Estimating Heterogeneous Treatment Effects in Randomized Control Trials

Christopher P. Adams[*]
Federal Trade Commission
Email: cadams@ftc.gov

May 16, 2014

Abstract

This paper shows that heterogeneous treatment effects can be estimated using data usually available in RCTs under mild assumptions. The paper adapts a methodology used in computer science and signal engineering called non-negative matrix factorization. The paper presents necessary and sufficient conditions for the factorization to be unique. The factorization is implemented as a continuously updating general method of moments estimator and analyzed using Monte Carlo simulations. This estimator is used to measure heterogeneous treatment effects of adjuvant chemotherapies for colon cancer using data from a trial run in the late 1980s. The results suggest that the value of adjuvant chemotherapy varies substantially across the population. For approximately 80% of patients, the addition of adjuvant chemotherapy has little effect on survival, while for 20% of patients, the proportion who survive past 4 years increase thirty five percentage points. The second group tends to have more lymph node involvement and less common histological differentiation of the tumor. However, these treatment effects are not estimated with precision.

[*] I'm very grateful to a number of colleagues that have talked to me about this problem including Matthew Chesnes, Tom Koch, John Rust, Henry Schneider, Nikos Sidiropoulos, Nathan Wilson and participants at the FTC and Georgetown seminars. The views expressed in this article are those of the author and do not necessarily reflect those of the Federal Trade Commission. All remaining errors are my own.

1 Introduction

In his seminal work, Rubin claims that our interest as scientists and policy makers is measuring the difference between the agent's outcome on the proposed treatment and his or her outcome on the alternative treatment (Rubin, 1974). Rubin defines this difference as the "causal effect." Here it will be called the individual treatment effect. If we knew the treatment effect for each individual we would be able to tailor the policy or treatment to that individual. Our policies would be significantly more efficient and effective. Even if we only knew the distribution of the individual treatment effect, it would be of great benefit. We would have a better idea of whether the treatment helped a significant proportion of the population. If treatment effects are correlated with patient observables such as age, gender or results of genetic tests, then we can provide more targeted and efficient treatments. If treatment outcomes are negatively correlated across patients, then we know that finding correlations between patient observables and treatment outcomes is of great value.

The problem is that we do not get to observe the individual treatment effect. We cannot directly measure the distribution of the individual treatment effect because for each individual we only observe the outcome for the treatment that they were exposed to. We can never observe an individual's outcome for the treatment that they did not receive. We can observe the factual outcome, we cannot observe the outcome that is counter to the fact, the counter-factual outcome.

Rubin's solution to this basic identification problem is to suggest that we ask a different question. Rubin argues that we should be interested in the "typical" treatment effect, which he defines as the "average treatment effect." He then points out that if we have data that is unconfounded, such as from ideal randomized control trial, the average treatment effect is identified. The average treatment effect is the average difference between the individual's outcome on the proposed treatment and the individual's outcome on the alternative treatment. Because averages are linear operators, this difference is equal to the difference between the average outcome on the proposed treatment and the average outcome on the alternative treatment. The average difference is equal to the difference in the averages. Given that the data is unconfounded, we have unbiased estimates of the average outcome for each treatment and thus an unbiased estimate of the difference in the averages and thus the average difference.

Rubin's solution is the statistical equivalent of looking for car keys under the street lamp. There may be good reasons for estimating the average treatment effect, but the fact that it is "typical" is not one of them. There is no requirement that a majority of population, a plurality of the population or even a significant number of the population have an individual treatment effect that is the same *sign* as the average treatment effect.[1] Manski (2013) provides a formal argument for estimating the average treatment effect. Manski shows that if the social planner has particular (theoretically standard) preferences then she will want to maximize the average treatment effect.

Adams (2013) presents an alternative solution. While it may be not be possible to directly observe the distribution of treatment effects it may be possible to bound the distribution. Tian and Pearl (2000) show that Manski's idea of bounding the average treatment effect can be generalized to the joint distribution over individual treatment outcomes. The author's call these the "natural bounds" as the only maintained assumption is that probabilities lie between 0 and 1, which they do naturally (Manski, 1990). In the case where we have access to data that is unconfounded, such as data from an ideal randomized control trial, then the Fréchet-Hoeffding bounds are sharp for the joint distribution of individual treatment outcomes (**?**). Adams (2013) discusses various assumptions for sharpening the bounds further, such as the use of subset analysis (Fan and Park, 2009), instrumental variables (Manski and Pepper, 2000) and behavioral assumptions (Heckman and Honore, 1990).

This paper suggests a third possibility. It suggest that while the individual treatment effect cannot be observed directly, it may be inferred. It may be possible to infer the distribution from observed characteristics of the indi-

[1] Thanks to Bill Vogt for pointing this out to me.

vidual and the treatment. Variation in treatment outcomes can be thought of as being “caused” be some unobserved or “latent” characteristic of the individual. Thus identifying the distribution of the individual treatment effect is equivalent to identifying the distribution of the latent characteristic and identifying the functional relationship between the latent characteristic and the distribution of the outcome of interest. The observed distribution of treatment outcomes is a mixture of distributions of outcomes conditional on the latent characteristic.

Mixture models have been successfully applied to various problems in applied statistics (An et al., 2010; Hall et al., 2005; Allman et al., 2009; Cunha et al., 2010). The most common approach to non-parametric identification of mixture models is characterized by Kruskal (1977). In this approach there is assumed to be three conditionally independent signals of the latent variable. The author refers to the problem as a three-way array or tri-linear map. Each signal is an observed characteristic that is determined by the latent variable. If the latent variable is held fixed, then the three observed characteristics are assumed to be statistically independent. In the economics literature these signals are thought of as variables that are measured with error. For example in Cunha et al. (2010) the authors are interested in measuring latent characteristics of children that determine the effect of education resources on their cognitive and non-cognitive skill formation. The latent characteristics of interest are measured with “error” by various standardized tests given to the children. Kruskal (1977) presents sufficient conditions on the observed data (rank conditions) and the data generating process (conditional independence or exclusion restriction assumptions). While the original proof is difficult to penetrate, a number of authors have shown constructively that two observed conditional joint distributions can be decomposed via an eigen-decomposition into a distribution over the latent characteristics and the distributions of the outcome of interest conditional on each of the latent characteristics.[2]

This paper’s approach to non-parametric identification of a mixture model is distinct from the approach discussed above. Most importantly, the assump-

[2] See the proof in An et al. (2010) for example.

tion on the data generating process is substantially more general. Rather than requiring three conditionally independent signals, the approach requires only two conditionally independent signals. Rather than decomposing a three-way array (tri-linear map) the problem involves decomposing a two-way array (a matrix) or a bi-linear map. Henry et al. (2014) describes the assumption as an exclusion restriction. That is, the distribution of the outcome of interest conditional on the latent characteristic is assumed to be constant across all observations of some characteristic of the individual.

The benefit of the approach is that the assumption on the data generating process is much less restrictive than that used in previous work (An et al., 2010; Cunha et al., 2010). The cost of the approach is that requirements on the observed data are much more exacting. Not only do particular rank conditions need to hold, but the approach also requires "sparsity" conditions.[3] The issue is that matrix factorizations are generally *not* unique. Huang et al. (2013) survey the literature and present both necessary and sufficient conditions for a factorization called a "non-negative matrix" factorization to be unique. As suggested by the name, non-negative matrix factorization requires that all the elements of the matrix being factored and the matrix factors be non-negative (Lee and Seung, 1999). Non-negativity is a natural assumption for the application of this approach to mixture models because the matrices represent probability distributions.

Henry et al. (2014) show that if in addition to non-negativity we also (naturally) require certain rows and columns of the matrix to add to one, then it is possible to characterize the set of parameter values (conditional probability distributions) that can be generated as factors of the observed joint distribution. The authors show that the exclusion restriction on the data generating process leads to set identification of the parameters of interest. Henry et al. (2013, 2014) consider the case where there are two latent characteristics and present sufficient conditions on the observed data for point

[3] Sparsity refers to a feature of matrix factor where various cells are zero. It will be more intuitive to think of these conditions as requirements on conditional likelihood ratios. Certain observations will need to signal that the individual is a particular latent type with a very high probability.

identification. This paper presents necessary and sufficient conditions on the observed data for the case where there are K latent characteristics. If there are some values of the observed outcome of interest and some values of the observed characteristic of the individual that uniquely signal the individual's latent characteristic for each characteristic then the factorization is unique. Henry et al. (2013) present Monte Carlo analysis which shows how these likelihood ratios affect the sharpness on the bounds of the probability distributions of interest.

The estimation method is continuously updating general method of moments (Greene, 2000). The approach is very similar to the methods that have been suggested for non-negative matrix factorization (Lee and Seung, 1999) and suggested in Henry et al. (2013). The estimator is analyzed using Monte Carlo simulations of data similar to what may be available in RCTs. The Monte Carlo results point to issues with the optimization procedure interacting with "essential uniqueness." The identification result shows that the factorization is unique up to a rearranging (or relabeling) of the columns. This is a problem for the optimization procedure because the factors are arbitrarily defined. The procedure used in this paper assumes that "type 1" is the factor with the highest probability and "type 2" is the factor with the lowest probability. If these two probabilities are close to each other, i.e. the distribution is 50-50, then the optimization procedure will bounce between the two definitions and may not converge. If the outcomes of the Monte Carlo procedure are limited to those results which converged to a "small" sum of squares, then the results are unbiased. However, if all results are included, the estimate on the probability of the "type 1" is biased downwards.

The GMM estimator is used to estimate the heterogeneous treatment effect of adjuvant chemotherapy for colon cancer patients.[4] The empirical results suggest that chemotherapy given to non-metastatic cancer patients after surgery can help increase survival for all patients. For the majority of patients, approximately 80%, the addition of 5-Fu adjuvant chemotherapy to follow surgery, increases 4-year survival from 73% to 75%. However, for

[4] The study was conducted in the 1980s and considered the value of chemotherapy after surgery for stage III colon cancer patients.

some 20% of patients chemotherapy has a much larger effect on survival. For this group, 4-year survival increases from 0% to 51%.

The original paper finds that the treatment effect on survival varies with a number of patient characteristics including gender, tumor histology and the number of lymph nodes affected (Moertel et al., 1990). The estimator used here assumes there are *at most* two latent types of patients. The majority of patients are the first type. For these Type 1 patients, the effect of chemotherapy is minimal. This type of patient is more likely to have fewer than 4 lymph nodes affected and have their tumor be moderately well differentiated. For Type 2 patients, the effect of chemotherapy on survival is quite large. This type is more likely to have more than 4 lymph nodes affected and have their tumor be either well differentiated or poorly differentiated.

The factorization approach is potentially more general than subset analysis. The factorization approach allows there to be hidden types and allows the treatment effect to vary across the hidden types. This approach does not rule out the possibility that there is some observable characteristic of the patient that perfectly identifies the types. In general, various subsets may not perfectly identify the hidden types, but differences in treatment effects are due to variation in the mix of types in each subset.

The paper proceeds as follows. Section 2 presents the identification result. Section 3 presents the GMM estimator and the Monte Carlo analysis. Section 4 presents results of analysis of an RCT for adjuvant therapy for non-metastatic colon cancer. Section 5 concludes.

2 Identification

2.1 Conditionally Independent Signals

Consider a case where we observe two correlated variables (Y and S). It is assumed that these two variables are independently distributed conditional upon some unmeasured characteristic (U). Figure 1 represents the model. In our colon cancer example, Y represents patient survival after entry into the trial and S represents chronological age. Note that we are assuming that chronological age has no direct effect on survival, rather chronological age is a signal of (unmeasured) physical age which does directly affect survival. Another way to think about this assumption is to say that the patient's physical age is what determines their survival, but we only measure their physical age with error via observing the time from their birth to today. If we knew the patient's physical age, then knowing their birth year would provide no additional information for predicting survival.

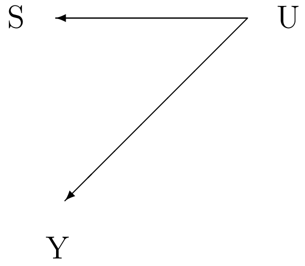

Figure 1: Two-Signal Graph

We can represent the graph with the following equation.

$$\Pr(Y < y, S < s) = \sum_{u \in \mathcal{U}} \pi(u) F_u(y) G_u(s) \tag{1}$$

where $\pi(u) = \Pr(U = u)$, $F_u(y) = \Pr(Y < y | U = u)$ and $G_u(s) = \Pr(S < s | U = u)$. The observed joint distribution is a mixture distribution over the distribution of Y and the distribution of S, which are both independent conditional on the unobserved characteristic U. Note that the graph and the equation incorporate this important assumption.

Assumption 1. $Y_i \perp S_i \mid U_i$

Henry et al. (2014) call this an exclusion restriction, where changes in S are not associated with changes in Y conditional on U.

If the possible outcomes in Y and S are finite we can write Equation (1) in matrix form. Let Y have I elements, S have J elements and U have K elements then.

$$\mathbf{P} = \mathbf{F}\mathbf{D}_{\pi}\mathbf{G}^{T} \tag{2}$$

where $\mathbf{P}$ is a $I \times J$, $\mathbf{F}$ is a $I \times K$ matrix, $\mathbf{D}_{\pi}$ is a diagonal $K \times K$ matrix and $\mathbf{G}^{T}$ is a $K \times J$ transposed matrix.. The matrix $\mathbf{P}$ represents the observed joint distribution over the two observed signals Y and S. This joint distribution is discretized into various cells representing the joint probability. For example the top left cell may be the probability that the patient dies in the first year and is less than 50 years old. The matrix $\mathbf{F}$ represents the probabilities of Y, with each column providing the probability of survival conditional on the latent characteristic. The diagonal matrix $\mathbf{D}_{\pi}$ represent the probabilities over the latent characteristics. Lastly, the matrix $\mathbf{G}$ is the conditional probability distribution over S, where each column is conditional on a different latent characteristic.

$$\begin{array}{l} \mathbf{P}_{ij} = [\Pr(y = Y_i, s = S_j)] \\ \mathbf{F}_{ik} = [\Pr(y = Y_i | u = U_k)] \\ \mathbf{D}_{kk} = [\pi_k] \text{ and } \mathbf{D}_{lm} = [0] \\ \mathbf{G}_{jk} = [\Pr(s = S_j | u = U_k)] \end{array} \tag{3}$$

Given this rewrite of our mixture distribution we can present the linear algebra result. It is always possible to factorize a rectangular matrix into two matrices.

$$\mathbf{P} = \mathbf{W}\mathbf{H} \tag{4}$$

where $\mathbf{W}$ is a $I \times K$ matrix and $\mathbf{H}$ is a $K \times J$ matrix. We can think of $\mathbf{W} = \mathbf{F}\mathbf{D}_{\pi}$ and $\mathbf{H} = \mathbf{G}^{T}$. There are numerous versions of this type of factorization and they are used for numerous purposes (Lee and Seung, 1999). The issue with this factorization is that it is "unique" up to any $K \times K$ matrix $\mathbf{A}$ of full-rank.

$$\mathbf{P} = \mathbf{W}\mathbf{A}\mathbf{A}^{-1}\mathbf{H} \tag{5}$$

or

$$\mathbf{P} = \tilde{\mathbf{W}}\tilde{\mathbf{H}} \tag{6}$$

where $\tilde{\mathbf{W}} = \mathbf{WA}$ and $\tilde{\mathbf{H}} = \mathbf{A}^{-1}\mathbf{H}$. That is to say, the factorization it is not particularly unique.

However, it has been shown that under certain conditions on the matrices $\mathbf{W}$ and $\mathbf{H}$, the factorization can be unique. One set of such conditions is non-negativity. That is we can factorize $\mathbf{P}$ constraining $\mathbf{W}$ and $\mathbf{H}$ to have all positive elements.

The results below will consider uniquely determining these matrices up to "permutations". That is re-ordering the columns of $\mathbf{W}$. Let $\mathcal{Q}$ be the set of permutation square matrices of full-rank.

Definition 1. *Let $\mathcal{Q}$ be such that for all $\mathbf{Q} \in \mathcal{Q}$*

1. *$\mathbf{Q}$ is $K \times K$,*
2. *$\mathbf{Q}$ is full rank, and*
3. *$\mathbf{Q}_{ij} = 1$ and $\mathbf{Q}_{i'j'} = 0$ for all $i' \neq i$, $j' \neq j$ and all $i, j \in \{1, ..., K\}$.*

Formally, the decomposition is "essentially unique" if it is unique up to re-ordering and re-scaling.

Definition 2. *(Uniqueness of NMF) (Huang et al., 2013) The NMF of $\mathbf{P} = \mathbf{WH}$ is (essentially) unique if $\mathbf{P} = \tilde{\mathbf{W}}\tilde{\mathbf{H}}$ implies $\tilde{\mathbf{W}} = \mathbf{WQD}$ and $\tilde{\mathbf{H}} = (\mathbf{QD})^{-1}\mathbf{H}$, where $\mathbf{D}$ is a diagonal matrix with its diagonal entries positive and $\mathbf{Q} \in \mathcal{Q}$ is a permutation matrix (see Definition 1).*

The factorization is unique if the $\mathbf{A}$ matrix is a diagonal matrix multiplied by a permutation matrix.[5]

The following is a necessary condition for uniqueness (see Theorem 3 in Huang et al. (2013)).

Theorem 1. *(Necessary Condition) Define*

$$\begin{aligned} \mathcal{I}_k &= \{i \in \{1, ..., I\} | \mathbf{W}_{i,k} \neq 0\} \\ \mathcal{J}_k &= \{j \in \{1, ..., J\} | \mathbf{H}_{k,j} \neq 0\} \end{aligned} \tag{7}$$

[5] Definition 2 is actually for the asymmetric case. Huang et al. (2013) present a number of results for the symmetric case.

If the NMF $\mathbf{P} = \mathbf{WH}$ *is unique, then there do not exist* $k_1, k_2 \in \{1, ..., K\}$*,* $k_1 \neq k_2$ *such that* $\mathcal{I}_{k_1} \subseteq \mathcal{I}_{k_2}$ *or* $\mathcal{J}_{k_1} \subseteq \mathcal{J}_{k_2}$*.*

Proof. See Appendix B of Huang et al. (2013). □

Theorem 1 is a necessary condition, meaning that the distributions conditional on type must satisfy this condition in order for the factorization to be unique. It puts a lower bound on what is required. The minimum number of zero or sparse cells that satisfy the condition of the theorem would be K with one in each column of $\mathbf{W}$ and each of K rows of $\mathbf{W}$, and the same for $\mathbf{H}$.

Huang et al. (2013) also present the sufficient conditions that have been presented in the literature. In general, these conditions imply that for each row of $\mathbf{H}$ there can *only* be one positive element. This means that for each type, the conditional distribution must put probability one on an observable characteristic of the patient. These are very onerous conditions that may not hold in practice. Below it is shown that the minimum number of zero cells that would be necessary for essential uniqueness are also sufficient when the rows of $\mathbf{H}$ add to 1 (as they do if the matrix represents conditional probability distributions).

The next lemma shows that adding up constrains $\mathbf{A}$ to be a matrix where the rows sum to 1.

Lemma 1. *If* $\mathbf{A}$*,* $\mathbf{H}$ *and* $\tilde{\mathbf{H}}$ *are matrices with the following properties*

1. $\tilde{\mathbf{H}} = \mathbf{A}^{-1}\mathbf{H}$
2. $\sum_{j=1}^{J} \mathbf{H}_{jk} = 1$ *for all* $k \in \{1, ..., K\}$*, and*
3. $\sum_{j=1}^{J} \tilde{\mathbf{H}}_{jk} = 1$ *for all* $k \in \{1, ..., K\}$

then

$$\sum_{k=1}^{K} \mathbf{A}_{kl} = 1 \text{ for all } l \in \{1, ..., K\} \tag{8}$$

Proof. By (1)

$$\mathbf{A}\tilde{\mathbf{H}}\mathbf{1} = \mathbf{H}\mathbf{1} \tag{9}$$

From (2) and (3)

$$\mathbf{A}\mathbf{1} = \mathbf{1} \tag{10}$$

□

The adding up constraint on $\mathbf{H}$ reduces the possible matrices that $\mathbf{A}$ can be. In particular, the rows of $\mathbf{A}$ must sum to 1. Also note that the same is true for $\mathbf{A}^{-1}$.

The following result provides sufficient conditions for uniqueness of the decomposition. The result shows that $\mathbf{A}$ can be close to the $\mathbf{I}$ or a particular permutation matrix when certain cells of $\mathbf{W}$ and $\mathbf{H}$ are close to zero. Note that the notation $\mathbf{B} \geq 0$ means that every element of $\mathbf{B}$ is positive.

Theorem 2. *If* $\mathbf{A}$, $\mathbf{W}$, $\tilde{\mathbf{W}}$, $\mathbf{H}$ *and* $\tilde{\mathbf{H}}$ *are matrices with the following properties,*

1. $\tilde{\mathbf{H}} = \mathbf{A}^{-1}\mathbf{H}$,
2. $\tilde{\mathbf{W}} = \mathbf{W}\mathbf{A}$,
3. $\sum_{j=1}^{J} \mathbf{H}_{kj} = 1$ *for all* $k \in \{1, ..., K\}$,
4. $\sum_{j=1}^{J} \tilde{\mathbf{H}}_{kj} = 1$ *for all* $k \in \{1, ..., K\}$,
5. $\mathbf{H} > 0$, $\mathbf{W} > 0$, *and*
6. $\tilde{\mathbf{H}} \geq 0$, $\tilde{\mathbf{W}} \geq 0$

then as $\mathbf{W}_{i_k k} \to 0$ *and* $\mathbf{H}_{k j_k} \to 0$ *for at least one* $i_k \in \{1, ..., I\}$ *and* $j_k \in \{1, ..., J\}$, *where* $i_k \neq i_{k'}$ *and* $j_k \neq j_{k'}$ *for all* $k \neq k' \in \{1, ..., K\}$, $\mathbf{A} \to \mathbf{Q}$ *where* $\mathbf{Q} \in \mathcal{Q}$.

Proof. Sketch of proof for the case where $K = 2$. The proof of the $K > 2$ case is presented in the appendix.

Step 1. By (1), (3) and (4) and Lemma 1 let

$$\mathbf{A} = \begin{bmatrix} 1+a & -a \\ -b & 1+b \end{bmatrix} \tag{11}$$

where $a, b \in \Re$. Note that

$$\mathbf{A}^{-1} = \frac{1}{det(\mathbf{A})} \begin{bmatrix} 1+b & a \\ b & 1+a \end{bmatrix} \tag{12}$$

where $det(\mathbf{A}) = 1 + a + b$.

Step 2. Let $\{a, b\} \in \mathcal{A}$ where each element are the parameters of $\mathbf{A}$ that satisfy the conditions of the theorem.

Case 1. Let $1 + a + b > 0$. In this case, (1), (2), (5) and (6) lead to four inequalities that will hold for all $\{a, b\} \in \mathcal{A}$ for all vectors associated with $\mathbf{W}$ and $\mathbf{H}$

1. $b \leq (1 + a)\frac{\mathbf{W}_1}{\mathbf{W}_2}$
2. $b \geq a\frac{\mathbf{W}_1}{\mathbf{W}_2} - 1$
3. $b \geq -a\frac{\mathbf{H}_2}{\mathbf{H}_1} - 1$
4. $b \geq -(1 + a)\frac{\mathbf{H}_2}{\mathbf{H}_1}$

where $\mathbf{W}_k$ refers to the kth column of $\mathbf{W}$ and $\mathbf{H}_k$ refers to the kth row of $\mathbf{H}$.

Figure 2 presents the set $\mathcal{A}$ as the interior of the four inequalities. We see that if there is an element of $\mathbf{W}_2$ and an element of $\mathbf{W}_1$ such that the ratio of the elements goes to zero, then the inequality described by (2) rotates inwards to the b-axis. Similarly if there is two *different* elements of the same vectors such that the ratio of $\mathbf{W}_1$ and $\mathbf{W}_2$ goes to zero, the inequality described by (1) rotates inwards to the a-axis. Similarly, for the other vectors so that $\{a, b\} \rightarrow \{0, 0\}$ for all $\{a, b\} \in \mathcal{A}$.

Case 2. Rearrange the columns of $\mathbf{A}$ and make the same argument as for Case (1). □

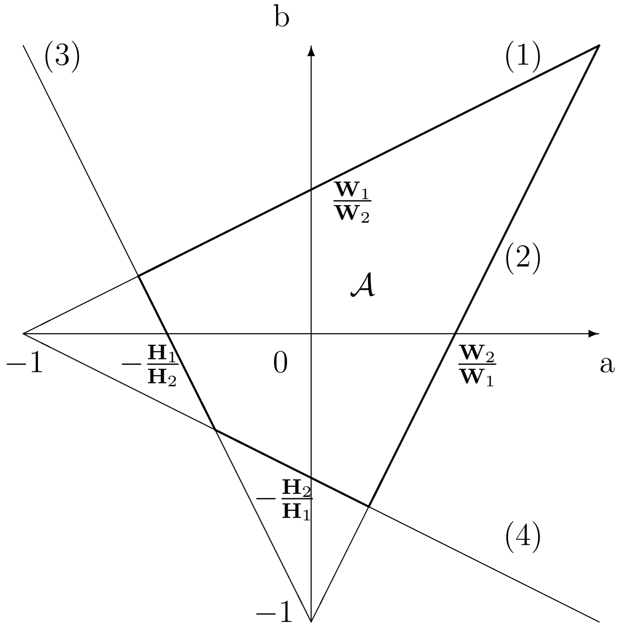
b
(3)
(1)
$\frac{\mathbf{W}_1}{\mathbf{W}_2}$
(2)
$\mathcal{A}$
-1
$-\frac{\mathbf{H}_1}{\mathbf{H}_2}$
0
$\frac{\mathbf{W}_2}{\mathbf{W}_1}$
a
$-\frac{\mathbf{H}_2}{\mathbf{H}_1}$
(4)
-1

Figure 2: The set $\mathcal{A}$

Theorem 2 presents sufficient conditions for uniqueness of the factorization when the matrix $\mathbf{H}$ is also constrained to have its rows add to 1, which is a natural condition if such rows are probability distributions. The sufficient conditions are much less onerous than the ones that have been presented in the literature previously and similar to the necessary condition that has been previously presented. The theorem also shows that if some of the cells are "close" to 0, then the factorization can be "close" to unique in the sense that all the matrices that satisfy the factorization are close to each other in the standard Euclidean metric.

This is an alternative approach to the proof of the result presented in Henry et al. (2014). Unlike Henry et al. (2014) the result also shows how observable variables need to change in order for the factorization to be unique (for point identification).[6]

[6] Henry et al. (2013) presents a similar result showing how conditions on the "tails" of

2.2 Identification of the Unconfounded Model

This section shows that the results presented above can be applied to the problem of identification of heterogeneous treatment effects in data from randomized control trials.

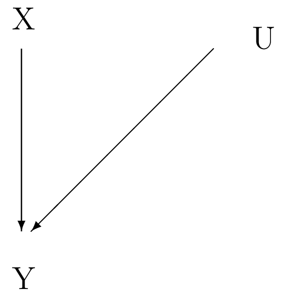

Figure 3: Unconfounded Graph

Figure 3 represents our stylized randomized control trial. Our treatment (X) is causing our outcome of interest (Y) and the "treatment effect" of X on Y is being mediated by some unobserved characteristic (U). With data from RCTs we are able to observed the probability distribution of Y conditional on the treatment X. If our outcome of interest is ordered such as survival then we can write the following equation.

$$\Pr(Y < y|X = x') = \sum_{u \in \mathcal{U}} \pi(u) F_u(y|x') \tag{13}$$

where $x' \in \mathcal{X}$ represents a particular treatment. Equation (13) shows that the probability we observe is the average of the survival distributions conditional on the treatment chosen and the unobserved characteristic. Our observed survival is "averaged out" over the unobserved characteristic (U).

the outcome distributions conditional on type need to change for the parameters to be point identified. Henry et al. (2014) also presents the identified set for the case where there are more than two unobserved types.

If we compare two treatments we can write the following equation where the observed difference in treatments is equal to the average of the conditional differences.

$$\begin{aligned}&\Pr(Y < y|X = x') - \Pr(Y < y|X = x'')\\ &= \textstyle\sum_{u \in \mathcal{U}} \pi(u)(F_u(y|x') - F_u(y|x''))\end{aligned} \tag{14}$$

This equation shows that if we observe that the LHS is positive, it does not mean that all the component parts of the RHS are positive.

If we were considering two different drug treatments in the adjuvant colon cancer setting and we found that the LHS is positive, we may conclude that the treatment x'' causes patients to live longer than the alternate treatment x'. However, it may be that for many patients the conditional probability of survival $(1 - F_u(y))$ is lower for x''. It is perfectly possible for the "average" effect of the treatment to be positive and the individual effect of the treatment to be negative for a large number of patients.

The solution to this identification problem is to use the result presented in the previous section.

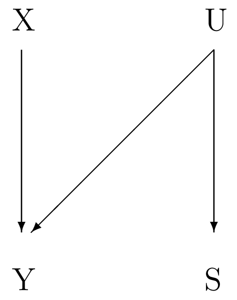

Figure 4: RCT-Signal Graph

Figure 4 represents a situation where we have a signal of the unobserved characteristic (U) that is mediating the treatment effect. In our colon cancer example, this signal may be chronological age. With this signal available we are able to observed the joint distribution of Y and S conditional on the treatment.

$$\Pr(Y < y, S < s | X = x') = \sum_{u \in \mathcal{U}} \pi(u) F_u(y|x') G_u(s) \tag{15}$$

Note that it is assumed that the signal S doesn't vary with the treatment (X). This is not strictly necessary but it makes the equation consistent with the graph.

It is straight forward to see that the sub-graph $U - Y - S$ of Figure 4 is exactly the same as the graph in Figure 1. So holding the treatment X constant, we can use the argument above to show that $\pi(u)$, F_u and G_u are identified from observing the joint distribution of the outcome of interest and the signal.

In the next section we use Equation (15) to create the moments for the GMM estimator.

Equation (15) suggests a natural test of the model. If the equation holds for any value of X then we can estimate the model separately for each treatment and test whether the distribution of S and U vary between the treatments. We can jointly test the assumptions of the model including the cardinality of U, the conditional independence of Y and S and the exclusion restriction that the treatment does not affect the distribution of U or S. Note that below we do not do this test, rather we make the additional exclusion restriction that the unconditional distribution of S does not vary with the treatment choice (X). We require random assignment into treatment, no selection into the trial itself and full-compliance with the treatment assignment for this assumption to be credible. Moertel et al. (1990) indicates that there seem to be no issues with treatment compliance, but no information is provided about selection into the trial itself.

3 Constrained NMF Estimator

This section presents a general methods of moments implementation of the constrained non-negative matrix factorization. The optimization problem is

$$\begin{aligned} \min_{\mathbf{F},\mathbf{D}_\pi,\mathbf{G}} \quad & ||\mathbf{P} - \mathbf{F}\mathbf{D}_\pi\mathbf{G}^T||_F^2 \\ \\ s.t. \quad & \mathbf{F}_{ik} \geq 0 \text{ for all } i,k \\ & \mathbf{D}_{\pi kk} \geq 0 \text{ for all } k \\ & \mathbf{G}_{jk} \geq 0 \text{ for all } j,k \\ & \textstyle\sum_{i=1}^{I} \mathbf{F}_{ik} = 1 \text{ for all } k \\ & \textstyle\sum_{j=1}^{J} \mathbf{G}_{jk} = 1 \text{ for all } k \\ & \textstyle\sum_{k=1}^{K} \pi_k = 1 \end{aligned} \tag{16}$$

The object being minimized is simply the elements of the resulting matrix squared and summed, i.e. sum of squares. We can think of the cells of the matrix $\mathbf{P}$ as particular "moments" of the joint distribution. It is thus equivalent to a GMM estimator and this paper implements the problem as a continuously updating GMM estimator with the parameters forced to satisfy the constraints.

For the Monte Carlo analysis consider the case where there are two types of patients, where 75% are Type 1 ($\pi_1 = 0.75$). There is an observed outcome (Y) and an observed characteristic of the patient (S). The researcher observes the unconditional probability distribution of outcomes ($P(Y,S)$) for a sample of size N. In this case the matrix $\mathbf{P}$ is 3×3 which is then factored into matrices $\mathbf{F}$ and $\mathbf{G}$ that are 3×2.

$$\mathbf{F} = \begin{bmatrix} 0.29 & 0.95 \\ 0 & 0.05 \\ 0.71 & 0 \end{bmatrix} \tag{17}$$

$$\mathbf{G} = \begin{bmatrix} 0 & 0.2 \\ 0.19 & 0.8 \\ 0.81 & 0 \end{bmatrix} \tag{18}$$

The previous section shows that for the factorization to be unique, each column of the two matrices must have cell that is 0 and that zero cell cannot be in the same row in the same matrix. Here, the cell is not actually zero, but is set to a probability of 0.0001. Conditional on the patient's unobserved type, there is a 1/10,000 probability of that outcome being observed.

Table 1: NMF Monte Carlo Results

	Actual	Mean	SD	Mean	SD
π_1	0.7500	0.7337	(0.0862)	0.7493	(0.0032)
π_2	0.2500	0.2663	(0.0862)	0.2507	(0.0032)
$\mathbf{F}_{11}$	0.2900	0.3522	(0.1648)	0.2891	(0.0034)
$\mathbf{F}_{21}$	0.0001	0.0607	(0.1487)	0.0001	(0.0001)
$\mathbf{F}_{31}$	0.7099	0.5871	(0.2316)	0.7109	(0.0034)
$\mathbf{F}_{12}$	0.9499	0.8184	(0.2581)	0.9488	(0.0026)
$\mathbf{F}_{22}$	0.0500	0.1229	(0.1847)	0.0506	(0.0025)
$\mathbf{F}_{32}$	0.0001	0.0558	(0.1043)	0.0006	(0.0007)
$\mathbf{G}_{11}$	0.0001	0.0689	(0.1752)	0.0001	(0.0001)
$\mathbf{G}_{21}$	0.1899	0.2588	(0.1806)	0.1903	(0.0030)
$\mathbf{G}_{31}$	0.8100	0.6723	(0.2566)	0.8096	(0.0030)
$\mathbf{G}_{12}$	0.1999	0.2108	(0.1094)	0.1998	(0.0044)
$\mathbf{G}_{22}$	0.8000	0.6841	(0.2201)	0.7980	(0.0051)
$\mathbf{G}_{32}$	0.0001	0.0892	(0.1722)	0.0023	(0.0027)
SOS	0.0000	0.2507	(0.4348)	0.00003	(0.00003)
T		100		56	

Table 1 presents the results from the Monte Carlo analysis assuming 30,000 observations with 100 runs. The first column is the actual values from the model. The second two columns presents summary statistics for all 100 runs. The first column in the set is the average value for each estimated parameter value and the second column is the standard deviation over the parameter values. The remaining columns are for the case where the cases are limited to those with "low" sum of squares.

The table shows that the when all cases are presented the results are biased, however when only the results from the cases with low sum of squares are presented, the bias goes away. The bias occurs because there are natural corner solutions to the problem. As mentioned above, the estimator is unique only up to permutations, i.e. relabeling the latent types. If the estimator assigns even probability weights to each of the types it will get stuck and will not converge to the global optimum.

The next section takes the GMM implementation of the non-negative matrix factorization to an actual data set from a randomized control trial.

4 Heterogeneous Effect of 5-Fu Chemotherapy

Today the standard of care for stage III colon cancer is surgery to remove the tumor and affected lymph nodes followed by an "adjuvant" chemotherapy consisting of 5-Fu and oxalplatin.[7] However, in the early 1990s it was not clear what the benefit was of giving chemotherapy to patients who had already had the cancer removed. A study reported in the New England Journal of Medicine showed a big difference between the survival outcomes who received "observation" after surgery and the patients that received a 5-Fu chemotherapy (Moertel et al., 1990). According to the study authors, "Among the patients with Stage C [III] disease, therapy with levamisole plus fluorouracil reduced the risk of cancer recurrence by 41 percent (P < 0.0001). The overall death rate was reduced by 33 percent (P = 0.006). ... We conclude that adjuvant therapy with levamisole and fluorouracil should be standard treatment for Stage C [III] colon carcinoma."

Although the analysis presented is an "average" for the population and provides little information about how effective the treatment is for various patients, the authors confidently conclude that all patients should receive ad-

[7] See `http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-treating-by-stage-colon`. The term "adjuvant" refers to the idea of giving chemotherapy after curative surgery prior to observing an cancer recurrence.

juvant chemotherapy. The authors conclusions have generally been adopted by the medical profession, albeit using different protocol than the ones studied here. The authors are aware of the possibility that the effectiveness may vary across the population. To that end, the authors conduct what they term "exploratory subset analysis." The authors suggest that tumor characteristics and lymph node involvement are associated with a larger treatment effect.

In the data 56% of patients in the observation group survive 4 years, while 68% of patients in 5-Fu group survived 4 years. There is a tendency to interpret the results as saying an *extra* 12 percentage point of patients would have survived if they had been given 5-Fu instead of surgery alone. Often, there is no recognition of the possibility that the results are perfectly consistent with mixing over two groups of patients. One group may have even greater survival benefit from 5-Fu while another group has their survival probability reduced by 5-Fu.

Here we consider the same data analyzed by Moertel et al. (1990).[8] However, it is assumed that there exist at most two types of colon cancer patients and these patients may have different treatment effects. Tumor histology and the number of affected lymph nodes are assumed to be a signal of the underlying patient type. Note that the analysis does not force their be variation in treatment effects.

The results presented below are estimated assuming that conditional distributions of observable patient characteristics and the distribution of types are the same across the three treatments. Both assumptions seem reasonable since the patients were randomly assigned to treatments and there is no evidence of large deviations from the treatment assignment or biased attrition. Unfortunately, the small sample sizes do not allow the distributions to be estimated separately for each treatment. Thus it is not possible to run the test suggested by the previous section.

Table 2 presents the results from the non-negative matrix factorization assuming that there are no more than two hidden types and assuming that

[8] This data is available in R.

Table 2: Heterogeneous Treatment Effect of Colon Cancer Drugs [5th percentile, 95th percentile]

	Obs	Lev	Lev & 5-Fu
Type 1	0.77 [0.66,0.89]		
Less 1 Year	0.04 [0.00,0.08]	0.04 [0.00,0.07]	0.05 [0.00,0.36]
1-3 Years	0.14 [0.00,0.22]	0.20 [0.07,0.27]	0.13 [0.00,0.69]
3-4 Years	0.09 [0.00,0.14]	0.05 [0.00,0.09]	0.07 [0.00,1.00]
More 4 Years	0.73 [0.58,0.89]	0.71 [0.62,0.89]	0.75 [0.00,0.89]
Type 2	0.23 [0.11,0.34]		
Less 1 Year	0.20 [0.00,0.29]	0.28 [0.11,0.51]	0.17 [0.00,0.98]
1-3 Years	0.71 [0.52,1.00]	0.47 [0.29,0.79]	0.32 [0.00,1.00]
3-4 Years	0.09 [0.00,0.18]	0.11 [0.00,0.22]	0.00 [0.00,0.67]
More 4 Years	0.00 [0.00,0.12]	0.15 [0.00,0.14]	0.51 [0.00,1.00]
Sum of Squares	0.04512 [0.05,0.38]		

the joint distribution of tumor histology and affected lymph nodes is independent of survival conditional on the patient's type. The first set of three columns is for Type 1 patients. The analysis estimates that 77% of patients in the trial are Type 1. The first four rows presents the proportion of Type 1 patients who survive less than 1 year, between 1 year and 3 years, between 3 years and 4 years and over 4 years respectively. For Type 1 patients receiving "observation", the probability of surviving more than 4 years is 73%. This survival rate decreases slightly to 71% for patients receiving levamisole and then slightly increases again to 75% for patients receiving both levamisole and 5-Fu. The last bottom set of rows presents the same results for Type 2 patients. 23% of the patients in the trial are Type 2 patients. None of these patients survive 4 years when they receive "observation." This proportion increases to 15% for Type 2 patients receiving levamisole, and increases dramatically to 51% for Type 2 patients who receive the combination of levamisole and 5-Fu. The table presents the 5th and 95th percentiles for the bootstrapped standard errors with 100 runs. The results suggest some very large variation in the measured effectiveness of the 5-Fu trial arm.

Table 3: Observed Patient Characteristics By Type for Nodal Involvement and Histological Differentiation [5th percentile, 95th percentile]

	Type 1	Type 2
< 4, Mod. Well	0.65	0.00
	[0.56,0.67]	[0.00,0.30]
< 4, Well & Poor	0.16	0.54
	[0.13,0.23]	[0.38,0.67]
> 4, Mod. Well	0.18	0.11
	[0.14,0.21]	[0.00,0.19]
> 4, Well & Poor	0.02	0.35
	[0.00,0.06]	[0.11,0.45]

Table 3 presents the remainder of the results from the factorization. It presents the proportion of patients that are Type 1 and have certain observable characteristics based on the number of lymph nodes affected by the cancer and the histology of the tumor. 65% of Type 1 patients are had less than 4 lymph nodes affected and their tumor was moderately well differentiated. The table shows that for 87% of Type 2 patients their tumor was either well differentiated or poorly differentiated. The 5th and 95th percentile of the bootstrapped standard errors are in square brackets underneath the parameter estimates.

The results presented in Table 2 show that the treatment effect associated with 5-Fu varies dramatically across the population analyzed by the randomized control trial. For the majority of patients, the addition of 5-Fu to the chemotherapy regime has almost no effect on survival. However, for a minority of patients, the addition of 5-Fu has a dramatic effect. 4 year survival increases about thirty five percentage points. The results presented in the table also show that the patients affected by the use of 5-Fu tend to have tumors that are either well differentiated or poorly differentiated and are more likely to have more than 4 lymph nodes affected by the cancer. These results are consistent with the fact that for stage II patients, who generally have no lymph node involvement, there is little evidence that 5-Fu increases overall survival.[9]

Figure 5 presents the Kaplan-Meier plot for those patients who had less than 4 lymph nodes affected and whose tumor is moderately well differentiated. From Tables 3 we see that 100% of these patients are Type 1. The figure illustrates that there is very little difference in outcomes across the three trial arms for this group of patients.

Figure 6 presents the Kaplan-Meier survival plot for patients who had more than 4 lymph nodes affected and whose tumor was either poorly differentiated or well differentiated. From Tables 2 and 3 we can see that most of these patients are Type 2 and we also see that for this group there is a large difference in survival outcomes across trial arms.

The previous section discussed the necessary and sufficient conditions for results to be unique. There must be a zero cell (or small cell) in at least one of the first four rows for each type and that cell cannot be in the corresponding column. In addition there must be a zero cell in the second four rows for each type of patient and it cannot be in the same row. We see that for Type

[9] See NCI discussion on the evidence for adjuvant therapy for stage II colon cancer patients at `http://www.cancer.gov/cancertopics/pdq/treatment/colon/HealthProfessional/page7#Reference7.11` (last accessed 1/26/14)

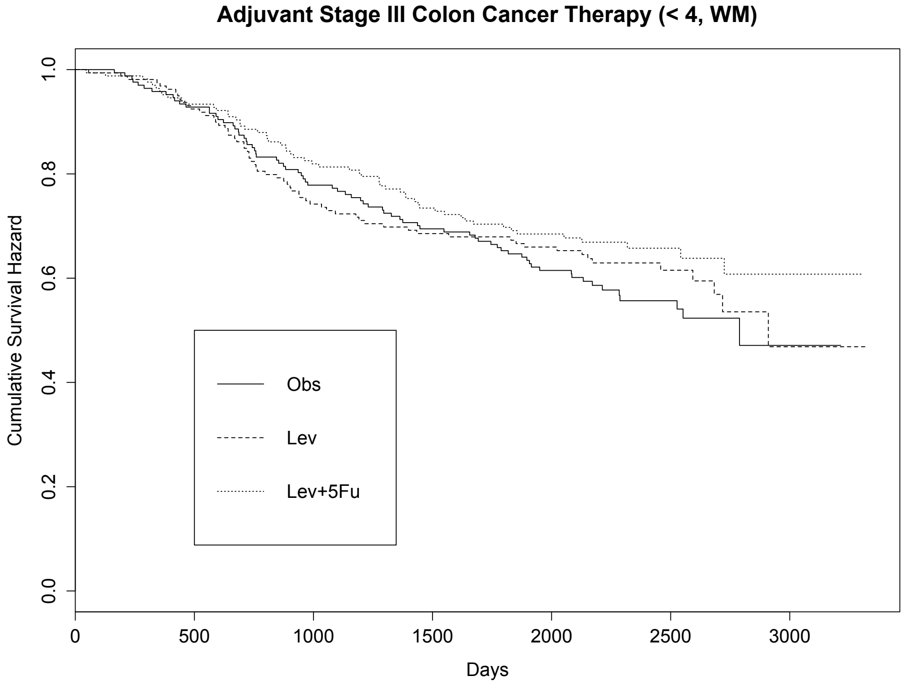

Figure 5: Kaplan-Meier plot for patients with less than 4 nodes implicated and a moderately well differentiated tumor. All these patients are Type 1.

2, there are three cells rounded to zero, two in the first four rows and one in the last four rows. In the case of Type 1 there are no cells rounded to zero. These results suggest we should have some concern about variation in the estimated results due to the non-uniqueness of the factorization.[10]

5 Conclusion

The paper explores an idea originally developed in computer science and signal processing called non-negative matrix factorization for estimating het-

[10] Note that variation presented from the bootstrapped results would incorporate some of the variation due to the non-uniqueness of the factorization.

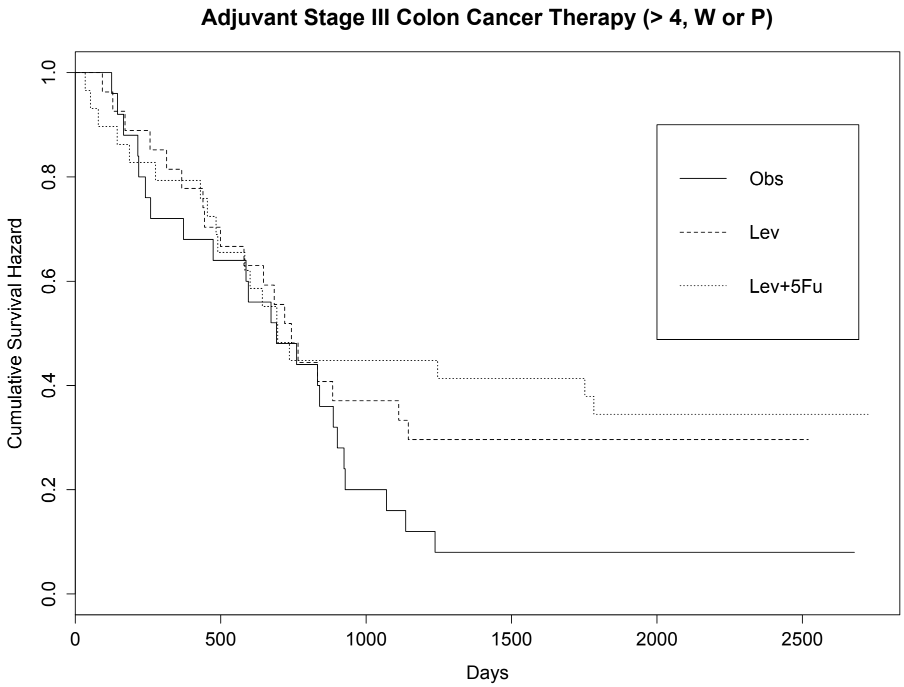

Figure 6: Kaplan-Meier plot for patients with more than 4 nodes implicated and either well or poorly differentiated tumor. 16% of these patients are Type 1 and 84% are Type 2.

erogeneous treatment effects. Non-negative matrix factorization is a method for determining “hidden” factors such as hidden patient types that may be associated with variation in treatment effectiveness. The idea is that there may exist signals of the hidden types and the joint distribution of those signals can reveal the underlying distributions conditional on the hidden types. Unfortunately, in general NMF is not unique. The paper presents sufficient conditions for the factorization to be unique for the case where there are two hidden types and when the factored matrix are additionally constrained to represent probabilities. It shows that as probability of certain observed characteristics conditional on one hidden type (but not the other) goes to zero, then the factorization becomes unique up to relabeling. The method is tested on a data set from a randomized control trial conducted in the 1980s analyzing the effect on survival of adjuvant chemotherapy for stage III colon cancer patients. The analysis shows that the adjuvant therapy has no effect for about 80% of patients while for about 20% of the patients the therapy dramatically increases the likelihood that a patient will survive over 4 years after surgery to remove the tumor. However, the treatment effect is very imprecisely measured.

References

Christopher P. Adams. Inference bounds on the joint distribution of treatment outcomes. Federal Trade Commission, June 2013.

E S Allman, C Matias, and J A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37: 3099–3132, 2009.

Yonghong An, Yingyao Hu, and Matthew Shum. Estimating first-price auctions with an unknown number of bidders: a misclassification approach. *Journal of Econometrics*, 157:328–341, 2010.

Flavio Cunha, James J. Heckman, and Susanne M. Schennach. Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3):883–931, May 2010.

Yanqin Fan and Sang Soo Park. Sharp Bounds on the Distribution of Treatment Effects and their Statistical Infernece. Vanderbilt University, February 2009.

William Greene. *Econometric Analysis*. Prentice Hall, fourth edition, 2000.

Peter Hall, Amnon Neeman, Reza Pakyari, and Ryan Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678, 2005.

James Heckman and B Honore. The Empirical Content of the Roy Model. *Econometrica*, 58:1128–1149, 1990.

Marc Henry, Koen Jochmans, and Bernard Salanié. Inference on mixtures under tail restrictions. Penn State, December 2013.

Marc Henry, Yuichi Kitamura, and Benard Salanié. Partial identification of finite mixtures in econometric models. *Quantitative Economics*, 5:123–144, 2014.

Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 2013. Forthcoming.

J. B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions with application to arithmetic complexity and statistics. *Linear Algebra Applications*, 18:95–138, 1977.

D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Charles Manski. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings*, 80:319–323, 1990.

Charles Manski. *Public Policy in an Uncertain World.* Harvard University Press, 2013.

Charles Manski and John Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68:997–1010, 2000.

C G Moertel, T R Fleming, J S Macdonald, D G Haller, and J A Laurie. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine*, 322(6):352–358, 1990.

Donald B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 1974.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, February 2000.

6 Appendix

Proof of Theorem 2.

Proof. The proof is in two steps. In Step (1), it is shown that if the $\Re^{K(K-1)}$ parameter space associated with $\mathbf{A}$ is limited to the axes, then it converges to the origin. Step (2) shows that if that space converges to the origin, then so does the whole parameter space associated with $\mathbf{A}$ converges to the origin.

Step 0. Let $\mathbf{A}$ be denoted as

$$\mathbf{A} = \begin{bmatrix} 1+\sum_{k=2}^{K} a_{1k} & -a_{12} & \cdots & -a_{1K} \\ -a_{21} & 1+\sum_{k\neq 2} a_{2k} & \cdots & -a_{2K} \\ \vdots & & & \vdots \\ -a_{K1} & \cdots & -a_{K(K-1)} & 1+\sum_{k=1}^{K-1} a_{Kk} \end{bmatrix} \tag{19}$$

Let $a \in \mathcal{A} \subset \Re^{K(K-1)}$ be a vector such that its corresponding parameters are for an $\mathbf{A}$ that satisfies the conditions of Theorem 2, where $a = \{a_{12}, ..., a_{1K}, a_{21}, a_{23}, ..., a_{K(K-1)}\}$.

Let $a^{lm} \in \mathcal{A}^* \subset \mathcal{A}$ be such that $a^{lm} = \{0, ..., 0, a_{lm}, 0, ..., 0\}$. That is, $\mathcal{A}^*$ is the subset of $\mathcal{A}$ where the element sits on a particular axis of the space.

Step 1. Consider $a^{12} \in \mathcal{A}^*$. In this case

$$\mathbf{A} = \begin{bmatrix} 1+a_{12} & -a_{12} & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & 0 & 1 \end{bmatrix} \tag{20}$$

and

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & a_{12} & 0 & \cdots & 0 \\ 0 & 1 + a_{12} & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & & 0 & 1 \end{bmatrix} \tag{21}$$

where $det(\mathbf{A}) = 1 + a_{12}$. By (2) and (6)

$$\mathbf{W}_1(1 + a_{12}) \geq 0 \tag{22}$$

and

$$-\mathbf{W}_1 a_{12} + \mathbf{W}_2 \geq 0 \tag{23}$$

By (1) and (6)

$$\mathbf{H}_1 + \mathbf{H}_2 a_{12} \geq 0 \tag{24}$$

and

$$\mathbf{H}_2(1 + a_{12}) \geq 0 \tag{25}$$

If there exists a $\mathbf{W}_{i1} > 0$ and $\mathbf{H}_{2j} > 0$ then $a_{12} \in [-\frac{\mathbf{H}_{1j'}}{\mathbf{H}_{2j}}, \frac{\mathbf{W}_{i'2}}{\mathbf{W}_{i2}}]$. So $a_{12} \to 0$ as $\{\mathbf{W}_{i'2}, \mathbf{H}_{1j'}\} \to \{0, 0\}$. This can be shown for any $a^{lm} \in \mathcal{A}^*$.

Step 2. *Claim: If for $a^* \in \mathcal{A}^*$, $a^* = \{0, 0, 0, ..., 0\}$, then for all $a \in \mathcal{A}$, $a = \{0, 0, 0, ..., 0\}$.*

Suppose not. Let $a \in \mathcal{A}$ and $a \neq 0$.

Case 1. By (2) and (6) and assumption the following equality holds

$$-\mathbf{W}_1 b - \mathbf{W}_k c + \mathbf{W}_l \geq 0 \tag{26}$$

where $c > 0$.

$$b \leq \frac{\mathbf{W}_l - \mathbf{W}_k c}{\mathbf{W}_1} \tag{27}$$

If for all $\mathbf{W}_l$, $\mathbf{W}_k$ and $\mathbf{W}_1$ given the data, the inequality holds, then it will still hold when $c = 0$ for some $b > 0$. A contradiction. □