann_arbor	springfield	grand_rapids	elizabeth
[51]	eugene	milwaukee	cleveland	new_haven	dc
[56]	boulder	henderson	buffalo		

Fig 4: List of the cities used in the training dataset surveys.

[1]	kansas	minneapolis	baltimore	phoenix	fort_wayne
[6]	indianapolis	manchester	st_louis	st_paul	norfolk

Fig 5: List of the cities used in the test dataset surveys.

9. Survey Detail

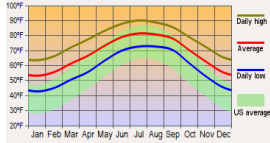
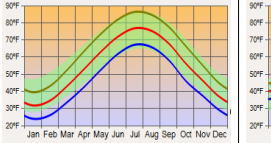
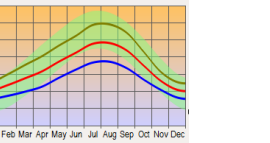
Figure 6 shows examples of surveys we handed out to respondents. Each survey contains three random living arrangements. The living arrangements are sampled randomly from a large collection that we generated from actual city and zipcode region data. To generate a living arrangement, we take the following steps:

1. We select a random city and a random zipcode region. All features of the living arrangement except *salary*, *commute time*, and *rent* are simply the features of the selected city and the zipcode region.
2. We generate a random salary $s = s_{base} \cdot c_{adj} \cdot c_{noise}$ where $s_{base} \sim Unif\{69K, 80K\}$. $c_{adj} = \left(\frac{\text{national average income}}{\text{region average income}} \right)^{0.15}$ is the regional wealth adjustment; richer regions yield higher salaries. $c_{noise} \sim N(0, 0.15^2)$ is a Gaussian multiplier noise.
3. We generate a random rent $r = r_{base} c_{noise}$ where r_{base} is the average rent of the zipcode region and $c_{noise} \sim N(0, 0.15^2)$ is a Gaussian multiplier noise.

• **Gender:** (circle one): M F Other N/A

• **Major:** Social Science Physical Science Math/Stats Language Biology Economics Other Undecided

You have just completed your Bachelor's degree. You have a choice of three different jobs in three different locations in the US. All other considerations being equal, which one of the following living situations do you most prefer?

	Situation A	Situation B	Situation C
Salary	69,000\$	79,000\$	88,000\$
Rent	690\$	710\$	760\$
Commute Time	28 min	24 min	26 min
Temperature			
City Population (Chicago: 3,000,000)	837,000	1,550,000	603,000
City Robbery Rate (per 100,000) (New York City: 250)	163	519	159
Diabetes Rate (national: 11%)	10.4%	10.4%	7.0%
Neighborhood Income (national: 44,000)	41,000\$	23,000\$	56,000\$
Neighbor. %College Grad (national: 32%)	10.6%	11.5%	54.7%
Neighbor. %Unemploy (national: 8.1%)	11.2%	17.6%	10.3%
Response (check one)			

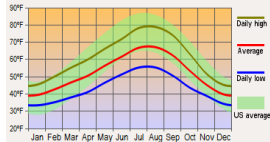
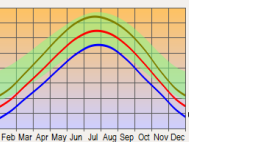
Survey ID: 1

1

• **Gender:** (circle one): M F Other N/A

• **Major:** Social Science Physical Science Math/Stats Language Biology Economics Other Undecided

You have just completed your Bachelor's degree. You have a choice of three different jobs in three different locations in the US. All other considerations being equal, which one of the following living situations do you most prefer?

	Situation A	Situation B	Situation C
Salary	84,000\$	76,000\$	64,000\$
Rent	1,100\$	1,500\$	870\$
Commute Time	12 min	34 min	26 min
Temperature			
City Population (Chicago: 3,000,000)	165,000	636,000	2,710,000
City Robbery Rate (per 100,000) (New York City: 250)	103	303	498
Diabetes Rate (national: 11%)	8.2%	8.0%	8.4%
Neighborhood Income (national: 44,000)	57,000\$	61,000\$	54,000\$
Neighbor. %College Grad (national: 32%)	22.1%	49.7%	36.1%
Neighbor. %Unemploy (national: 8.1%)	10.2%	6.5%	7.5%
Response (check one)			

Survey ID: 2

1

Fig 6: Example surveys