Sufficient Representations for Categorical Variables

Jonathan Johannemann
jonjoh@stanford.edu

Vitor Hadad
vitorh@stanford.edu

Susan Athey
athey@stanford.edu

Stefan Wager
swager@stanford.edu

Stanford University

Abstract

[[[REDO]]] We propose a solution via what we call the sufficient latent state assumption which seeks to describe the relationship between covariates X_i, response Y_i, and observable groups G_i. We explore how this assumption can be used to develop sufficient representations of dimension $k \ll |\mathcal{G}|$ for universally consistent estimators. We then show promising results for these representations in both simulated and empirical datasets.

1 Introduction

Many regression problems involve data collected from a number groups that may be statistically relevant. For example, in a medical setting we may want to model health outcomes using data on patients from several hospitals, and acknowledge that different hospitals may have idiosyncratic effects on patients that are not explained by other covariates. Similar considerations arise when working with data on students from different schools, voters from different zip-codes, employees at different firms, etc.

One of the most wide-spread approaches to this problem is via fixed effect modeling, as follows. Suppose that we observe n samples (X_i, G_i, Y_i) for $i = 1, \cdots, n$, where $X_i \in \mathbb{R}^p$ is a set of subject-specific covariates, $G_i \in \mathcal{G}$ is a categorial variable that records group membership and $Y_i \in \mathbb{R}$ is the respond of in interest, and that we want to estimate

$$\mu(x, g) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x, G_i = g\right]. \tag{1}$$

Then, the simple fixed effects approach starts by positing a model

$$\mu(x, g) = x\beta + \alpha_g, \tag{2}$$

and then estimating the coefficients β and α_g via ordinary least squares regression. More sophisticated extensions of this approach may involve considering non-linear transformations of x, interactions between group membership and the covariates x, and/or regularization [Angrist and Pischke, 2008, Diggle et al., 2002, Wooldridge, 2010].

Fixed effects modeling, however, does not always perform well with complex non-linear signals or when the number of groups $|\mathcal{G}|$ is large. The model (2) is quite rigid and may not be able to represent rich signals while; and, at the same time, the large number of α_g

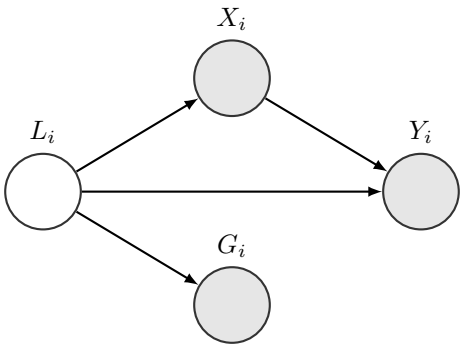

Figure 1: Causal graph depicting the key assumption that Y_i and X_i are independent of group membership G_i conditionally on latent state L_i. The grayed-out nodes are observed.

parameters in the model (2) may result in problems for statistical inference [Neyman and Scott, 1948]. In other words, the model (2) may have too many parameters to be stable all while lacking the degrees of freedom to fit the signal well.

The goal of this paper is to develop more parsimonious approach for representing group membership avoids the above problems. Specifically, we seek a mapping ψ that embeds group membership G_i into a k-dimensional space without losing any predictive information, i.e.,

$$\psi : \mathcal{G} \to \mathbb{R}^k, \quad \mu(x, g) = f(x, \psi(g)), \tag{3}$$

such that k is small (in particular, $k \ll |\mathcal{G}|$) and the function $f(\cdot, \cdot)$ is still easy to learn. Given such a mapping, the problem (1) becomes a routine regression problem with $(p + k)$-dimensional real-valued features $(X_i, \psi(G_i))$, and we can use out-of-the-box statistical learning software on it.

In order to obtain a useful representation of group membership G_i, we of course need to assume something about the relationship between G_i and the target outcome Y_i. The core assumption made in this paper is what we call the *sufficient latent state assumption* depicted in Figure 1: group membership G_i has no direct causal effect on Y_i, but may be associated with latent variables L_i that do have a direct effect on Y_i. For example, in the case of patients spread across hospitals, we assume that hospitals themselves do not directly *cause* health outcomes affect Y_i; however, hospitals may still be *predictive* of Y_i through their association with latent causal variables. For example, patients may have unobserved characteristics, e.g., severity of disease or socioeconomic resources, that both affect Y_i and lead the patient to self-select into different hospitals. Our main result is that, under this sufficient latent state assumption, practical representations of the form (3) exist and can be learned from data.

The principle of representing high-cardinality categorical variables as real-valued vectors has played an important role in many different areas. For example, in natural language processing, it now common to start more complex analyses with a pre-processing step that represent words as vectors that capture the way in which the words are used in context [Mikolov et al., 2013, Pennington et al., 2014]. Meanwhile, in the literature on panel data analysis, our approach is perhaps most closely related to a proposal of Bonhomme and Manresa [2015] where individual time series belong to discrete clusters and we have only one fixed effect per cluster (rather than one per time series). Bonhomme and Manresa [2015]

then fit this model via a k-means like algorithm that alternates clustering and estimation with per-cluster fixed effects.[1] The resulting framework provides a means for introducing lower dimensional, sufficient representations of categorical variables.

Our paper is structured as follows. We begin by reviewing similar problem settings in the fixed effects literature and the drawbacks of using existing methods in 1.1. In Section 2, we introduce the primary lemma which seeks to describe the true information we wish to extract from categorical variables. In Section 3, we expand on lemma 1 to develop methods that utilize this insight. In Sections 4 and 5, we run simulated and observational experiments with our proposed methods and follow up with discussion on how realized performance compared to our expectations.

1.1 Related Work

Traditionally, the discussion of how best to account group membership G_i in a non-parametric regression has focused on how to different ways to encode G_i in a way that can be given as an input to statistical software. One simple way to do so is via one-hot encoding: $\omega : \mathcal{G} \to \{0, 1\}^M$ such that the j-th entry of $\omega(g)$ is 1 if and only if g corresponds to the j-th element in $\mathcal{G}$, and where $M := |\mathcal{G}|$. Note that linear regression on one-hot encoded features $(X_i, \omega(G_i))$ exactly recovers the standard fixed effects model (2).

As discussed above, however, one-hot encoding may lead to undesirably high-dimensional problems when $|\mathcal{G}|$ is large. [2] In the Appendix (6.4), we present multiple encoding methods that similarly project $\omega : \mathcal{G} \to \mathbb{R}^{|\mathcal{G}|}$. These methods do not utilize information from the covariates X_i or response Y_i and suffer from the same pitfalls that come with high dimensional representation of the observed groups $\mathcal{G}$. The primary difference for these methods are the user's interpretation of the encoded variables which are commonly constructed as the comparison of the mean effect of a subset $\mathcal{G}' \in \mathcal{G}$ relative to the mean effect of the set $\mathcal{G} \backslash \mathcal{G}'$ or one of its subsets.

The problem of fixed effects is especially challenging with sparsity-seeking methods such as the lasso [Hastie et al., 2015] or decision trees [Breiman et al., 1984], and related ensemble methods such as random forests [Breiman, 2001] or gradient-boosted trees [Friedman, 2001]. Sparsity-seeking methods will set the contribution of features to zero unless there is strong evidence that the features matters for prediction, and it is difficult for rare levels of G_i to produce sufficient evidence to get a non-zero contribution to the model via their one-hot features. The end result is that sparsity seeking methods may largely ignore high-cardinality one-hot encoded factors.

Another prevalent way of working with categorical variables with decision trees is to consider full factorial splits that allow for arbitrary grouping of the levels of the categorical variable. For a variable with $M := |\mathcal{G}|$ levels, this allows for $2^{M-1} - 1$ potential splits. Breiman et al. [1984] showed that we can optimize over this exponential set of potential splits in time that scales linearly in M; however, from a statistical point of view, such factorial splits are prone to very strong overfitting when the number of levels is large.

[1] Our approach is not directly comparable to either of these methods, as we do not focus on textual data, and do not assume that the latent state L_i can be consistently estimated (in contrast, Bonhomme and Manresa [2015] assume that they have access to long enough time series that their clustering step is consistent which, in our setting, would be equivalent to assuming that L_i can be recovered).

[2] Another slightly more subtle difficulty is that when the categorical variable has many levels, the individual features $\omega(G_i)_j$ become very sparse (i.e., they are usually 0 and only very rarely 1). Many approaches to statistical learning work better with features whose variance roughly captures their range than with such spiky features.

Recently, Cerda et al. [2018] consider a related problem of representing "dirty" categorical variables that might arise if, e.g., several categorical levels are just misspellings of each other, and propose using a low-dimensional embedding that exploits lexicographic similarity (i.e., factors with similar spellings are arranged close to each other). In this paper, we use information in the X_i, rather than lexicographic information, to construct an embedding; however, the high-level conclusion that we can achieve meaningful gains by using auxiliary information to embed categorical variables in a low-rank space remains.

We also note, it is sometimes possible to achieve strong results by randomly projecting a one-hot representation of the categorical variables into $\mathbb{R}^k$ [Rahimi and Recht, 2008]. We also consider this approach but find that, at least in our experiments, we can achieve better performance using carefully crafted representations that leverage continuous covariates X_i.

2 Representing Groups with Sufficient Latent State

Our *sufficient latent state* assumption presented in the introduction and depicted in the causal graph 1 implies that the distribution of the outcome Y_i only depends on the observable factor G_i through some unobservable latent variable $L_i \in \{\text{good}, \text{poor}\}$. In other words, if we knew the value of L_i, then also knowing G_i would give us no additional information about the outcome. For a simple example, one may posit that a patient's underlying health status ($L_i \in \{\text{good}, \text{poor}\}$) may simultaneously determine to which hospital they are admitted (G_i), what symptoms (X_i) they exhibit, and what health outcomes outcomes (Y_i) they attain. Conditioned on the underlying health status, the hospital cannot provide any additional information about any of the other variables. Conversely, learning their hospital is only helpful inasmuch it allows us to infer something about their health status.

The following lemma states that we can say further characterize *how* the information about the categorical variable G_i enter the model: the conditional expectation function of the outcome depends only on the *conditional probabilities of the latent variable given the observable category.*

Lemma 1. *Suppose that the latent state L_i is discrete with k possible levels, and that the probabilistic structure require by the sufficient latent state assumption (Figure 1) holds. Then,*

$$\psi : \mathcal{G} \to \mathbb{R}^k, \quad \psi_l(g) = \mathbb{P}\left[L_i = l \,\middle|\, G_i = g\right] \tag{4}$$

provides a sufficient representation of G_i in the sense of (3):

$$\mu(x,\, g) = \frac{\sum_{l=1}^{k} \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, L_i = l\right] \mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] \psi_l(g)}{\sum_{l=1}^{k} \mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] \psi_l(g)}. \tag{5}$$

Expression (5) formalizes the intution laid out in the previous paragraph. The information associated with the category only enters the conditional expectation via the set of probabilities $\mathbb{P}\left[L_i = \ell \,\middle|\, G_i = g\right]$. If there are only K latent groups, then each category can be represented in a lossless manner by K dimensional vector of probabilities. An immediate consequence of this result is that if we knew ψ and gave training examples $((X_i,\, \psi(G_i)),\, Y_i)$ to any universally consistent learner, the learner would eventually recover the optimal prediction function $\mu(\cdot)$. To continue the example at the top of this section, the identity of the hospital enters the model through the probability that a patient is in good or poor health given the hospital.

The dependence of the conditional expectation function μ on the latent variable probabilities ψ via (5) is non-linear; however, we will retain consistency if we use an expressive enough method for learning on $((X_i, \psi(G_i)), Y_i)$. Methods known to be universally consistent include k-nearest neighbors [Stone, 1977], various tree-based ensembles [Biau et al., 2008], and neural networks [Faragó and Lugosi, 1993].

The discussion above may seem to imply that we need to estimate $\psi(g)$ directly. However, this quantity depends on the unobservable variable L_i and its identification is impossible without further assumptions and more sophisticated approach. Instead, we pursue a different approach based on the following fact: if $h : \mathbb{R}^K \to \mathbb{R}^K$ is a function that possesses a left-inverse $h^\dagger \circ h = \mathrm{id}$, then $h \circ \psi$ must also be a sufficient representation, since we can define $f(x, \psi(g)) = \tilde{f}(x, h^\dagger(h(\psi(g))))$. Therefore, in this paper, we focus on simple approaches that just rely on finding simple functions sufficients representations of the form $h(\psi(g))$ and can be estimated using observable data X_i and G_i.

3 Categorical variable encoding methods

Below we introduce methods that exploit the structure mentioned in the previous section. For an overview of other categorical encoding methods already in use, please see section 6.4 in the Appendix. All of the methods below take the form of removing the categorical column and replacing it with a set of columns that can be proven to encode the same information.

3.1 Means encoding

For our first method, we drop the categorical variables G_i and substitute in the average value of the continuous regressors X_i given the categorical variable. Figure 2 shows an illustration.

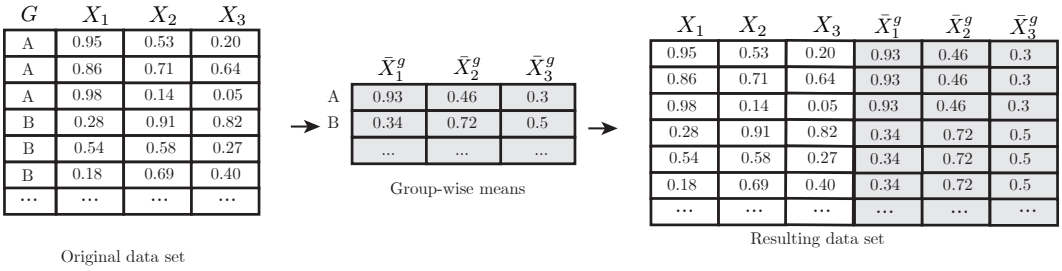

Figure 2: Implementation example of the *means* encoding.

This representation is easily interpretable, and it is simple to implement efficiently. This method may be particularly applicable in instances where the number of regressors p is small, and in particular if $p \ll |\mathcal{G}|$, since then the dimensionality reduction is more dramatic as compared to traditional encoding methods such as one-hot encoding. Conversely, if the dimension of X_i is larger than the number of levels of the latent factor, then $\omega(g)$ may be a needlessly high-dimensional representation of g. However, due to its simplicity, using this representation may still be desirable in practice.

Lemma 2. *Under the conditions of Lemma 1, suppose in addition that the matrix A defined by $(A)_{tj} := \mathbb{E}\left[X_{it} \,\middle|\, L_i = g\right]$ is left-invertible. Then, $\omega(g) = \mathbb{E}\left[X_i \,\middle|\, G_i = g\right]$ is sufficient*

in the sense of (3)*:*

$$\mu(x,\, g) = \frac{\sum_{l=1}^{k} \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, L_i = l\right] \mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] (A^{\dagger}\omega(g))_l}{\sum_{l=1}^{k} \mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] (A^{\dagger}\omega(g))_l} \tag{6}$$

Algorithm 1 Means Encoding Method

procedure MEANSENCODING(X, G)
  $\Omega \leftarrow 0_{n \times p}$
  $\bar{X} \leftarrow 0_{|\mathcal{G}| \times p}$
  for g in 1:$|\mathcal{G}|$ **do** ▷ Compute the group averages
    $\bar{X}_g \leftarrow \frac{1}{|\{i:G_i=g\}|} \sum_{i:G_i=g} X_i$
  for i in 1:n **do** ▷ Populate with group averages
    $\Omega_{i,\cdot} \leftarrow \bar{X}_{G_i}$
  return Ψ

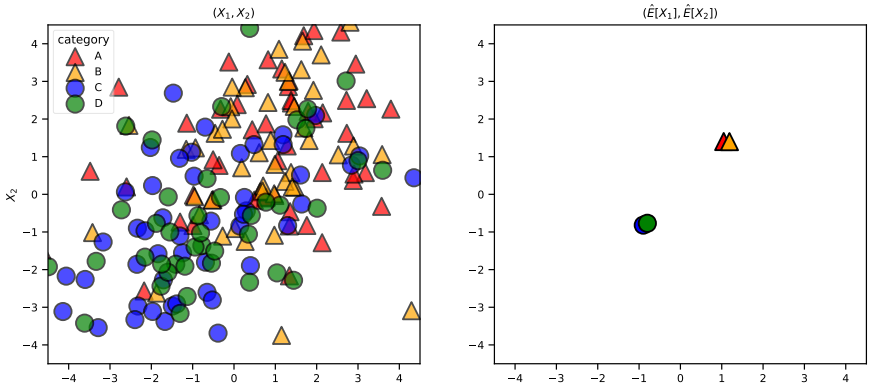

Figure 3: Intution for the *means* encoding on an illustrative data: the averages of the continuous variables (X_1, X_2) may reveal that the categories belong to distinct latent groups.

3.2 Low-rank encodings

The *means* encoding method described above may efficiently summarize the effect of the categorical variables if the continuous covariates are reasonably low-dimensional so that $p \ll M$. When p is large, it might be beneficial to use lower-dimensional representation of the conditional means. As opposed to depending on the covariates as is, we propose utilizing methods such as PCA or Sparse PCA Zou et al. [2006] which can provide informative linear combinations of the original X_i.

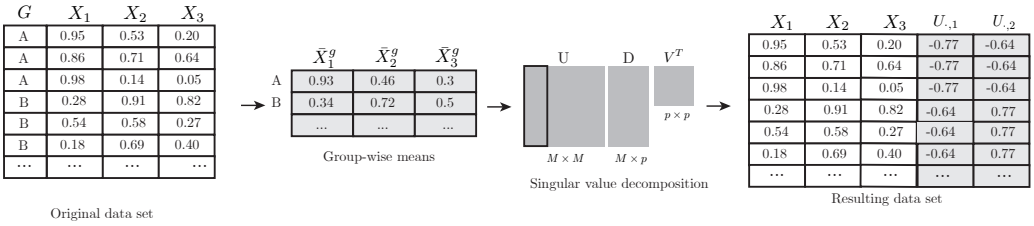

Figure 4: Implementation example of the *low-rank* encoding with singular value decomposition. Alternatively, we could also have used sparse PCA in place of SVD.

We suggest two approaches that use common matrix factorization methods. The first method is as follows:

$$\psi : G_i \mapsto U_{i,1:k} \tag{7}$$

where U_i is the left singular matrix in the singular value decomposition of the transpose of $E[X_i|G_i]$. Each row in U corresponds to an observable group G_i and, through cross validation, one can choose an appropriate value k to approximate the relevant columns in the $|\mathcal{G}| \times |\mathcal{G}|$ matrix.

Second, we propose the Sparse PCA encoding method which is similar but one can induce sparsity in the principal component vectors and the matrix B satisfies the objective Zou et al. [2006]:

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{i=1}^{n} ||X_i - AB^T X_i||^2 + \lambda \sum_{j=1}^{k} ||\beta_j||^2 + \sum_{j=1}^{k} \lambda_{1,j} ||\beta_j||_1 \tag{8}$$

$$\text{s.t.} \quad A^T A = I_{k \times k} \tag{9}$$

By either picking the number of principal components that reconstruct 95% of the original variance in X or using cross validation to select the number of principal components, we can achieve an encoding with dimensionality $d < p$. We provide the outline for the procedure in **??** which is the same as **??** except for the carefully chosen number of principal components as opposed to the original covariates X.

While SVD can provide additional dimensionality reduction, there are caveats that come with the method. First, this process assumes that the first k columns of the left singular matrix comprise of the relevant vectors to separate the latent groups. In Cook et al. [2007], Cook talks about the idea of using the first r number of principal components in regression in the hope that all relevant information with respect to the response is kept. The same concept holds for the latent groups' relevant covariates and therefore there is the possibly that those variables are contained in the last few singular vectors. Therefore, for this method, we highly recommend using cross validation to ensure its efficacy.

Next, tree-based models such as random forest Breiman [2001] or xgboost Chen and Guestrin [2016], which fit to data by considering singular covariates at any point in the model, may have difficulty taking advantage of these approaches due to the potential rotation of the original covariate space. This can result in separability in two or more dimensions but not one dimension.

Furthermore, the advantage of using sparse PCA over SVD is that the rotation in X is restricted since the orthogonal vectors are penalized. This results in sparse, more interpretable principal component vectors and reduces the potential for difficulties arising from

rotations of the original covariate space. However, the tradeoff for these benefits is that sparse PCA requires additional tuning of the λ parameters.

Algorithm 2 Low Rank Encoding Method

procedure LowRankEncoding(X, G, k)
  $M \leftarrow$ MeansEncoding(X, G) ▷ Compute the group averages
  $U, D, V^T \leftarrow SVD(M)$
  $S \leftarrow 0_{n \times k}$
  for i in 1:n **do** ▷ Populate with left singular matrix truncated rows
    $S_{i,\cdot} \leftarrow U_{G_i, 1:k}$
  return S

Algorithm 3 Sparse Low Rank Encoding Method

procedure SparseLowRankEncoding(X, G, k)
  $M \leftarrow$ MeansEncoding(X, G) ▷ Compute the group averages
  $A, B \leftarrow SPCA(M)$
  $Z \leftarrow M \cdot B_{\cdot, 1:k}$
  $S \leftarrow 0_{n \times k}$
  for i in 1:n **do** ▷ Populate with sparse principal components rows
    $S_{i,\cdot} \leftarrow Z_{G_i,\cdot}$
  return S

3.3 Encoding by multinomial logistic regression coefficients

For our third and last method, we propose estimating the conditional probability of category membership by multinomial logistic regression parametrized by coefficients $\{\theta_g\}_{g \in \mathcal{G}}$ (including an intercept)

$$P(G_i|X_i) = \Lambda_\theta(G_i = g|X_i) = \frac{\exp(\tilde{X}_i^T \theta_g)}{\sum_{g'} \exp(\tilde{X}_i^T \theta_{g'})} \qquad \tilde{X}_i := (1, X_i) \tag{10}$$

and then use the $(p+1)$-dimensional vector of coefficients associated with g^{th} category to represent it.

The main inspiration for this method is the natural language processing *word2vec* model first presented by Mikolov et al. [2013]. In that work, the authors propose two methods to represent words (categories) in a large corpus as relatively low-dimensional real-valued vectors. In one of these methods, each word is initially assigned two representations: as a center word v_w, and as a surrounding *context* word v_c. Then, the authors posit that the optimal representation is that maximizes the log-probability of the inner product of the two representations $v_w^T v_c$ for all pair of words (w, c) that co-occur near each other. Our method works in an analogous way if we consider the continuous vectors $\tilde{X}_i := (1, X_i)$ as "contexts", and let $\theta_g \in \mathbb{R}^p$ represent each category, since then maximizing the log-probability of the inner product $\tilde{X}_i^T \theta_g$ is the same as maximizing the multinomial logistic regression above.

Our method can also be seen to be a variation on the *inverse regression* method of Taddy et al. [2014]. Those authors make the following proposition:

$$y_i \perp\!\!\!\perp x_i \mid v_i \implies y_i \perp\!\!\!\perp x_i \mid \Phi' x_i$$

where $x_i = \Phi \cdot v_i + \epsilon_i$ and Φ is a $p \times K$ matrix which maps relevant information in y_i to x_i. In our case, we map all relevant information from G to X. Furthermore, for the case without subject effects, Taddy states that Φ can alternatively be interpreted as the *population average effect* ov v on x which has the same interpretation as our Means encoding method.

Lemma 3. *Under the conditions of Lemma 1, suppose in addition that A in (30) is left-invertible, and that $\mathbb{P}\left[G_i = g \,\middle|\, X_i\right]$ is the multinomial logit distribution with coefficients $\{\theta_g\}_{g \in \mathcal{G}}$ containing an intercept. Then, the vector $\theta_g \in \mathbb{R}^{p+1}$ is sufficient in the sense of (3):*

$$\mu(x,\, g) = \frac{\sum_{l=1}^{k} \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, L_i = l\right] \mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] (A^\dagger f(\theta_g))_l}{\sum_{l=1}^{k} \mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] (A^\dagger f(\theta_g))_l} \tag{11}$$

$$\textit{where} \quad f(\theta_g) := (E[X_i \Lambda_\theta(G_i | X_i)^T])_{\cdot, g} \tag{12}$$

4 Experiments

In the following section, we explore each method's effectiveness relative to one hot encoding across simulated and real world data sets. We apply two typically used methods random forests and xgboost. Both methods have a natural interpretation of separating A_{jl} with splits into buckets that correspond to the different latent groups which may be similar socioeconomic status, health condition, or consumer preference.

4.1 Simulations

We consider two simulations designs that share the distributions of latent groups L_i, observable groups G_i and covariates X_i, but whose outcome models for Y_i differ.

As an example of a real-life problem that our choice of simulation design tries to illustrate, consider the problem of predicting health outcomes Y_i from patient characteristics X_i and hospital G_i, with the latent groups L_i representing the patients' unobservable health status. Depending on their health care needs, healthier patients may concentrate on different hospitals. For example, patient with chronic diseases may select larger hospitals, whereas patients with more harmless diseases may choose to visit their local clinic.

Latent groups, observable groups and continuous covariates Latent groups L_i is drawn uniformly from the set of available groups, which we identify with integers.

$$L_i \sim \text{Uniform}(\{1, ..., |\mathcal{L}|\}) \tag{13}$$

Next, observable groups G_i are drawn according to the following rule. First, we partition the set of possible observable groups $\mathbb{G}$ into equally-sized sets $\{\mathbb{G}_\ell\}_{\ell=1}^{|\mathcal{L}|}$. Then, we draw the observable group G_i so that observations that were assigned latent group L_1 have higher

probability of falling into observable group $\mathbb{G}_1$, those in L_2 likely belong to $\mathbb{G}_2$, and so on. In symbols,

$$P(G_i = g \mid L_i) = \begin{cases} \frac{p_{L_i}}{|\mathbb{G}_{L_i}|} & \text{if} \quad g \in \mathbb{G}_{L_i} \\ \frac{1-p_{L_i}}{|\mathbb{G}^C_{L_i}|} & \text{otherwise} \end{cases} \qquad \text{where} \quad p_{L_i} > 0.5 \tag{14}$$

Observable covariates X_i are Normally distributed with unit variance, but some of their means are shifted depending on the latent group to which they belong.

$$(E[X_i|L_i])_j = \begin{cases} +1 & \text{with prob.} \quad 0.05 \\ -1 & \text{with prob.} \quad 0.05 \\ 0 & \text{with prob.} \quad 0.9 \end{cases} \qquad Var(X_i|L_i) = I_{p\times p} \tag{15}$$

Outcomes The effect of each latent group is a mean shift that is shared among the latent group.

$$\alpha_{L_i} \sim \text{Uniform}([-1, 1]) \tag{16}$$

In the *linear setup*, the outcome model is described as follows.

$$\beta_{L_i} = [U_1, \cdots, U_{\lceil 0.6p \rceil}, 0, ..., 0]^T \tag{17}$$
$$U_k \sim \text{Uniform}(\{-1, +1\}) \tag{18}$$
$$Y_i = X_i\beta_{L_i} + \eta_L\alpha_{L_i} + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon) \tag{19}$$

The parameter η_L is chosen so that $Var(\eta_L(\alpha_{L_i})) = Var(X_i\beta)$, that is, that the portion of the signal coming from the covariates and from the latent group are comparable. Lastly, σ_ϵ is adjusted so that the signal to noise ratio is kept at a fixed level.

In the *Interactions setup*, we introduce simple non-linearities in the outcome model via interactions between covariates and the observable group.

$$j_k \sim \text{Uniform}([1, ..., p]) \quad \text{for } k \in \{1, \cdots, \lfloor\sqrt{p}\rfloor\} \tag{20}$$
$$\beta_k \sim \text{Uniform}([-1, 1]) \tag{21}$$
$$Y_i = \eta_e(X_{i,j(1)}\alpha_{L_i}) + \sum_{k=1}^{\lfloor\sqrt{p}\rfloor} \beta_n \cdot X_{j(2k)} \cdot X_{j(2k+1)} + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon) \tag{22}$$

Again, similarly to the previous simulation, the parameter η_e is chosen so that the variance of the first and second terms in equation 22, and σ_ϵ is chosen so that the signal-to-noise ratio is fixed at a predetermined value.

4.2 Simulation Results

We simulated one hundred datasets for each combination of parameters on Table 1. For each simulated dataset, we estimated the outcome using the various methods described in Section 3, and then evaluated the predictions by their mean squared error.

Parameter	Values
Observations (n)	5000
Continuous covariates (p)	20
Rhos (ρ)	0.25
Signal to noise ratio	.5
Own-group probability (p_L)	.9
Latent groups ($\|\mathcal{L}\|$)	2, 10
Observable categories ($\|\mathcal{G}\|$)	100, 500

Table 1: Simulation parameters to create grid of inputs.

Results for the simulations are provided in 6. We find that the methods described above which seek to estimate the latent groups consistently outperform methods which require adding $|\mathcal{G}|$ additional columns for the Regression Forest and, for larger numbers of latent groups, XGBoost as well. We then find that the permutation, fisher, and multiple permutation methods are on average much better than the methods that add $|\mathcal{G}|$ columns but still slightly fall behind the methods that estimate the latent groups. Also, in most cases we find that multiple permutation does better than fisher and permutation.

For the methods that do take advantage of the low rank structure, we notice that the main improvement in performance for Regression Forests occurs due to the immense reduction in dimensionality. While the performance improvements over one hot encoding for 100 observable groups ranges from 2-5%, performance improvement can approach 15-18% for 500 observable groups. Intuitively, we find that this benefit is generally less prevalent for 2 latent groups for both XGBoost and Regression Forests due to the lesser complexity of the underlying relationship as defined by the conditional independence graph in 1. Furthermore, we find that MNL tends to do worse which appears to be due to the fact that the underlying method requires model estimation and is therefore more data-intensive than other methods. For the XGBoost based evaluation, due to the model's ability to express more complex functions than normal tree-based methods, we find that there is little or negative benefit when adding our encoding methods when the number of latent groups is rather small. However, this changes when there are a larger number of latent groups, but not by a substantial amount.

4.3 Empirical Applications

We also evaluate these methods on publicly available datasets that are accessible on Kaggle. We run 4 fold, stratified cross validation on the datasets to avoid the case where there are categorical variables in the test set which are not contained in the training set.

Pakistan Educational Performance In Hemani [2017], Hemani consolidated a series of surveys from Alif Ailaan, a nonprofit organization in Pakistan that focuses on improving education across the country. The objective of the surveys was to provide an objective means of comparing school systems across cities and provinces in order to spark competition between local governments to spur educational reform.

The dataset used in the analysis below contains $n = 580$ and 504 after removing null valued rows which can be broken down into $|\mathcal{G}| = 127$ cities from 2013 to 2016. The number of additional covariates p is 20 and our estimation methodology is further elaborated on in

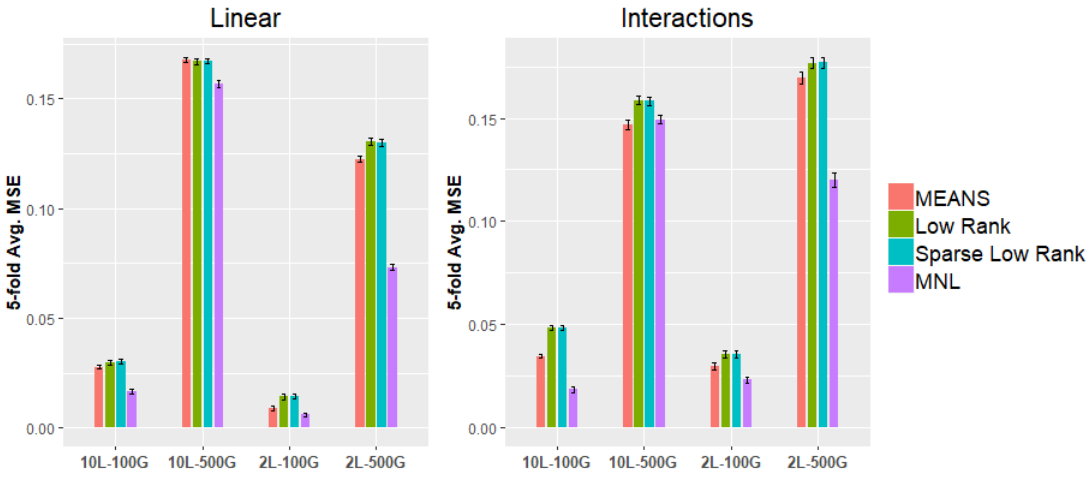

Figure 5: The percent improvement for each setup and varied latent groups and observable categories with fixed rho, signal to noise ratio, and own-group probability.

Appendix 6.5.

Ames Housing The objective of the Ames Housing Dataset De Cock [2011] was to act as a more complex alternative to the Boston Housing Dataset Harrison and Rubinfeld [1978]. De Cock's aim was to use this dataset for the final project in his regression course which would allow students to more extensively showcase what they had learned.

The Ames Housing Dataset has $n = 2,930$ individual home sales in Ames, Iowa from 2006 to 2010. The dataset has 80 covariates and our categorical variable "neighborhood" has $|\mathcal{G}| = 25$.

King County House Sales The King County House Sales Dataset from harlfoxem [2016] is a data set containing the record of 21,613 home sales in King County, Washington between May 2014 to May 2015. The author does not provide much additional information aside from it being a good dataset to test regression models. The dataset is relatively popular with over 169,000 views and 28,000 downloads at the time of this paper.

The data itself came with 21 covariates including the sale price of the house. We treat the "zipcode" covariate, which has $|\mathcal{G}| = 70$, as the categorical variable.

4.4 Empirical Results

We can see that, for both XGBoost and Regression Forests, on average there is an improvement over one hot encoding. The only setting which is more consistently negative is the King County dataset. For the XGBoost simulations, we see that there is a bigger performance improvement for Ames which is a relatively higher dimensional problem compared to the "King County" and "Pakistan" datasets. For the Regression Forest based evaluation, we see that primarily the methods which provide the most dimensionality reduction are the best performers. This could be likely due to the fact that one hot encoding provides a 25-dimensional increase in the modeling problem while methods like Means and MNL add

Figure 6: The percent improvement relative to one hot encoding for each new method on observational data.

p additional dimensions where p is decently larger than the number of unique categories.In both setups, the rotation of the covariate space in King County appears to be unfavorable given that the low rank and sparse low rank methods are the worst performers.

5 Conclusion

In this paper, we explore the task of mapping high cardinality categorical variables G_i to a k-dimensional space without loss of information relevant to our response Y_i. To do this, we make an assumption about the relationship between G_i and Y_i which we call the *sufficient latent state assumption*. This assumption provides us with the basis for creating encoding methods which can be used by universally consistent estimators to extract sufficient representations of G_i. Among our recommendations for encoding methods, we provide encoding methods which are interpretable or focus more on reducing the size of the $\mathbb{R}^k$ representation. We find that these methods tend to outperform one hot encoding and other traditional approaches to modeling with categorical variables as $|\mathcal{G}|$ increases. Ultimately, we see this work as a foundation to more easily working with categorical variables as modern applications seek out signal in all forms of data.

For future lines of research, we believe the next important task will be to figure out how to deal with more than one high cardinality categorical variable for a given prediction problem. Otherwise, additional areas of interest include more directly trying to estimate 4 or using more sophisticated approaches for capturing the structure in 1. Approaches include non-negative matrix factorization or objective functions potentially similar to 9 with constraints to generate probability matrices. Finally, we also aim to further explore more methods in natural language processing and their interpretations in settings with high cardinality categorical variables.

6 Appendix

6.1 Proof of Lemma 1

Proof. To show the equivalence of 1 and 5, we begin by expanding 1 as

$$\mu(x,\, g) = \sum_{l=1}^{L} \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\ G_i = g,\ L_i = l\right] \mathbb{P}\left[L_i = l \,\middle|\, X_i = x,\, G_i = g\right] \tag{23}$$

Now, the conditional independence assumptions encoded in our graph imply that the expectation term simplifies to

$$\mathbb{E}\left[Y_i \,\middle|\, X_i = x,\ G_i = g,\ L_i = l\right] = \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\ L_i = l\right] \tag{24}$$

while the second term can be rewritten using Bayes rule as

$$\mathbb{P}\left[L_i = l \,\middle|\, X_i = x,\, G_i = g\right] = \frac{\mathbb{P}\left[X_i = x \,\middle|\, L_i = l\right] \mathbb{P}\left[L = l \,\middle|\, G_i = g\right]}{\sum_{l'=1}^{L} \mathbb{P}\left[X_i = x \,\middle|\, L_i = l'\right] \mathbb{P}\left[L = l' | G_i = g\right]} \tag{25}$$

Combining the above, we see that the mapping μ only depends on the categorical variable through the function $\psi^*(g) = \mathbb{P}\left[L = l' | G_i = g\right]$. Therefore, $\psi^*(g)$ it is a sufficient representation as defined in (3). □

6.2 Proof of Lemma 2

Proof. Begin by noting that conditioanl expectations can be computed as a linear combination of the sufficient statistics discussed in Lemma 1.

$$\mathbb{E}\left[X_i \,\middle|\, G_i = g\right] = \sum_{l=1}^{K} \mathbb{E}\left[X_i \,\middle|\, L_i = l\right] \mathbb{P}\left[L_i = l \,\middle|\, G_i = g\right] \tag{26}$$

$$= \sum_{l=1}^{K} \mathbb{E}\left[X_i \,\middle|\, L_i = l\right] \psi_l(g) \tag{27}$$

or, in matrix form,

$$\Omega = A\Psi \tag{28}$$

where these matrices are defined as

$$\Omega = \begin{bmatrix} \mathbb{E}\left[X_1 \,\middle|\, G = g_1\right] & \cdots & \mathbb{E}\left[X_1 \,\middle|\, G = g_M\right] \\ \vdots & \ddots & \\ \mathbb{E}\left[X_p \,\middle|\, G = g_1\right] & \cdots & \mathbb{E}\left[X_p \,\middle|\, G = g_M\right] \end{bmatrix}_{p \times M} \tag{29}$$

$$A = \begin{bmatrix} \mathbb{E}\left[X_1 \,\middle|\, L = l_1\right] & \cdots & \mathbb{E}\left[X_1 \,\middle|\, L = l_K\right] \\ \vdots & \ddots & \\ \mathbb{E}\left[X_p \,\middle|\, L = l_1\right] & \cdots & \mathbb{E}\left[X_p \,\middle|\, L = l_K\right] \end{bmatrix}_{p \times K} \tag{30}$$

$$\Psi = \begin{bmatrix} \mathbb{P}\left[L = l_1 \,\middle|\, G = g_1\right] & \cdots & \mathbb{P}\left[L = l_1 \,\middle|\, G = g_M\right] \\ \vdots & \ddots & \\ \mathbb{P}\left[L = l_K \,\middle|\, G = g_1\right] & \cdots & \mathbb{P}\left[L = l_K \,\middle|\, G = g_M\right] \end{bmatrix}_{K \times M} \tag{31}$$

The sufficient representation for the category $\psi(g) = \Psi_g$ lies on the linear span of the set of columns of A. Since A has a left-inverse $A^\dagger$ such that $A^\dagger A = I$, we can retrieve the representations by matrix multiplication.

$$\psi(g) = (\Psi)_{\cdot,g} = A^\dagger(\Omega)_{\cdot,g} =: A^\dagger\omega(g) \tag{32}$$

Since $\omega(g)$ is also a sufficient representation for the categories. □

6.3 Proof of Lemma 3

The coefficients $\hat{\theta}$ satisfies the multinomial logit maximum likelihood problem

$$\hat{\theta} = \arg\max_\theta \frac{1}{n}\sum_i \log \Lambda_\theta(G_i|X_i) \tag{33}$$

The first-order conditions for each coefficient, for each component of the sum (33) is

$$\partial_{\theta_{tj}} \log \Lambda_\theta(G_i|X_i) = \begin{cases} X_{it} - \Lambda_\theta(j|X_i)X_{it} & \text{If } G_i = j, \text{ i.e if } i \text{ belongs to the } j \text{ category} \\ 0 - \Lambda_\theta(j|X_i)X_{it} & \text{otherwise} \end{cases} \tag{34}$$

More succintly, the coefficients must satisfy the following $p \times M$ system of equations

$$\frac{1}{n}\sum_i X_i D_i^T = \frac{1}{n}\sum_i X_i \Lambda_\theta(\cdot|X_i)^T \tag{35}$$

where D_i is the $M \times 1$ one-hot encoding of observation i's observable category (i.e., an $M \times 1$ vector that has $M - 1$ zeros and a 1 in the entry corresponding to G_i). The expression $\Lambda_\theta(\cdot|X_i)$ represents the $M \times 1$ vector of multinomial logistic probabilities for each category.

Asymptotically, the $(t, j)^{th}$ entry of the object on the left-hand side of (35) converges in probability to

$$E[X_{it}1\{G_i = j\}] = E[X_{it}|G_i = j]P(G_i = j) \tag{36}$$

Meanwhile, the corresponding entry on the right-hand side converges to

$$E[X_{it}\Lambda_\theta(G_i = j|X_i)] \tag{37}$$

In matrix form, these can be written as

$$\Omega\Gamma = E[X_i\Lambda_\theta(\cdot|X_i)] \tag{38}$$

where Ω that was defined in (29), and Γ is an $M \times M$ diagonal entries such that $\Gamma_{gg} = P(G_i = g)$. Now, the decomposition (28) allows us to rewrite (38) as

$$\Psi = A^\dagger E[X_i\Lambda_\theta(\cdot|X_i)^T]S^{-1} \tag{39}$$

provided that A has full column rank.

Therefore, following Lemma 1 again, the g^{th} column of the matrix on the right-hand side of (39) is a sufficient representation for category g.

$$\Psi = A^\dagger E[X_i\Lambda_\theta(\cdot|X_i)^T] \propto A^\dagger E[X_i \exp(X_i\theta_g)] \tag{40}$$

6.4 Additional Encoding Methods

For a more in-depth treatment, see Venables [2016]. Note that several of the methods below are simple linear transformations of each other and should yield equivalent levels of performance in theory. However, as we will see in sections 4.2 and 4.3, in practice the resulting performance can differ substantially.

One-hot or dummy This is the most common categorical encoding, and it is the method we take to be our main baseline, against which we will compare all other methods. It expands out the categorical column into $k-1$ columns where k is the number of unique elements in the set of categorical levels in the column. Each column is binary 1 or 0 depending on whether the corresponding level was observed in the original categorical column. [Murphy, 2012, sec 2.3.2]

	b	c	d	e
a	0	0	0	0
b	1	0	0	0
c	0	1	0	0
d	0	0	1	0
e	0	0	0	1

Deviation Similar to one-hot encoding except that the k^{th} unique element's row that is the reference level is now set to all values of -1. This means that categorical levels are being compared to the grand mean of all of the levels instead of the mean of a given level with respect to the reference level.

	b	c	d	e
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1
e	-1	-1	-1	-1

Difference Compares a given level to the mean of the levels that precede it.

	b	c	d	e
a	-0.5	-0.333	-0.25	-0.2
b	0.5	-0.333	-0.25	-0.2
c	0.0	0.667	-0.25	-0.2
d	0.0	0.000	0.75	-0.2
e	0.0	0.000	0.00	0.8

Helmert Compares levels of a chosen categorical variable to the mean of the subsequent levels uniquely observed thus far.

	b	c	d	e
a	0.80	0.00	0.00	0.00
b	-0.20	0.75	0.00	0.00
c	-0.20	-0.25	0.67	0.00
d	-0.20	-0.25	-0.33	0.50
e	-0.20	-0.25	-0.33	-0.50

Repeated Effect Columns are encoded to represent a cumulative comparison of subsequent levels with previous ones.

	b	c	d	e
a	0.8	0.6	0.4	0.2
b	-0.2	0.6	0.4	0.2
c	-0.2	-0.4	0.4	0.2
d	-0.2	-0.4	-0.6	0.2
e	-0.2	-0.4	-0.6	-0.8

Permutation Assigns a unique integer to each category. Note that even when the categories do not possess an intrinsic ordering, some mappings may yield better results if they happen to be aligned with the true average effect the category has on the outcome variable.

	perm
a	5
b	3
c	4
d	1
e	2

Multi-Permutation (Multi-Perm) Following the intuition above, with a larger number of columns we might find more interesting permutations. Hence, we also experiment with four random integer mappings at once.

	perm1	perm2	perm3	perm4
a	1	5	4	2
b	2	3	5	3
c	3	1	2	4
d	4	4	1	1
e	5	2	3	5

Fisher taken from Hastie et al. [2009], we order the categories by increasing mean of the response.

For the following five methods, we use information about the continuous covariates to construct the mapping ψ.

6.5 Estimation details

Below we provide additional details to better clarify how new methods related to Means were estimated and remove basic problems in the data that complicate training models.

For PCA Means, we select the first k principal components which generate 95% of the variance in X, group by the unique categorical levels, and take the means of those principal components $z_{1:k}$. For SPCA Means, we use the default hyperparameters in the "sparsepca" R package.

In the Ames dataset, we remove the features "PoolQC", "GarageQual", and "GarageYrBlt" due to almost being perfectly correlated with other features and remove the missing data rows as well. In the King County House Sales dataset, we remove the ID and date of the house sale as the covariates. Finally, in the Pakistan dataset, many of the covariates are almost perfectly correlated such as "% girls enrolled", "% boys enrolled", and "gender parity score." As a result, we removed such covariates resulting in the following list of covariates: Education score, Toilet, Province, Population, School infrastructure score, Total number of schools, Primary Schools with single teacher, Primary Schools with single classroom, Pakistan Economic Growth, Number of secondary schools, Electricity, No Facility, City, Global Terrorism Index - Pakistan, Complete Primary Schools, Building condition satisfactory, Drone attacks in Pakistan, Drinking water, Boundary wall, Bomb Blasts Occurred, % Complete Primary Schools, % Boys Enrolled.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, pages 1–18, 2018.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

R Dennis Cook et al. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1–26, 2007.

Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

Peter J Diggle, Patrick J Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

András Faragó and Gábor Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

harlfoxem. House sales in king county,usa, 2016. URL https://www.kaggle.com/harlfoxem/housesalesprediction/.

David Harrison, Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. New York: Springer, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

Mesum Raza Hemani. Pakistan education performance dataset, 2017. URL https://www.kaggle.com/mesumraza/pakistan-education-performance-dataset/.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

Kevin Murphy. Machine learning, a probabilistic perspective, 2012.

Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. Heterogeneous treatment effects in digital experimentation. *arXiv preprint arXiv:1412.8563*, 2014.

WN Venables. codingmatrices: Alternative factor coding matrices for linear model formulae [software], 2016.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, pages 265–286, 2006.