ESSAYS ON MEASUREMENT ERROR
IN FINITE-VALUED VARIABLES

NANDITA GOPALKRISHNA GAWADE

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
ECONOMICS
ADVISER: BO E. HONORÉ

SEPTEMBER 2012
This dissertation examines the nonparametric identification of parameters of interest in measurement error models when the mismeasured variable is finite-valued. The maintained assumption is that two or more observed variables are independent conditional on the unobserved variable that is misclassified.

This conditional independence assumption is the generalization to a nonparametric setting of a common exclusion restriction used in the linear errors-in-variables (EIV) model. For example, the classical EIV model in a linear regression context assumes that measurement error is uncorrelated with the true regressor and with the error in the regression equation. The second assumption means that the measurement error is uncorrelated with the outcome once we have controlled for the true regressor. In a nonlinear model with non-additive errors, we would require that the regression measurement error and the outcome are independent conditional on the true regressor.

The conditional independence assumption, although strong, is common in practice. The popularity of this assumption is in part due to the fact that, when there are three or more observed variables, conditional independence leads to nonparametric point identification of the joint distribution of observable variables and the latent variable. In this dissertation, we use the conditional independence assumption to frame the problem as a finite mixture model in which the components are product distributions.

The first chapter is concerned with partial identification in measurement error models when the mismeasured variable is discrete. In particular, we are interested in the joint
distribution of an outcome $Y$ and a regressor $X^*$, but we observe the joint distribution of $Y$ and a surrogate $X$. We explore the implications of assuming nondifferential measurement error or conditional independence of the outcome and the surrogate given the true regressor. The conditional independence assumption results in a factoring of the joint distribution of the observables. The chief observation of this paper is that, when all variables are discrete, this is a problem of matrix factorization. In particular, we obtain a complete description of the identified set when both $X$ and $X^*$ are binary and improve on the bounds existing in the literature. In contrast, the identified set can be very complicated when $X^*$ can take three or more values.

In the second chapter, we examine the identified set for the nonparametric regression of continuous $Y$ on misclassified binary $X^*$. The results obtained in Chapter 1 continue to hold even when $Y$ is continuously distributed. The identified set can be completely characterized by two scalar parameters, namely the infimum and supremum of the ratio of observed densities of $Y$ conditional on the two values of $X$. We also consider a simple strategy for estimating these two parameters, and examine the properties of this strategy using a simulation study.

The third chapter is a note on the point identification obtained when we add a third observed variable to the measurement error model described in Chapter 1. We are now concerned with the problem of decomposing a three-dimensional tensor. We show that the point identification obtained in this case is related to the uniqueness of tensor decompositions, first proved by Kruskal (1977).
In the final chapter, we observe that a $K$-variate finite mixture of product distributions with $R$ components can be represented by the stochastic factorization of an order-$K$ tensor into $R$ rank-1 tensors. When the number of components $R$ is not known, an object of interest is the smallest number of components that is consistent with the observed distribution. We show that this number is equal to the nonnegative rank of the order-$K$ tensor representing the observed distribution. We find lower bounds for the nonnegative tensor rank that are based on the matrix rank of flattened versions of the tensor. In general, these bounds are not sharp. We obtain sufficient conditions under which these bounds are attained, and provide examples of probability arrays where these bounds are strictly lower than the number of components.
ACKNOWLEDGEMENTS

Above all, I am indebted my adviser Bo Honoré for giving his time and attention to my work. The completion of this thesis would not have been possible without his guidance, encouragement and support. I am also grateful to Andrés Aradillas-López, Laura Chioda, and Andriy Norets for reading different versions of this thesis and for helpful suggestions. For comments and discussion, I would like to thank participants in the Econometrics Seminar at Princeton University, and also Kaveh Akram and Robert Meyer, my colleagues at the Value-Added Research Center.

I also want to thank my friends who have encouraged me through graduate school and beyond. Finally, I want to thank my family and especially my parents, Gopalkrishna and Anuradha Gawade, who are eventually responsible for everything. This thesis is dedicated to them.
For my parents.
## Contents

Abstract ................................................................. iii

Acknowledgements ....................................................... vi

List of Figures ........................................................... x

1 Measurement Error in Finite-Valued Explanatory Variables: Implications of Conditional Independence .......................... 1

   1.1 Introduction ....................................................... 1

      1.1.1 Related Literature .......................................... 6

   1.2 Conditional Independence and the Identified Set ................. 8

   1.3 The Identified Set for Binary $X^*$ ................................ 15

      1.3.1 Regression of $Y$ on $X^*$ .................................. 21

      1.3.2 Continuous $Y$ ............................................... 22

   1.4 Geometric Properties of the Identified Set in Higher Dimensions .... 24

   1.5 Conclusion ...................................................... 31

2 Identification and Estimation of Bounds in a Regression with a Binary Misclassified Regressor ........................................ 38
List of Figures

1.1 A geometric interpretation of the conditional independence assumption . . . 10
1.2 The identified set when both $Y$ and $X^*$ are binary. . . . . . . . . 17
1.3 The line passing through $\pi_1$ and $\pi_2$. . . . . . . . . . . . . . . . . . . . . . 19
1.4 $\tilde{A}(\Pi)$ is connected up to re-labelling of the values of $X^*$. . . . . . . . . 26
1.5 Finiteness does not imply identification up to permutations. . . . . . . . . 30
2.1 $Y|X^*$ is Normal with a location shift . . . . . . . . . . . . . . . . . . . . . . 48
2.2 $Y|X^*$ is Normal with a scale change . . . . . . . . . . . . . . . . . . . . . . 50
2.3 Identified set for $\beta$ as a function of its true value . . . . . . . . . . . . . . 55
2.4 Identified set for $\beta$ as a function of $\sigma$ . . . . . . . . . . . . . . . . . . . . . 56
2.5 Identified set for $\beta$ as a function of the marginal distribution of $X^*$ . . 57
2.6 Identified set for $\beta$ as a function of the misclassification probabilities . . . 58
4.1 Rank and nonnegative rank of matrices . . . . . . . . . . . . . . . . . . . . . . . 87
4.2 Face of the 3-simplex (Double-dashed lines are the rank-1 matrices) . . . . 88
4.3 Rank and nonnegative rank of an order-3 tensor can be different . . . . 89
4.4 Tensor rank can be bigger than mode-$k$ ranks . . . . . . . . . . . . . . . . 90
Chapter 1

Measurement Error in Finite-Valued Explanatory Variables: Implications of Conditional Independence

1.1 Introduction

Measurement error in explanatory variables has been widely studied in statistics and econometrics. It is known that, in general, the effects of interest are not identified in the absence of a model describing the measurement error and of additional information such as a validation sample.

The classical errors-in-variable (EIV) model in a linear regression context makes two assumptions on the measurement error:

i. Measurement error is uncorrelated with the true regressor, and

ii. measurement error is uncorrelated with the error in the regression equation.

These assumptions yield well-known bounds on the coefficient from the infeasible regression
of the outcome on the true regressor. First, measurement error leads to an attenuation in the relationship between the outcome and the observed regressor, i.e. the coefficient on the observed regressor is biased towards zero. Second, the inverse of the coefficient from the reverse regression of the observed regressor on the outcome is an upper bound for the true coefficient of interest.

In this chapter, we are interested in obtaining bounds for the relationship between an outcome and a regressor when the latter is finite-valued. Consider, for example, the effect on wages of belonging to a labour union when union membership is observed with error, or the effect of prenatal smoking behavior of the mother on infant birthweight when smoking status is misreported.

The first assumption of the linear EIV model is no longer valid in the case of discrete random variables. Measurement error (or misclassification) is now necessarily correlated with the true value of the regressor. This is most apparent in the case of a binary $0 - 1$ variable where measurement error can take three possible values $-1, 0$ and $1$. For a given true value, measurement error can take only two of its possible values, and so is dependent on the true variable.

The second assumption of the linear EIV model needs to be strengthened for application in a nonparametric setting. The essence of this assumption is that measurement error does not contain any information about the outcome once we have controlled for the true regressor. The observed error-ridden variable is assumed to be conditionally independent of the dependent variable given the true underlying variable. This assumption is know as
‘redundancy’ or ‘nondifferential error’ in the measurement error literature. The goal of this chapter is to examine the identifying power of the conditional independence assumption when the mismeasured variable is discrete.

This paper employs a completely nonparametric approach in that it specifies no underlying model describing the relationship between the outcome $Y$ and the misclassified explanatory variable $X^*$, or error model describing the relationship between $X$ and $X^*$. It asks instead what can be learned about the conditional distribution of $Y$ given $X^*$ when we observe only the conditional distribution of $Y$ given $X$, the observed error-ridden measure of $X^*$, and assume nondifferential measurement error.

We start by assuming that all observed variables are discrete. The main observation used here is that the conditional independence assumption translates to a factoring of the joint distribution of the observed variables. The joint distribution of $N$ discrete variables can be represented by an $N$-dimensional tensor. So conditional independence relations among discrete variables are related to the problem of tensor decomposition.

For instance, we can view the conditional distributions $\Pr(Y|X)$, $\Pr(Y|X^*)$ and $\Pr(X^*|X)$ as probability matrices whose columns add up to one. The conditional independence assumption leads to a factoring of the matrix $\Pr(Y|X)$ into the two matrices $\Pr(Y|X^*)$ and $\Pr(X^*|X)$. The problem of identifying the underlying distributions is thus a problem of matrix factorization. The set of probability matrices that satisfy this condition is quite complicated. In general, it is not finite and $\Pr(Y|X^*)$ is not identified.

Nevertheless, we would like to understand properties of this set. Is the identified set
connected? If it is connected, then we can find a path between any two candidates for the joint distribution of \((Y, X, X^\ast)\). On the other hand if it is disconnected, then we can think of the disjoint components as corresponding to different explanations for the relationship between the outcome and the true regressor. If the disjoint components are ‘far apart’, we can use this information to formulate additional assumptions on the true relationship between \(Y\) and \(X^\ast\) (or on the measurement error process) that could lead to rejection of one or more of the components.

Are the identified set or its components convex? If yes, then the interior of the set consists of distributions for \((Y, X, X^\ast)\) that can be written as mixtures of the distributions on the boundary of the set. If, in addition, the identified set is a convex polytope, then all admissible candidate distributions can be written as mixtures of the distributions at the vertices of the identified set. In this case, it is possible to completely describe the identified set using only a finite number of parameters.

We find a complete description of the identified set for \(\Pr(Y\mid X^\ast)\) when both \(X\) and \(X^\ast\) are binary. We find tight bounds for the conditional distributions \(\Pr(Y\mid X^\ast)\) and \(\Pr(X^\ast\mid X)\), and for all functionals of them. In particular, we bound the regression of \(Y\) on \(X^\ast\) given by the marginal effect \(\theta^\ast \equiv E[Y\mid X^\ast = 1] - E[Y\mid X^\ast = 2]\). These bounds use all the information contained in the observed distribution and improve on bounds existing in the literature, such as Bollinger (1996) and Mahajan (2003).

We also show that, in the binary \(X^\ast\) case, the identified set consists of the union of two disjoint rectangles corresponding to the two possible ‘labelings’ of the values of \(X^\ast\). We can
reduce the set to a single rectangle by assuming that the misclassification is not so bad as
to reverse the values of $X$ and $X^*$, i.e, a success is more likely to be reported as a success
than misreported as a failure. Because the identified set is a rectangle, we can explicitly
characterize it using a finite number of parameters, namely the vertices of the rectangle.
Finally, we extend the partial identification result to the case where $Y$ is continuous.

In contrast to the binary case, the identified set can be very complicated when $X^*$
takes more than two values. It may be connected or consist of a finite number of disjoint
components. Unlike the binary case, the disjoint components need not correspond only to
switching the values of $X^*$. When the disjoint components are ‘far apart’ we can still use
assumptions limiting the extent of measurement error to reject one or more of them.

In the case of binary $X^*$, we can show that $\Pr(Y|X^*)$ is identified to a re-labelling
of $X^*$ if and only if the support of the observed distribution $\Pr(Y|X)$ shifts with $X$. In
this case, the only admissible misclassification is when the values of $X^*$ are recoded to
obtain $X$, and $X$ is perfectly correlated with $X^*$. Surprisingly, this is no longer true when
$X^*$ can take three or more values. Even when the support of $Y$ shifts with $X$ and the
identified set is finite, the admissible distributions for $(Y,X,X^*)$ can correspond to very
different explanations for the measurement error process and for the relationship between
the outcome and the true regressor.

Finally, in both cases, we find that the identifying power of the conditional independence
assumption depends on the strength of the relationship between the observed variables, $Y$
and $X$. When the distribution of $Y$ changes a lot with $X$, the identified set is smaller.
The limiting case is when the support of $Y$ shifts with $X$ and the identified set is finite. This compares nicely with a result for the classical EIV model. As mentioned earlier, the coefficient on the true regressor lies in the interval bounded from below by the coefficient from the regression of $Y$ on $X$ and from above by the inverse of the regression of $X$ on $Y$. The length of this interval depends on the $R^2$ of these regressions, which is a measure of the strength of the linear relationship between $Y$ and $X$. See the discussion in Hausman (2001) for details and for examples of when we might expect a strong relationship between the observed variables.

1.1.1 Related Literature

The problem of linear regression in the presence of a binary misclassified regressor was first studied by Aigner (1973) who showed that the least squares estimate was biased towards zero. Bollinger (1996) showed that the least squares estimate is the tight lower bound for the true slope parameter, and also obtained an upper bound that depends on the slope from the inverse regression. In a nonparametric setting, Mahajan (2003) obtained sharp bounds for the probabilities of interest when both $Y$ and $X^*$ are assumed to be binary. Black, Berger, and Scott (2000) obtained bounds for the slope parameter in a linear regression model with two noisy reports of the regressor.

Black, Berger, and Scott (2000) also show that additional information in the form of repeated measurements can be used to obtain point identification of the parameters of interest. This was shown independently in Kane, Rouse, and Staiger (1999). Mahajan (2006) obtained point identification of the nonparametric regression function and the misclassifi-

All of the aforementioned papers assume nondifferential measurement error. Molinari (2008) describes how to obtain bounds for a mean regression with a misclassified regressor under any set of assumptions, specifically without assuming nondifferential misclassification. Deng and Hu (2009) use assumptions on the nonparametric regression model to bound the effect of a binary misclassified regressor with arbitrary measurement error.

The contribution of the current paper is the observation that the assumption of nondifferential measurement error is related to the problem of matrix factorization or tensor decomposition. We use this observation to obtain sharp bounds on the effect of a binary misclassified regressor. These bounds use all the information in the observed distribution. We show that the identified set for the probabilities of interest can be quite complicated when the misclassified regressor takes three or more values. We describe the identified set in two cases - when the distribution of the outcome does not change much with the observed regressor, and at the other extreme when the support of the outcome shifts with the observed regressor.
1.2 Conditional Independence and the Identified Set

In this section, we define a theoretical framework for the identification problem.

Let $Y$ be an outcome variable taking values in $\{1, \ldots, I\}$ and let $X^*$ be an explanatory variable taking values in $\{1, \ldots, R\}$. Suppose we are interested in $P(Y|X^*)$ but observe $P(Y|X)$, where $X$ is an imperfect measure of $X^*$ taking values in $\{1, \ldots, J\}$. For most of this chapter, we will assume that $J = R$, i.e. the imperfect measure can take the same number of values as the true explanatory variable.

Consider the conditional independence assumption $Y \perp X|X^*$. This means that the observed measure $X$ affects the outcome $Y$ only through the true explanatory variable $X^*$. The goal of this paper is to examine the identifying power of this assumption.

The conditional independence assumption leads to the following relationship between the observed and the unknown probabilities:

$$P(Y = i|X = j) = \sum_r P(Y = i|X^* = r)P(X^* = r|X = j).$$

(1.2.1)

The joint distribution of any two discrete random variables can be written in the form of a probability matrix with positive entries that add up to one. Similarly, the conditional distribution of an $M$-valued random variable given an $N$-valued random variable can be represented by an $M \times N$ probability matrix whose columns are positive and add up to one. We will call such matrices as conditional probability matrices.

Define the conditional probability matrices $\Pi, A$ and $B$ as follows:

$$\pi_{ij} = P(Y = i|X = j); \quad a_{ir} = P(Y = i|X^* = r); \quad b_{rj} = P(X^* = r|X = j)$$
Equation (1.2.1) then translates to the following matrix equation.

$$
\pi_{ij} = \sum_r a_{ir} b_{rj} \text{ or } \Pi = AB
$$

(1.2.2)

Here $\Pi$ is the $I \times R$ matrix of observed conditional probabilities. The unknown conditional probabilities are contained in the $I \times R$ matrix $A$ and the $R \times R$ matrix $B$.

Conditional independence of $Y$ and $X$ given $X^*$ means that the matrix $\Pi$ can be factorized into two component matrices $A$ and $B$. The identification problem is then to find all pairs of conditional probability matrices $(A, B)$ that satisfy equation (1.2.2). In this paper, we treat the entries of $A$ as the parameters of interest and those of $B$ as nuisance parameters.

The space of probability distributions of $Y$ can be represented by the standard $(I - 1)$-dimensional simplex given by

$$
\Delta^{I-1} = \left\{ p \in \mathbb{R}^I : p_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{I} p_i = 1 \right\}.
$$

Every column of $\Pi$ and of $A$ is a probability distribution of $Y$. Therefore every column of $\Pi$ and of $A$ is a point in $\Delta^{I-1}$. Similarly, every column of $B$ is a point in $\Delta^{R-1}$. The set of all possible conditional probability matrices $\Pi$ (and of $A$) is given by $(\Delta^{I-1})^R$. The corresponding set for $B$ is given by $(\Delta^{R-1})^R$.

We can now define the identified set as follows:

$$
\mathcal{A}(\Pi) \equiv \{ A \in (\Delta^{I-1})^R : \text{there exists } B \in (\Delta^{R-1})^R \text{ such that } \Pi = AB \}
$$

(1.2.3)

We are now ready to state our first proposition, which is a geometric description of the identified set defined above. A proof is given in the appendix.
Proposition 1.2.1. A probability matrix $A \in (\Delta^{I-1})^R$ is in the identified set $A(\Pi)$ if and only if every column of $\Pi$ lies in the convex hull of the columns of $A$.

An immediate corollary of Proposition 1.2.1 is that the identified set is closed under permutation of columns of $A$. This can also be deduced from (1.2.2) because we can always write $AB$ as $APP^T B$. Here $P$ is any permutation matrix and hence orthogonal to itself.

A second observation is that $A = \Pi$ is always in the identified set, because we cannot rule out the possibility that there is no measurement error. Again, this fact is immediately obvious from (1.2.2) which always has the solution $A = \Pi$ and $B = I_R$.

The above two observations are better understood with the help of a simple example.

Example 1.2.1. Let $Y$ be a three-valued variable and let $X$ and $X^*$ be four-valued variables. Here $I = 3$ and $R = 4$. The 2-dimensional simplex is a triangle in $\mathbb{R}^3$ as shown in Figure 1.1. Each point in the triangle $\Delta^2$ is a valid probability distribution for $Y$.

The points $\pi_1, \pi_2, \pi_3$ and $\pi_4$ in $\Delta^2$ represent the distribution of $Y$ conditional on
$X = 1, X = 2, X = 3$ and $X = 4$ respectively. These points are contained in the convex hull of the points $a_1, a_2$ and $a_3$. Let $a_4$ be any point in $\Delta^2$. The matrix $A = [a_1, a_2, a_3, a_4]$ is a valid solution for the conditional distribution $P(Y|X^*)$. Here, $a_1$ is the distribution of $Y$ conditional on $X^* = 1$ and so on.

Any matrix formed by a permutation of the columns of $A$ is also in the identified set. For instance, the matrix $\tilde{A} = [a_2, a_3, a_4, a_1]$ is a valid solution for the conditional distribution $P(Y|X^*)$, with $a_2$ as the distribution of $Y$ conditional on $X^* = 1$. Thus, permutations of this matrix give us $4!$ possible values for the conditional distribution $P(Y|X^*)$.

Further, note that the matrix $\Pi = [\pi_1, \pi_2, \pi_3, \pi_4]$ and all permutations of it are valid solutions for the conditional distribution of $Y$ given $X^*$. The corresponding misclassification matrices would be $4 \times 4$ permutation matrices, signifying that there is no measurement error other than a relabeling of values between $X$ and $X^*$.

In this example, and indeed in general, there are infinite such matrices whose columns contain the observed probabilities in their convex hull.

Let us first discuss in some detail one of the sources of this non-identification, as it were. Given a matrix $A$ that is in the identified set, the matrix obtained by permuting the columns of $A$ is also in the identified set. This form of multiplicity of solutions is known as ‘aliasing’. It is due to the fact that the different values of $X^*$ are merely labels and, in the absence of further assumptions, we cannot meaningfully distinguish between them.

Hu (2008) provides alternative assumptions on the underlying model or on the misclassification probabilities that ‘fix’ a particular order of the values of $X^*$. Assumptions 2.4 and
2.5 of that paper take the form of monotonicity restrictions on the relationship between $Y$ and $X^*$ i.e., the underlying model. On the other hand, assumptions 2.6 and 2.7 are restrictions on the relationship between $X$ and $X^*$ that lead to determination of the columns of the misclassification matrix.

Monotonicity of the underlying model may be suggested by economic theory. For instance, we may expect the probability of low birth weight to increase with the frequency of maternal smoking. Given a solution for the conditional distribution of $Y$ given $X^*$ that satisfies this restriction, no permutation of this solution can satisfy the same monotonicity condition. Moreover, this restriction will also rule out all permutations of conditional distributions with maximum probability of low birthweight at, say, a moderate level of maternal smoking. In other words, monotonicity of the underlying relationship between $Y$ and $X^*$ can have more content than to merely fix the aliasing problem. The same is true of any restrictions on the misclassification probabilities that are suggested by a priori knowledge of the measurement error process.

The problem considered in Hu (2008), in Mahajan (2006), and also in Chapter 3 of this thesis is such that the conditional independence assumption gives us identification up to permutations of the unknown distributions. In this case, the assumptions considered in Hu (2008) in order to fix the aliasing do exactly that and no more. In the present case, however, these assumptions imply additional restrictions on the identified set, and so we will refrain from invoking them.

Let us now discuss the more general implications of Proposition 1.2.1. There are two
kinds of restrictions on the space in which $A$ can lie. The first comes from the fact that $A$ is a conditional probability matrix and the second from the fact that $A$ must satisfy the conditional independence assumption.

Proposition 1.2.1 suggests an interesting geometric representation of these restrictions. Think of $\Pi$ as $R$ points in the $(I - 1)$-simplex. These points give us a convex polytope with at most $R$ vertices. The identified set $\mathcal{A}(\Pi)$ is then the set of all convex polytopes with $R$ or fewer vertices that ‘contain’ the given polytope and are ‘contained’ in the $(I - 1)$-simplex.

The rank of the observed conditional probability matrix $\Pi$ is a measure of the dependence between $Y$ and $X$. Suppose that $\Pi$ has the minimum possible rank of one. This means that the columns of $\Pi$ are equal. The distribution of $Y$ given $X$ does not change with $X$ and, in fact, the two variables are independent. In the above example, this will occur when the points $\pi_r$ coincide. In this case, the identified set $\mathcal{A}(\Pi)$ is the largest that it can be under the assumption of conditional independence alone but it is still strictly smaller than $(\Delta^{I-1})^R$.

One case where the dimension of the problem is significantly reduced is when $\text{rank}(B) = R < I$. We are now limited to those cases where the dependent variable takes more values than the regressor. For instance, the binary choice model with a many-valued discrete misclassified regressor is excluded. Once we assume that $Y$ takes many values compared to $X^*$, the assumption that $B$ is full rank is only restrictive when the observed matrix $\Pi$ has less than full rank.
Let $\tilde{A}(\Pi)$ denote the identified set when $B$ has full rank.

$$\tilde{A}(\Pi) \equiv \{A \in (\Delta^{I-1})^R : \text{there exists } B \in (\Delta^{R-1})^R \text{ with } \text{rank}(B) = R \text{ such that } \Pi = AB\}$$

(1.2.4)

The assumption that $B$ is invertible is a statement about the distribution of $X^*$ given $X$. When $X^*$ and $X$ are binary, this assumption is equivalent to assuming that $X^*$ and $X$ are not independent. The rank of $B$ tells us how informative the observed measure $X$ is about the true regressor $X^*$. When $B$ has full rank, every value that $X$ takes is informative about the distribution of $X^*$ and $X$ is like a particularly good instrument.

**Proposition 1.2.2.** Let $\text{rank}(B) = R$. Then $A \in \Delta^{I-1}$ is in the identified set $A(\Pi)$ if and only if

i. every column of $A$ lies in the affine hull of the columns of $\Pi$, and,

ii. every column of $\Pi$ lies in the convex hull of the columns of $A$.

One implication of Proposition 1.2.2 is that if $Y \perp X$ then we must have $Y \perp X^*$. Of course this is obvious because $B$ having full rank implies that $X$ and $X^*$ are not independent.

Proposition 1.2.2 tells us that the columns of $A$ lie in the intersection of a $(\text{rank}(\Pi) - 1)$-dimensional subspace and the $(I - 1)$-dimensional simplex. This means that our ‘search space’ for the columns of $A$ is limited to a $(\text{rank}(\Pi) - 1)$-dimensional convex polytope instead of an $(I - 1)$-dimensional one. As we shall see, this fact is especially useful when $R = 2$ and the columns of $A$ are constrained to lie on a line segment.

The above discussion suggests that the set $A(\Pi)$ is very hard to describe. We are inter-
ested in convexity, connectedness and dimension of this set. We would like, if possible, to obtain a parametric representation of this set. In the following section, we obtain complete answers to all of these questions in the case where \( X^* \) is binary. In section 1.4, we will demonstrate that matters become quite complicated in higher dimensions. We describe the identified set in limiting cases corresponding to the strength (or weakness) of the relationship between the observed variables. However, a complete characterization of the identified set in higher dimensions is beyond the scope of this paper.

1.3 The Identified Set for Binary \( X^* \)

In this section, we characterise the identified region for \( P(Y|X^*) \) when \( X^* \) takes only two values. We begin with the simplest possible example in which both the outcome \( Y \) and the regressor \( X^* \) are binary. We then describe the identified set when \( Y \) can take more than two values and also in the limiting case of continuous \( Y \).

Example 1.3.1. In this example, we consider the simplest possible case in which all variables are binary.

Suppose that \( Y, X \) and \( X^* \) all take values in \( \{1, 2\} \). The table below gives the observed conditional distribution of \( Y \) given \( X \) and the implied bounds for \( P(Y|X^*) \) under the assumption that \( P(Y = 1|X^* = 1) > P(Y = 1|X^* = 2) \).

| \( Y \) | \( P(Y|X = 1) \) | \( P(Y|X = 2) \) | \( P(Y|X^* = 1) \) | \( P(Y|X^* = 2) \) |
|---|---|---|---|---|
| \( Y = 1 \) | 0.7 | 0.6 | [0.7,1] | [0,0.6] |
| \( Y = 2 \) | 0.3 | 0.4 | [0,0.3] | [0.4,1] |

The observed probabilities \( P(Y|X) \) are a convex combination of the unknown probabilities \( P(Y|X^*) \). This, together with the assumption to restrict the size of the measurement
error, means that $P(Y = 1|X^* = 1) \geq 0.7$ and $P(Y = 1|X^* = 1) \leq 0.6$. We get the outer bounds from the fact that all parameters are probabilities.

We now describe this problem in geometric terms. The 1-simplex is the line segment joining the points (0,1) and (1,0). It suffices to bound the probabilities $P(Y = 1|X = 1)$ and $P(Y = 1|X = 2)$ given by the first row of $A$. The identified set is therefore a subset of the unit square. It is simply the union of the two rectangles given by

$$[0, \min(\pi_{11}, \pi_{12})] \times [\max(\pi_{11}, \pi_{12}), 1] \cup [\max(\pi_{11}, \pi_{12}), 1] \times [0, \min(\pi_{11}, \pi_{12})].$$

Note that these bounds are exactly the ones obtained in Mahajan (2003). There are several points to be noted.

i. Suppose $Y$ and $X$ are independent so that $\pi_{11} = \pi_{12} = \tilde{\pi}$. Then the two rectangles touch at the point $(\tilde{\pi}, \tilde{\pi})$ and the identified set is connected. In all other cases, i.e. when $Y$ and $X$ are not independent, the two rectangles are disjoint and the identified set is disconnected.

ii. Even when connectedness does not hold, the two components of the identified set are convex.

iii. Finally, the two components correspond to permutations of the columns of $A$ or to a switching of the values of $X$.

These nice properties of the identified set are not limited to the above example but in fact extend to all cases in which the columns of $\Pi$ as well as $A$ are constrained to lie on a line. Suppose $X^*$ is binary and $X$ and $X^*$ are not independent. Consider the observed
conditional probability matrix \( \Pi \). If \( \pi_1 = \pi_2 \), i.e. \( X \perp Y \), then by Proposition 1.2.2 we must have \( X^\ast \perp Y \), and \( P(Y|X^\ast) = P(Y|X) \). Of course, in this case, there are no restrictions on the misclassification probabilities \( B = P(X^\ast|X) \) except the ones imposed by the dependence of \( X \) and \( X^\ast \).

Consider the more interesting case in which \( \pi_1 \neq \pi_2 \). The affine hull of two distinct points in \( \mathbb{R}^l \) is the line passing through them.

\[
p(t) = t\pi_1 + (1-t)\pi_2 \quad t \in \mathbb{R}.
\]  

By part (i) of Proposition 1.2.2 the columns of \( A \) must lie on this line. However they cannot be any two points on this line because \( A \) is a probability matrix which satisfies the conditional independence assumption. Let \( a_r = p(t_r) \) for \( r = 1, 2 \). Note that

\[
A = \Pi \begin{bmatrix}
t_1 & t_2 \\
1 - t_1 & 1 - t_2
\end{bmatrix} = \Pi B^{-1}
\]

\[
a_r = t_r \pi_1 + (1 - t_r) \pi_2 \quad r = 1, 2.
\]
$t_1$ and $t_2$ are elements of the inverse of the misclassification probability matrix. We will show that the restrictions imposed on $A$ imply that $(t_1, t_2)$ lies in the union of two disjoint rectangles.

The first set of constraints comes from the fact that the $a_{ir}$’s are probabilities and therefore nonnegative.

$$t_r \pi_{i1} + (1 - t_r) \pi_{i2} \geq 0 \Rightarrow \begin{cases} t_r \geq \frac{-\pi_{i2}}{\pi_{i1} - \pi_{i2}} \pi_{i1} - \pi_{i2} > 0 \\ t_r \leq \frac{\pi_{i2}}{\pi_{i2} - \pi_{i1}} \pi_{i2} - \pi_{i1} > 0 \end{cases} \quad (1.3.3)$$

The inequalities (1.3.3) must hold for every value $i$ that $Y$ takes and for $r = 1, 2$. Define

$$t_+ = \min_{\pi_{i2} - \pi_{i1} > 0} \frac{\pi_{i2}}{\pi_{i2} - \pi_{i1}} \geq 1 \quad \text{and} \quad t_- = \max_{\pi_{i1} - \pi_{i2} > 0} \frac{-\pi_{i2}}{\pi_{i1} - \pi_{i2}} \leq 0. \quad (1.3.4)$$

We have $t_- \leq t_r \leq t_+$ for $r = 1, 2$. The parameters $t_+$ and $t_-$ are well-defined when $\pi_1 \neq \pi_2$. The points $p(t_+)$ and $p(t_-)$ are the extreme points of the line segment that is the intersection between the line $p(t)$ and the $(I - 1)$-simplex.

The second set of constraints comes from part (ii) of Proposition 1.2.2. Since $\pi_1$ and $\pi_2$ lie in the convex hull of $a_1$ and $a_2$, we must have either $t_1 \leq 0$ and $t_2 \geq 1$, or vice versa. Thus $(t_1, t_2)$ must lie either in $[t_-, 0] \times [1, t_+]$ or in $[1, t_+] \times [t_-, 0]$.

**Theorem 1.3.1.** Suppose $X^*$ and $X$ are binary. Suppose $X \not\perp Y$. Then

$$P(Y|X^* = r) = t_r P(Y|X = 1) + (1 - t_r) P(Y|X = 2) \quad r = 1, 2.$$  

$$(t_1, t_2) \in [t_-, 0] \times [1, t_+] \cup [1, t_+] \times [t_-, 0].$$

where $t_-$ and $t_+$ are as defined in equation (1.3.4). These bounds are sharp.

With a slight abuse of notation, we can say that the columns of $A$ lie in the union of two disjoint rectangles given by $[p(t_-), \pi_2] \times [\pi_1, p(t_+)] \cup [\pi_1, p(t_+)] \times [p(t_-), \pi_2]$. This set has
all the nice properties from Example 1.3.1. Although it is disconnected, it consists of two convex sets which correspond to a switching of the values of $X^*$ and it can be completely described by the two scalar parameters $t_+$ and $t_-$. 

In Example 1.3.1, when $Y$ is also binary, the points $p(t_+)$ and $p(t_-)$ are simply $(1,0)$ and $(0,1)$ (or vice versa). However, when $Y$ can take more than two values, typically we can do better than this. The parameters $t_-$ and $t_+$ defined in (1.3.4) depend on all values of $Y$. Each additional value $i$ of $Y$ gives us more information about the probability ratio $P(Y|X=1)/P(Y|X=2)$, which in turn can improve our bounds on the misclassification probabilities and on $P(Y|X^*=r)$.

**Example 1.3.2.** Suppose that $X$ and $X^*$ are binary and that $Y$ can take one of three values. The table below gives the observed conditional distribution of $Y$ given $X$ and the implied bounds for $P(Y|X^*)$.

| $Y$ | $P(Y|X=1)$ | $P(Y|X=2)$ | $P(Y|X^*=1)$ | $P(Y|X^*=2)$ |
|-----|------------|------------|--------------|--------------|
| $Y=1$ | 0.7 | 0.6 | (0.7,0.75) | (0.5,0.6) |
| $Y=2$ | 0.2 | 0.1 | (0.2,0.25) | (0.1,0.1) |
| $Y=3$ | 0.1 | 0.3 | (0.1,0.1) | (0.3,0.5) |

Suppose now that we are interested only in the events $Y=1$ and $Y \neq 1$ conditional on $X^*$. Ignoring the information contained in the higher values of $Y$ leads to worse bounds for $P(Y=1|X^*)$. 

---

Figure 1.3: The line passing through $\pi_1$ and $\pi_2$. 

\[ p(t_-) \quad a_1 \quad \pi_2 = p(0) \quad \pi_1 = p(1) \quad a_2 \quad p(t_+) \]
Even though we are interested in a subset of the range of $Y$, it is useful to look at the entire range of $Y$ in order to calculate the correct bounds for the misclassification probabilities, and hence for the conditional distribution of $Y$.

At this point, it is natural to ask whether we get any additional information when the observed variable $X$ can take more values than the true variable $X^*$, because $X$ and $Y$ play somewhat analogous roles in our set-up. This question is more sensible if you think of the problem not as one of misclassification but as a general latent class problem, and consider the case where the number of latent classes is fewer than the number of possible values of the observed characteristic.

| $Y$ | $P(Y|X = 1)$ | $P(Y|X = 2)$ | $P(Y|X^* = 1)$ | $P(Y|X^* = 2)$ |
|-----|--------------|--------------|----------------|----------------|
| $Y = 1$ | 0.7          | 0.6          | (0.7,1)        | (0.0,6)        |
| $Y \neq 1$ | 0.3         | 0.4          | (0.0,3)        | (0.4,1)        |

In Figure 1.3, an additional value of $X$ corresponds to another point $\pi_3$ on the line $p(t)$. All three observed points $\pi_j$ must lie on the same line as they are each constrained to be a convex combination of the two unobserved points $a_1$ and $a_2$. The third point leads to a smaller identified set if it is closer to the boundary of the simplex than the two existing points.

Additional values of $X$ do not lead to an improvement in the ‘outer’ bounds, but they may contain information that lets us rule out less extreme distributions for $Y|X^*$ leading to better ‘inner’ bounds.

Finally, in the case where both $Y$ and $X$ can take more than two values and $X^*$ is binary, we can imagine a test for the conditional independence assumption. As noted above, the binary nature of $X^*$ requires that the observed points $\pi_j$ lie on the same line. A violation
would mean that $Y$ and $X$ are in fact not independent conditional on $X^*$.

### 1.3.1 Regression of $Y$ on $X^*$

Now that we have sharp bounds on the conditional distribution of $Y$ given $X^*$, we can use these to bound functionals of the distribution. We are particularly interested in the regression of $Y$ on $X^*$, which in the case of binary $X^*$ is simply the difference between the two conditional means. It follows from Theorem 1.3.1 that

$$E[Y|X^* = 1] - E[Y|X^* = 2] = |t_1 - t_2| (E[Y|X = 1] - E[Y|X = 2]).$$  \hfill (1.3.5)$$

The expression on the left-hand side is the slope parameter of the simple linear regression of $Y$ on $X^*$ (Aigner 1973, Bollinger 1996) and also the treatment effect from Lewbel (2007). The true slope is a multiple of the observed slope from the regression of $Y$ on $X$. The bounds on $(t_1, t_2)$ imply that the multiplier $|t_1 - t_2|$ lies in the interval $[1, t_+ - t_-]$, that is, the true slope is at least as large as the observed slope from the regression of $Y$ on $X$.

This is the so-called ‘attenuation effect’ of misclassification in a binary explanatory variable, where the observed slope is a lower bound for the true slope parameter.

Note that the bounds apply when we replace $Y$ by any function $\tau(\cdot)$ of $Y$. Note also that we obtain point identification of the regression of $\tau(Y)$ on $X^*$ when it is zero. The true variable has no effect on the expected value of $\tau(Y)$ if and only if the observed variable has no effect on it.

One function we could consider is the indicator function that $Y$ takes a certain value $i$. The expectation of this function is the probability that $Y = i$. For every value $i$ of $Y$ we
have

\[ P[Y = i | X^* = 1] - P[Y = i | X^* = 2] = s \left( P[Y = i | X = 1] - P[Y = i | X = 2] \right) \quad s \in [1, t_+ - t_-]. \]

Thus the probability mass functions of \( Y \) conditional on \( X = 1 \) and \( X = 2 \) are ‘closer’ to each other than those of \( Y \) conditional on \( X^* = 1 \) and \( X^* = 2 \).

Given (1.3.6), it is not surprising that nondifferential measurement error leads to attenuation when the misclassified explanatory variable is binary. Consider the two subpopulations corresponding to \( X^* = 1 \) and \( X^* = 2 \) and let \( \theta_r \) be the mean of \( Y \) in group \( r \). Suppose \( \theta_1 > \theta_2 \) i.e, individuals in group 1 have on average larger values of \( Y \) than those in group 2. When misclassification is nondifferential, the two groups are mixed in such a way that the mean of \( Y \) in the observed group 1 is lower than the true mean \( \theta_1 \) while the mean of \( Y \) in the observed group 2 is higher than \( \theta_2 \). The observed difference in means is thus smaller than the true difference and the effect of \( X^* \) on \( Y \) is attenuated.

The focus of this paper is the upper bound of the attenuation effect \((t_+ - t_-)\). The two parameters \( t_+ \) and \( t_- \) correspond respectively to the minimum and maximum of the probability ratios \( \pi_{i1}/\pi_{i2} \).

### 1.3.2 Continuous \( Y \)

We can perform a similar exercise for continuous \( Y \) taking values in \( Y \subseteq \mathbb{R} \). Let \( f(\cdot | j) \) denote the conditional density of \( Y \) given \( X = j \) and \( f^*(\cdot | r) \) denote the conditional density of \( Y \) given \( X^* = r \). Assume that \( X^* \) and \( X \) are not independent, so that \( \mathbf{B} \) is invertible.
As before, we can write

\[ f^*(y|r) = t_r f(y|1) + (1 - t_r) f(y|2) \geq 0 \quad y \in \mathcal{Y}, r = 1, 2. \quad (1.3.7) \]

where \( t_1 \) and \( t_2 \) are elements of \( B^{-1} \). Define

\[ t_+ = \inf \{ y : f(y|1) < f(y|2) \} \frac{f(y|2)}{f(y|1)} \geq 1 \]

\[ t_- = \sup \{ y : f(y|1) > f(y|2) \} \frac{-f(y|2)}{f(y|1)} \leq 0. \quad (1.3.8) \]

We can bound \((t_1, t_2)\) by the rectangles \([t_-, 0] \times [1, t_+] \cup [1, t_+] \times [t_-, 0]\). In turn, these bounds imply restrictions for the mean regression of \( Y \) on \( X^* \)

\[ E[Y|X^* = r] = t_r E[Y|X = 1] + (1 - t_r) E[Y|X = 2] \quad r = 1, 2. \quad (1.3.9) \]

with \((t_1, t_2) \in [t_-, 0] \times [1, t_+] \cup [1, t_+] \times [t_-, 0]\). Note that this is strictly better than the observation that the observed conditional means \((E[Y|X = 1], E[Y|X = 2])\) are convex combinations of the true conditional means \((E[Y|X^* = 1], E[Y|X^* = 2])\). The latter only provides us with the ‘inner’ bounds for the regression given by the observed means and corresponding to setting \( t_1 = 0 \) and \( t_2 = 1 \) or vice versa.

The ‘outer’ bounds given by (1.3.8) follow from the additional observation that the true conditional densities \((f^*(\cdot|1), f^*(\cdot|2))\) can be written as affine combinations of the observed conditional densities. It is easy to see from (1.3.7) and from the discrete \( Y \) case that for every value \( \tilde{y} \) of \( Y \), the value of \( f(\tilde{y}|\cdot) \) implies restrictions for the misclassification probabilities \( P(X^* = 1|X = \cdot) \) and therefore for the entire conditional distribution of \( Y \) given \( X^* \) and not merely \( f^*(\tilde{y}|\cdot) \).

In Chapter 2, we go into more detail on the identification and estimation of bounds for
the nonparametric regression of a continuous outcome on a binary misclassified regressor, when the misclassification is non-differential.

1.4 Geometric Properties of the Identified Set in Higher Dimensions

In this section, we study the geometric properties of $\mathcal{A}(\Pi)$ when $\Pi$ has more than two columns – i.e. when $X^*$ can take more than two values. In the previous section, we saw that the identified set has some very nice properties when $X^*$ is binary. It has exactly two connected components corresponding to the two permutations of the values that $X^*$ can take. Each component is convex; in fact, it is a rectangle and can be described completely using only two parameters. In contrast, none of these properties hold when $X^*$ can take more than two values.

In what follows, we assume that $\text{rank}(B) = R$, so that the results of Proposition 1.2.2 hold. Let $C$ be the intersection of the affine hull of the columns of $\Pi$ and the simplex $\Delta^{I-1}$. Recall that $C$ is the convex polytope in which the columns of $\Pi$ and $A$ are constrained to lie. For example, when $R = 3$, $C$ is a single point if $\text{rank}(\Pi) = 1$, a line segment if $\text{rank}(\Pi) = 2$, or a convex polygon with at most $I$ vertices if $\text{rank}(\Pi) = 3$.

First note that the set $\bar{\mathcal{A}}$ is easy to characterize when $\text{rank}(\Pi) = 2$. In this case, the columns of $\Pi$ and $A$ are constrained to lie on a line and we can proceed in a similar way to Section 2. The conditional independence assumption implies restrictions for exactly two columns of $A$. The remaining $R - 2$ columns can be anywhere on the line segment $C$. The
set \( \bar{\mathcal{A}} \) is given by the union of permutations of the \( R \)-dimensional rectangle

\[
C_- \times C_+ \times C \times \cdots \times C
\]

where \( C_- \) and \( C_+ \) are the line segments on either side of the columns of \( \Pi \). It is easy to see that this set is connected but not convex.

When \( \text{rank}(\Pi) = 2 \), all the information in \( \Pr(Y|X) \) that is useful in learning about \( \Pr(Y'|X^*) \) under the conditional independence assumption is contained in only two of its columns or only two values of the observed regressor \( X \). Further, if we assume that \( B \) is invertible, we can learn about the unknown distribution \( \Pr(Y|X^*) \) for exactly two values of \( X^* \). So, we could as well have assumed that \( X^* \) and \( X \) are binary.

Now suppose that \( \text{rank}(\Pi) \geq 3 \).

One question we might ask is when is the conditional independence assumption least informative, or when is \( \bar{\mathcal{A}} \) the ‘biggest’.

Consider the construction in Figure 1.4 with \( \text{rank}(\Pi) = R = 3 \). We have drawn \( C \) to be a triangle but it can have up to \( I \) edges. Let \( \Pi \) denote the triangle that is formed by the columns of \( \Pi \) and \( A \) denote the triangle that is formed by the columns of \( A \). The identified set \( \bar{\mathcal{A}} \) is the set of all triangles \( A \) that contain the triangle \( \Pi \) and are contained in \( C \).

Let \( S_\pi \) be the smallest circle that contains the triangle \( \Pi \). Let \( A_{\text{reg}} \) be an equilateral triangle that contains \( S_\pi \). Let \( S_\alpha \) be the circle obtained by rotating \( A_{\text{reg}} \) around the centre of \( S_\pi \). Suppose \( \Pi \) is such that \( S_\alpha \) lies inside \( C \) for some \( A_{\text{reg}} \). This is true when the columns of \( \Pi \) are close to each other and far from the boundary of \( C \).

We claim that, while still containing \( \Pi \),
Figure 1.4: $\tilde{A}(\Pi)$ is connected up to re-labelling of the values of $X^*$.

i. any equilateral triangle can be rotated on $S_a$, and

ii. any triangle $A$ that contains $\Pi$ can be shrunk or stretched to some equilateral triangle on $S_a$.

These operations give us a path through the set of triangles that contain $\Pi$. Thus the set $\tilde{A}(\Pi)$ is a connected set up to re-ordering of the vertices of $A$ or permutation of columns of $A$. Moreover, it can be shown that some (but not all) of the re-orderings of the vertices of $A$ are in fact connected through rotations of the triangle $A_{reg}$ on the circle $S_a$.

Let $l, m$ and $n$ be the vertices of an equilateral triangle on $S_a$. Then we can find a path in $\tilde{A}(\Pi)$ from the triangle $(l, m, n)$ to the triangle $(m, n, l)$ by rotating it on the circle $S_a$. Similarly any two permutations of these vertices that come from a rotation of the triangle $(l, m, n)$ are connected to each other by a path. On the other hand, we can show that there is no path between the triangles $(l, m, n)$ and $(l, n, m)$. Note that $(l, n, m)$ is obtained by a reflection and not a rotation of $(l, m, n)$.$^1$

In the above discussion we assumed that $\Pi$ had full column rank. If instead $\text{rank}(\Pi) =$

---

$^1$The result used here and proved in the appendix is that the set of regular $N$-simplices with vertices on the $(N-1)$-sphere consists of exactly two connected sets.
$3 < R$, $\Pi$ is now a convex polygon with at most $R$ vertices. Suppose as before that $\Pi$ is small and close to the centre of the containing polygon $C$. We can still find equilateral triangles $A_{\text{reg}}$ that contain $\Pi$ and have vertices on a circle $S_a$ that is contained in $C$. However the matrix $A_{\text{reg}}$ corresponding to $A_{\text{reg}}$ now has $R - 3$ free columns that can lie anywhere in $C$. As a result, we can find paths between any two permutations of the columns of the matrix $A_{\text{reg}}$. Consequently, the identified set $\bar{A}(\Pi)$ is a connected set.

Claim 1.4.1. Let $\text{rank}(\Pi) \geq 3$. If the columns of $\Pi$ are ‘close’ to each other and ‘far’ from the boundary of $C$ then the following statements hold.

i. If $\text{rank}(\Pi) < R$, $\bar{A}$ is connected but not convex.

ii. If $\text{rank}(\Pi) = R$, $\bar{A}$ consists of exactly two connected components. These components correspond to ‘even’ and ‘odd’ permutations of the columns of $A$. These components are not convex.

Note that when $\text{rank}(\Pi) = 2$, the conclusions of Claim 1.4.1 are always true regardless of how far the columns of $\Pi$ are from each other and from the end-points of the line segment $C$. In this case, the connectedness and convexity properties of $\bar{A}(\Pi)$ do not depend on the relative position of the columns of $\Pi$. In contrast, in higher dimensions, as the columns of $\Pi$ move apart, the identified set can change in interesting ways.

Under the conditional independence assumption, the information about $A$ contained in $\Pi$ comes from those columns of $\Pi$ that are the vertices of their convex hull. Suppose there is a column $\pi_j$ that can be written as a convex combination of $\{\pi_r, r \neq j\}$. Then any $A$ that contains $\{\pi_r, r \neq j\}$ in the convex hull of its columns also contains $\pi_j$. When
\text{rank}(\mathbf{\Pi}) = 2$, the convex hull of its columns is a line segment with exactly two columns of $\mathbf{\Pi}$ at its end-points or vertices. On the other hand, say when \text{rank}(\mathbf{\Pi}) = 3, the convex hull of its columns is a polygon with anywhere between 3 and $R$ columns of $\mathbf{\Pi}$ at its vertices.

Moreover, when the observed variable $X$ can take more values that the true variable $X^*$, the convex hull of the columns of $\mathbf{\Pi}$ can have more than $R$ vertices. An additional value of $X$ amounts to another opportunity to change the distribution of $Y$ conditional on the observed variable. If this new conditional distribution for $Y$ is not a convex combination of the existing distributions, it can help to rule out more candidates for the distribution of $Y$ given $X^*$.

Thus, the rank of $\mathbf{\Pi}$ is not a perfect measure of the information contained in $\mathbf{\Pi}$. We also care about the distance of the columns of $\mathbf{\Pi}$ from each other and from the boundary of the simplex. When these are close to each other and far from the boundary, the distribution of $Y$ does not change much as we change the values of $X$, and $X$ is relatively uninformative about the relationship between $Y$ and $X^*$. Again, we can draw an analogy with an instrumental variable that is not very correlated with the outcome.

Now consider the other extreme in which $\mathcal{A}(\mathbf{\Pi})$ is finite. It is easy to see that a necessary condition for finiteness is that the columns of $\mathbf{\Pi}$ must lie on the boundary of the containing polytope $C$.

\textit{Claim 1.4.2.} i. $\mathcal{A}(\mathbf{\Pi})$ is finite only if the columns of $\mathbf{\Pi}$ are on the boundary of $C$.

ii. Furthermore, we conjecture that if every column of $\mathbf{\Pi}$ has exactly one zero and every row of $\mathbf{\Pi}$ has at most one zero, then $\mathcal{A}(\mathbf{\Pi})$ is finite.
Part ii of Claim 1.4.2 suggests a sufficient condition for finiteness of $A(\Pi)$. This condition is a statement about the position of zeroes in the observed distribution $\Pr(Y|X)$. It says that for any value $j$ of $X$, the observed distribution $\Pr(Y|X = j)$ has exactly $I - 1$ points of support. Moreover, the support of $Y$ cannot be the same for any two distinct values of $X$.

Recall that when $X^*$ is binary, we obtain finiteness if and only if the columns of $\Pi$ are the end-points of the line segment $C$. This condition is also a statement about how the support of $\Pr(Y|X)$ changes with $X$. In this case, the solutions are given by

$$A = \Pi = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}, \quad B = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad A = \begin{bmatrix} \pi_2 & \pi_1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$ 

Thus, when $X^*$ is binary, if the support of $\Pr(Y|X = 1)$ is not contained in the support of $\Pr(Y|X = 2)$ and vice versa, we obtain identification of the true conditional distribution $\Pr(Y|X^*)$ up to permutations of the values of $X^*$. Moreover, the true conditional distribution is equal to the observed conditional distribution up to re-labelling.

Claim 1.4.2 tells us that such a condition on the support of $\Pr(Y|X)$ does give us finiteness of the identified set for $\Pr(Y|X^*)$ even in higher dimensions. However, it does not always give us identification up to permutations of the values of $X^*$.

*Example* 1.4.1. Suppose $Y$ can take 4 values and $X^*$ can take 3 values. Consider the

---

2The sufficient condition in Claim 1.4.2 constrains the columns of $\Pi$ to lie on distinct $(\text{rank}(\Pi) - 2)$-dimensional faces of $C$. It possible to construct other sufficient conditions for finiteness based on changing the support of $Y$ given $X$. For example, we could constrain the columns of $\Pi$ to lie on distinct $(\text{rank}(\Pi) - 3)$-dimensional faces of $C$. 

(a) Slicing $\Delta^3$.

(b) Finiteness.

(c) Identification up to Permutations.

Figure 1.5: Finiteness does not imply identification up to permutations.

The identified set $\bar{A}(\Pi_1)$ contains only permutations of the matrix $\Pi_1$ itself. On the other hand, $\Pi_2$ admits another solution.

The difference between $\Pi_1$ and $\Pi_2$ lies in the way that the affine hulls of their respective columns intersect the simplex. $\Pi_1$, $\Pi_2$ and their containing polytopes are plotted in Figure 1.5. $C_2$ is a triangle that contains $\Pi_2$, thus the vertices of $C_2$ are admissible values for the columns of $A$. On the other hand, $C_1$ is a quadrilateral which only admits the columns of $\Pi_2$ as values for the columns of $A$. 

$$
\Pi_1 = \begin{bmatrix}
1/3 & 0.1 & 0.1 \\
1/3 & 0 & 0.8 \\
1/3 & 0.8 & 0 \\
0 & 0.1 & 0.1
\end{bmatrix}, \quad \Pi_2 = \begin{bmatrix}
1/3 & 0.8 & 0.8 \\
1/3 & 0 & 0.1 \\
1/3 & 0.1 & 0 \\
0 & 0.1 & 0.1
\end{bmatrix}
$$

$$
\Pi_2 = \begin{bmatrix}
1/3 & 1/3 & 15/17 \\
2/3 & 0 & 0 \\
0 & 2/3 & 0 \\
0 & 0 & 2/17
\end{bmatrix}, \quad \begin{bmatrix}
1/2 & 0 & 3/20 \\
1/2 & 3/20 & 0 \\
0 & 17/20 & 17/20
\end{bmatrix}
$$

following observed probability matrices $\Pi_1$ and $\Pi_2$. 

The identified set $\bar{A}(\Pi_1)$ contains only permutations of the matrix $\Pi_1$ itself. On the other
Now, when $X$ takes more than 3 values, there may be no triangle in $C_1$ that contains all columns of $\Pi_1$. This suggests an over-identification test for the conditional independence assumption.

As the dimension of the containing polytope $C$ increases, we can find other admissible values for $A$ that need not be the vertices of $C$. This is interesting because - first, it illustrates the richness of the identified set $\tilde{A}(\Pi)$ in higher dimensions and second, it shows that even the very restrictive case where the support of $\Pr(Y|X)$ shifts with $X$ is not sufficient to guarantee that there is no misclassification.

However, when the elements of the finite identified set are far apart, we can conceive of restrictions on size of the misclassification probabilities that will eliminate all but one candidate distribution for $(Y, X, X^*)$.

### 1.5 Conclusion

In this paper, we have formulated the problem of nondifferential measurement error in a discrete explanatory variable as a problem of factorizing the conditional probability distribution of the outcome and the error-ridden measure. When all variables are discrete, this is a problem of matrix factorization.

We have completely characterized the set of solutions to this factorization problem when the mismeasured regressor is binary. We have shown that, in contrast to the binary case, this set can be quite complicated in higher dimensions. In particular, we have shown that this set can consist of topologically separated components which are not merely the result of a re-labelling of the values of the true regressor.
In conclusion, we would like to re-draw attention to a parallel between the classical EIV model and the problem considered in this paper. The width of the bounds on the parameter of interest in the linear EIV model depends on the strength of the linear relationship between the observed variables, which is characterized by the regression $R^2$. Similarly, the size of the identified set in the current setting depends on the strength of the relationship between the observed variables. However, there is no single number that characterizes the dependence between $Y$ and $X$.

We have seen that various features of the observed distribution $\Pi$ are relevant - the rank of $\Pi$, the distance between its columns, their distance from the boundary of the containing polytope and the shape of the containing polytope itself. Unlike classical EIV, these features can sometimes lead to an empty identified set, suggesting a test for over-identification.
Appendix

Proofs from Section 2.

Recall the following notions from linear algebra.

**Definition 1.5.1.** Let $P$ be a real-valued $M \times N$ matrix. The **convex hull** of the columns of $P$ is

$$\text{conv}(P) \equiv \left\{ \sum_n \lambda_n p_{.n} : \lambda_n \in \mathbb{R}_+ \text{ and } \sum_n \lambda_n = 1 \right\}.$$

**Definition 1.5.2.** Let $P$ be a real-valued $M \times N$ matrix. The **affine hull** of the columns of $P$ is

$$\text{aff}(P) \equiv \left\{ \sum_n \lambda_n p_{.n} : \lambda_n \in \mathbb{R} \text{ and } \sum_n \lambda_n = 1 \right\}.$$

**Proposition 1.2.1**

**Proof.** Suppose $A \in \mathcal{A}(\Pi)$. Then the $j$-th column of $\Pi$ can be written as

$$\pi_j = Ab_j \text{ or } \begin{bmatrix} P(Y = 1|X = j) \\ \vdots \\ P(Y = I|X = j) \end{bmatrix} = \begin{bmatrix} P(Y = 1|X^* = 1) & \cdots & P(Y = 1|X^* = R) \\ \vdots & \ddots & \vdots \\ P(Y = I|X^* = 1) & \cdots & P(Y = I|X^* = R) \end{bmatrix} \begin{bmatrix} P(X^* = 1|X = j) \\ \vdots \\ P(X^* = R|X = j) \end{bmatrix}$$

Conversely, if $\pi_j$ lies in the convex hull of the columns of $A$, then we can find $b_j \in \Delta^{R-1}$ such that $\pi_j = Ab_j$. □

**Proposition 1.2.2**

**Proof.** The ‘only if’ part follows from Proposition 1.2.1. Suppose there is $B \in (\Delta^{R-1})^R$
such that \( \mathbf{\Pi} = \mathbf{A}\mathbf{B} \) and \( \mathbf{B} \) has full rank \( R \). Let \( \mathbf{T} = \mathbf{B}^{-1} \). Then we have

\[
\mathbf{A} = \mathbf{\Pi}\mathbf{T} \quad \text{or} \quad a_r = \sum_{j=1}^{R} t_{rj} \pi_j.
\]

Moreover, since the columns of \( \mathbf{B} \) add up to one, we have \( \sum_{j} t_{rj} = 1 \) for every \( r \). Hence \( a_r \in \text{aff}(\mathbf{\Pi}) \) for every \( r \). \( \square \)

**Proofs from Section 4.**

**Lemma 1.5.1.** *The set of regular \( N \)-simplices with vertices on the \((N-1)\)-sphere consists of exactly two connected components.*

**Proof.** Let \( S^{N-1} \) denote the \((N-1)\)-dimensional sphere. Let \( \mathcal{A}_N \) be the set of regular \( N \)-simplices with vertices on \( S^{N-1} \).

Without loss of generality we can assume that \( S^{N-1} \subset \mathbb{R}^N \) is centered at the origin and has radius 1. The set \( \mathcal{A}_N \) is simply the set of \( N \times (N + 1) \) matrices

\[
\mathbf{A}_{\text{reg}} = \begin{bmatrix} a_0 & a_1 & \cdots & a_N \end{bmatrix}
\]

with the property that \( \|a_i\| = 1 \), \( \langle a_i, a_j \rangle = -\frac{1}{N} \) for \( i \neq j \) and \( a_0 + a_1 + \cdots + a_N = 0 \). The columns of \( \mathbf{A}_{\text{reg}} \) are the vertices of a regular simplex in \( \mathcal{A}_N \).

Now, let \( O(N) \) be the set of orthogonal matrices; i.e. the set of \( N \times N \) matrices \( \mathbf{R} \) with \( \mathbf{RR}^T = \mathbf{I}_N \). We claim that \( O(N) \) and \( \mathcal{A}_N \) are homeomorphic. As it is well-known (Horn and Johnson 1990) that \( O(N) \) has exactly two connected components, this claim will prove the lemma.
To prove the claim, note that if $A \in \mathcal{A}_N$ and $R \in O(N)$ then $RA$ is also in $\mathcal{A}_N$. Fix a matrix $\tilde{A} \in \mathcal{A}_N$. Define $f : O(N) \to \mathcal{A}_N$ as $f(R) = R\tilde{A}$. Define $g : \mathcal{A}_N \to O(N)$ as

$$g(A) = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix} \begin{bmatrix} \tilde{a}_1 & \cdots & \tilde{a}_N \end{bmatrix}^{-1} = A_{1:N} \tilde{A}_{1:N}^{-1}.$$  

To show that $g(A)$ is indeed orthogonal

$$g(A)g(A)^T = A_{1:N} \tilde{A}_{1:N}^{-1} \begin{bmatrix} \tilde{A}_{1:N}^{-1} \end{bmatrix}^T A_{1:N}^T$$

$$= A_{1:N} \begin{bmatrix} A_{1:N}^T \tilde{A}_{1:N} \end{bmatrix}^{-1} A_{1:N}^T$$

$$= A_{1:N} \begin{bmatrix} A_{1:N}^T A_{1:N} \end{bmatrix}^{-1} A_{1:N}^T = I_N$$

where the third equality follows from the property that $\|a_i\| = 1$ and $\langle a_i, a_j \rangle = -\frac{1}{N}$ for $i \neq j$.

Suppose rank($\Pi$) = $R$. Let $\Pi$ be the simplex given by the columns of $\Pi$ and $A$ be the simplex given by the columns of $A$. Recall that $C$ is the convex polytope given by the intersection of $\text{aff}(\Pi)$ and $\Delta^{I-1}$. The dimension of $C$, $\Pi$ and $A$ is $N = R - 1$.

Let $S_{\rho}(v)$ be the $(N - 1)$-sphere in $C$ with radius $\rho$ and centre at $v$.

**Lemma 1.5.2.** Let $v \in C$ and $\rho > 0$ be such that $\Pi$ is contained in the sphere $S_{\rho}(v)$ and $S_{N\rho}(v)$ is contained in $C$. Let $A_{\text{reg}}$ be a regular simplex inscribed in $S_{N\rho}(v)$. Then there is a path between $A$ and $A_{\text{reg}}$.

**Proof.** Without loss of generality, let the vertices of $A$ be on the boundary of $C$. Then at least one $(N - 1)$-face of $A$ does not intersect the interior of the $\rho$-ball with centre at $v$.

To see this, note that $S_{\rho}(v)$ is inscribed in $A_{\text{reg}}$. Suppose each face of $A$ intersects $S_{\rho}(v)$.  

35
Then at least one vertex of $A$ is closer to $v$ than the vertices of $A_{reg}$ i.e., $d(v, a_r) < N\rho$ for some $r$. But this contradicts our assumption that $S_{N\rho}(v)$ is contained in $C$.

Suppose that the vertices $a_1, \ldots, a_N$ span a face of $A$ that does not intersect the interior of the $\rho$-ball with centre at $v$, and let $a_{N+1}$ be the remaining vertex. For each $r = 1, \ldots, N$, there is a vertex $\alpha_r$ on the line segment between $a_r$ and $a_{N+1}$ such that the simplex spanned by $\alpha_1, \cdots, \alpha_N$ is tangent to $S_\rho(v)$. The path that moves each $a_r$ to $\alpha_r$ along the line segment remains in $A(II)$. Now we may stretch the $(N-1)$-simplex spanned by the $\alpha_r$ to a regular $(N-1)$-simplex which contains the $\alpha_r$ (and which therefore is also tangent to $S_\rho(v)$) and whose vertices, say $a'_1, \cdots, a'_N$, lie on $S_{N\rho}(v)$. Finally, move $a_{N+1}$ to the unique vertex $a'_{N+1}$ on $S_{N\rho}(v)$ which forms a regular $N$-simplex with the $a'_1, \cdots, a'_N$. This completes the proof.

Claim 1.4.1

Proof. Lemma 1.5.1 and Lemma 1.5.2 together prove part ii of Claim 1.4.1.

Claim 1.4.2

Proof. i. Suppose $\pi_1$ is in the interior of $\Delta^{l-1}$. Define $a(t) \in \mathbb{R}^l$ as follows.

$$a(t) = t\pi_1 + (1-t)\pi_2.$$
Note that $\pi_1 \in \text{conv}(a(t), \pi_2)$ for $t > 1$

$$\pi_1 = \frac{1}{t} a(t) + \left(1 - \frac{1}{t}\right) \pi_2.$$ 

Moreover, as $\pi_1$ is in the interior of $\Delta^{I-1}$, $a(t) \in \Delta^{I-1}$ for $t$ small enough. \hfill \square
Chapter 2

Identification and Estimation of Bounds in a Regression with a Binary Misclassified Regressor

2.1 Introduction

In this chapter, we revisit the subject of Chapter 1 in the case where $Y$ is continuous and $X^*$ is binary. We make an additional standard assumption: that the misclassification in $X^*$ is not serious enough to reverse the labeling.

We have already shown that the identified set for nonparametric identification of the joint distribution of continuous $Y$ and binary $X^*$, $X$ is completely described by two scalar parameters. In the next section, we re-prove this result with a slight change of notation that allows for better comparison with bounds existing in the literature. In addition, we explicitly derive the bounds for all parameters of interest.

The two scalar parameters mentioned above correspond to the infimum and supremum of the ratio of observed densities of $Y$ conditional on the two values of $X$. We examine the behavior of this density ratio and the resulting bounds using a series of examples in which
$Y$ given $X^*$ is assumed to be normally distributed. Specifically, we compare our bounds to those obtained by Bollinger (1996).

We propose a simple estimation technique which involves taking the minimum and maximum of the ratio of kernel density estimates over a trimmed set. We assess its performance using Monte Carlo simulations.

### 2.2 Model and Partial Identification

In this section, we will derive bounds for the joint distribution of $(Y, X^*)$ when $Y$ is a continuous variable with support $Y \subset \mathbb{R}$ and $X^*$ is a binary or indicator variable. We do not observe $X^*$ but a misclassified version of it, $X$, which is also assumed to be binary.

Here, we think of $Y$ as the response variable and $X^*$ as the predictor. The object of interest is their joint distribution, which is not directly observed because of measurement error in $X^*$.

We do not make any functional form assumptions on the relationship between $X^*$ and $Y$. However, because $X^*$ is binary, we can always write the following.

$$y = \alpha + \beta X^* + \epsilon$$

$$\alpha = E[y|X^* = 0]$$

$$\beta = E[y|X^* = 1] - E[y|X^* = 0]$$

By construction, the equation error is mean independent of $X^*$, however there are no other restrictions on the distribution of $\epsilon$.

We make the following two assumptions on the measurement error.

- **Assumption 1**: $Y$ is independent of $X$ conditional on $X^*$.
• Assumption 2: \( \Pr[X = 1|X^* = 1] > \Pr[X = 1|X^* = 0]. \)

In the measurement error literature, Assumption 1 is referred to as nondifferential measurement error, and \( X \) is called a surrogate for \( X^* \). The observed variable gives us no additional information about the response other than that which is contained in the true predictor.

There are two immediate observations to be made about Assumption 2. First, the strict inequality means that the surrogate is actually related to the true predictor. If we were equally likely to observe the two values of \( X \) given the truth, \( X \) would be independent of \( X^* \). Second, we are more likely to observe the correct classification, which means that \( X \) and \( X^* \) are positively correlated. Together with Assumption 1, this will mean that sign of the relationship between the response and the predictor is unaffected by the misclassification.

We now introduce some notation. Let \( f(y) \) be the marginal density of \( Y \), let \( f_r(y) \) be the density function of the distribution of \( Y \) conditional on \( X^* = r \) and let \( f_j(y) \) be the density function of the distribution of \( Y \) conditional on \( X = j \). Let \( P^* \) be the marginal probability of observing a success in \( X^* \), and \( P \) be the marginal probability of observing a success in \( X \). Let \( p \equiv \Pr[X = 1|X^* = 0] \) be the probability of observing a misclassified success i.e. observing a success when the true value is failure, and \( q \) be the probability of observing a misclassified failure.

Directly, and from Assumptions 1 and 2, we have the following relationships between
the observed and unobserved distributions.

\[ f(y) = P^* f_1^*(y) + (1 - P^*) f_0^*(y) \]  
\[ P = (1 - q) P^* + p (1 - P^*) \]  
\[ P f_1(y) = (1 - q) P^* f_1^*(y) + p (1 - P^*) f_0^*(y) \]  
\[ (1 - P) f_0(y) = q P^* f_1^*(y) + (1 - p) (1 - P^*) f_0^*(y) \]

In particular, we can write each of the observed conditional density functions as a distinct convex combination of the unobserved conditional densities. In matrix notation,

\[
\begin{pmatrix}
  f_1(y) \\
  f_0(y)
\end{pmatrix} = \begin{pmatrix}
  (1-q) P^* & p(1-P^*) \\
  q P^* & (1-p)(1-P^*)
\end{pmatrix} \begin{pmatrix}
  f_1^*(y) \\
  f_0^*(y)
\end{pmatrix}
\]

(2.2.5)

By Assumption 2, \( p + q < 1 \) and the above \( 2 \times 2 \) matrix of conditional probabilities of \( X^* \) given \( X \) is invertible. Let

\[
T = \begin{pmatrix}
  (1-q) P^* & p(1-P^*) \\
  q P^* & (1-p)(1-P^*)
\end{pmatrix}^{-1} = \begin{pmatrix}
  (1-p) P^* & -p(1-P) \\
  -q P^* & (1-q)(1-P^*)
\end{pmatrix}.
\]

(2.2.6)

Note that the rows of \( T \) add up to 1. Moreover, with Assumption 2, the off-diagonal entries are negative and the diagonal elements are greater that 1. Thus, the unobserved densities are affine combinations of the observed densities. Compare this with Equation (1.3.7) in Chapter 1.

Let \( t_1 \) and \( t_0 \) denote the diagonal elements of \( T \) that are greater than 1. Now we can write

\[
\begin{pmatrix}
  f_1^*(y) \\
  f_0^*(y)
\end{pmatrix} = T \begin{pmatrix}
  f_1(y) \\
  f_0(y)
\end{pmatrix} = \begin{pmatrix}
  t_1 & 1 - t_1 \\
  1 - t_0 & t_0
\end{pmatrix} \begin{pmatrix}
  f_1(y) \\
  f_0(y)
\end{pmatrix}
\]

(2.2.7)
We can use the above equation to construct bounds on \( t_1 \) and \( t_0 \). These in turn lead to bounds on all unobserved quantities of interest i.e., the joint distribution of \( Y \) and \( X^* \) given by \((f_1^*(y), f_0^*(y), P^*)\) and on the misclassification probabilities \( p \) and \( q \). We collect these in Theorem 1.

**Theorem 2.2.1.** Given Assumptions 1 and 2, we have

\[
1 \leq t_1 \leq \left( 1 - \inf_{y} \frac{f_1(y)}{f_0(y)} \right)^{-1} \equiv \underline{t}_1 \tag{2.2.8}
\]

\[
1 \leq t_0 \leq \left( 1 - \inf_{y} \frac{f_0(y)}{f_1(y)} \right)^{-1} \equiv \underline{t}_0 \tag{2.2.9}
\]

The identified set for \((f_1^*(y), f_0^*(y), P^*, p, q)\) obtains by varying \( t_1 \) in \([1, \underline{t}_1]\) and \( t_0 \) in \([1, \underline{t}_0]\), and using equations (2.2.2), (2.2.6) and (2.2.7). This identified set is sharp.

The bounds for each parameter are

\[
0 \leq p \leq \frac{(\underline{t}_1 - 1)P}{\underline{t}_1 - P} \tag{2.2.10}
\]

\[
0 \leq q \leq \frac{(\underline{t}_0 - 1)(1 - P)}{\underline{t}_0 - (1 - P)} \tag{2.2.11}
\]

\[
\frac{P}{\underline{t}_1} \leq P^* \leq 1 - \frac{1 - P}{\underline{t}_0} \tag{2.2.12}
\]

\[
f_1(y) \preceq f_1^*(y) \preceq \underline{t}_1 f_1(y) + (1 - \underline{t}_1) f_0(y) \quad \forall y \text{ s.t. } f_1(y) \preceq f_0(y) \tag{2.2.13}
\]

\[
f_0(y) \preceq f_0^*(y) \preceq \underline{t}_0 f_0(y) + (1 - \underline{t}_0) f_1(y) \quad \forall y \text{ s.t. } f_0(y) \preceq f_1(y) \tag{2.2.14}
\]
Proof. In this proof, we will derive the bounds for $t_1$ and $t_0$, and show that the resulting identified set for $(f_1^*(y), f_0^*(y), P^*, p, q)$ is sharp. The bounds for each parameter of interest follow directly.

By construction, a lower bound for each of $t_1$ and $t_0$ is 1. To see that this leads to a valid distribution for $(Y, X, X^*)$, we only need to note that this corresponds to the case of no misclassification. In this case, $T$ is the $2 \times 2$ identity matrix, and the misclassification probabilities, $p$ and $q$, are 0. The conditional distributions of $Y$ given $X^*$ are the same as those of $Y$ given $X$, i.e., $f_1^*(y) = f_1(y)$ and $f_0^*(y) = f_0(y)$ for all $y \in \mathcal{Y}$. The probability of a success in $X^*$ is equal to that of a reported success in $X$, i.e., $P^* = P$. Thus, in the absence of additional information, we cannot rule out the possibility that there is no misclassification.

Let us now consider an upper bound for $t_1$. Since $f_1^*$ is always positive, we have

$$t_1 f_1(y) + (1 - t_1) f_1(y) \geq 0 \quad \forall y \in \mathcal{Y} \tag{2.2.15}$$

This implies

$$-\inf_{\{y : f_1(y) < f_0(y)\}} \frac{-f_0(y)}{f_1(y) - f_0(y)} \leq t_1 \leq -\sup_{\{y : f_1(y) > f_0(y)\}} \frac{-f_0(y)}{f_1(y) - f_0(y)} \tag{2.2.16}$$

The lower bound on $t_1$ in (2.2.16) above is uninformative because it is negative, whereas we know that $t_1$ is greater than 1. Consider the upper bound. If

$$t_1 \leq \inf_{\{y : f_1(y) < f_0(y)\}} \frac{-f_0(y)}{f_1(y) - f_0(y)}$$

then

$$t_1 \leq \frac{1}{1 - \inf_y \frac{f_1(y)}{f_0(y)}} \equiv \bar{t}_1 \tag{2.2.17}$$
Note that the second infimum is taken over all \( y \), because the infimum of the density ratio is attained when \( f_1(y) < f_0(y) \).

Similarly, we have the following upper bound for \( t_0 \).

\[
t_0 \leq \frac{1}{1 - \inf_y \frac{f_0(y)}{f_1(y)}} \equiv \overline{t}_0
\]  

(2.2.18)

We can substitute \( t_1 \in [1, \overline{t}_1] \) and \( t_2 \in [1, \overline{t}_0] \) into (2.2.7) to get candidate density functions, \( \overline{f}_1(y) \) and \( \overline{f}_0(y) \). The functions obtained in this manner are valid pdfs; they are positive by construction, and integrate to 1 because they are affine combinations of the observed density functions, \( f_1(y) \) and \( f_0(y) \). This proves sharpness of the identified set.

\[\square\]

**Remark 2.2.1.**

i. The bounds from Theorem 1 lead to bounds on parameters of the distribution of \((Y, X^*)\). In particular we can bound the regression of \( Y \) on \( X^* \), which we consider in detail in a subsection. Bounds on such parameters are also considered in Aigner (1973), Bollinger (1996) and Henry, Kitamura, and Salanie (2010), among others.

ii. The lower bound for \((t_0, t_1)\) corresponds to the case of no misclassification. This was first shown in Aigner (1973). It illustrates the attenuation effect of non differential misclassification in a binary regressor.

iii. Henry, Kitamura, and Salanie (2010) consider the more general problem of nonparametric identification of finite mixture models, with an application to regression with a misclassified binary predictor. Their approach to identification is closest to the one
used here. They use non-negativity of the misclassification probabilities to obtain
bounds on two scalar parameters. They conclude that in the absence of additional
information, nothing can be learned about the misclassification probabilities. On the
other hand, we use non-negativity of the true conditional densities to obtain sharp
upper bounds on $t_0$ and $t_1$, and hence on $p$ and $q$.

iv. The upper bounds for $t_1$ and $t_0$ correspond to the maximum allowable misclassification
in either direction, given our assumptions. The maximum possible misclassification
occurs when the candidate densities $f_1^*(y)$ and $f_0^*(y)$ are exactly 0 in different regions
of $\mathcal{Y}$. An immediate implication is that we get point identification when we assume
that the support of $Y$ changes with of $X^*$.

v. The lower (upper) bound for $(t_0, t_1)$ does not correspond to the smallest (largest)
possible value for the marginal distribution of $X^*$, given in equation (2.2.12). The
lower bound for $P^*$ occurs with maximum possible misclassification of a success as a
reported failure and no misclassification in the other direction.

Theorem 1 shows that bounds on all parameters of interest can be calculated from two
real-valued functionals of the observed density ratio $f_1(y)/f_0(y)$. Define $r(y), \overline{r}, \underline{r}$ as follows:

$$r(y) = \frac{f_1(y)}{f_0(y)} \quad (2.2.19)$$

$$\overline{r} = \sup_y r(y) \quad (2.2.20)$$

$$\underline{r} = \inf_y r(y) \quad (2.2.21)$$

The supremum and infimum of the density ratio $r(y)$ correspond to $\overline{r}_0$ and $\overline{r}_1$ respectively.
Consequently, all parameters of interest can be calculated from $\tau$ and $r$. In the next section, we suggest an estimator for these parameters based on a density ratio estimate obtained by replacing the densities in (2.2.19) by their kernel estimates.

Let us now consider two examples that will help fix our ideas. First, note that the observed densities for $Y$ are mixtures of the two true conditional densities, $f^*_1(y)$ and $f^*_0(y)$. In the following examples, we consider respectively a location mixture and a variance mixture of two normals.

**Example 2.2.1.** Let $Y|X^* = 0 \sim N(1, 1)$ and $Y|X^* = 1 \sim N(3, 1)$. Let $p = q = 0.1$ and $P^* = 0.5$. Under Assumption 1, this is a complete description of the joint distribution of $(Y, X^*, X)$.

Figure 2.1a shows the different conditional densities of $Y$. The dotted lines are $f^*_0(y)$ and $f^*_1(y)$, the normal density functions with unit variance and means 1 and 3, respectively. The solid lines represent $f_0(y)$ and $f_1(y)$, which are mixtures of the two normals. $X^*$ is only a location shifter for $Y$, so $f^*_1$ dominates $f^*_0$ on the right tail, while the opposite is true on the left tail.

Misclassification is not too bad, so the same is true for $f_1$ and $f_0$, but to a lesser extent. Thus, the ratio $f_1(y)/f_0(y)$ approaches its supremum at the right tail and its infimum at the left tail. This is clear from Figure 2.2a which plots the ratio of the observed densities $f_1(y)/f_0(y)$.

We used numerical optimization in *Mathematica* to find the approximate sup and inf of
$r(y)$. We get the following upper bounds.

$$\bar{r} = 9, \quad \underline{r} = 0.111111, \quad \bar{t}_1 = \bar{t}_0 = 1.125.$$  

$$0 \leq p \leq 0.1, \quad 0 \leq q \leq 0.1, \quad 0.444444 \leq P^* \leq 0.555556$$

Note that the upper bounds for the misclassification errors are their true values. Thus, the upper (or outer) bounds for $f^*_1$ and $f^*_0$ are the density functions of $N(3,1)$ and $N(1,1)$ respectively.

In this example, the true parameter lies on the boundary of the identified set. In fact, this can be attributed to a characteristic of the density ratio of two normal distributions that differ only in their mean.

$$\frac{f^*_1(y)}{f^*_0(y)} \rightarrow 0 \quad \text{as} \quad y \rightarrow -\infty \quad \text{and} \quad \frac{f^*_0(y)}{f^*_1(y)} \rightarrow 0 \quad \text{as} \quad y \rightarrow \infty$$

This property is proved in Lemma 4 of Henry, Kitamura, and Salanie (2010) who use it as a strategy for point identification. Write the observed density ratio as

$$\frac{f_1(y)}{f_0(y)} = \frac{1 - P}{P} \frac{(1 - q)P^* f^*_1(y) + p(1 - P^*)f^*_0(y)}{qP^* f^*_1(y) + (1 - p)(1 - P^*)f^*_0(y)}$$

Using the fact that $P = 0.5$, we have

$$\frac{f_1(y)}{f_0(y)} \rightarrow \frac{p}{1 - p} \quad \text{as} \quad y \rightarrow -\infty \quad \text{and} \quad \frac{f_1(y)}{f_0(y)} \rightarrow \frac{1 - q}{q} \quad \text{as} \quad y \rightarrow \infty$$

In fact, in this example, we could have used the above relationship to calculate the sup and inf of the density ratio.
Figure 2.1: $Y \mid X^*$ is Normal with a location shift
Example 2.2.2. Let $Y|X^* = 0 \sim N(1, 1)$ and $Y|X^* = 1 \sim N(1, 2)$. As before, let $P^* = 0.5$, but increase the level of misclassification to $p = q = 0.2$. Under Assumption 1, this is a complete description of the joint distribution of $(Y, X^*, X)$.

As in the previous example, Figure 2.2a shows the different conditional densities of $Y$ and Figure 2.2b shows the observed density ratio. $X^*$ changes only the scale of $Y$, so $f_1^*$ dominates $f_0^*$ on both tails. The ratio $f_1(y)/f_0(y)$ approaches its supremum at the either tail and attains a global minimum at the mean $y = 1$.

Again using numerical optimization in Mathematica, we get the following bounds.

$$r = 4, \quad r = 0.666667, \quad t_1 = 3, \quad t_0 = 1.333333,$$

$$0 \leq p \leq 0.4, \quad 0 \leq q \leq 0.2, \quad 0.166667 \leq P^* \leq 0.625$$

The upper bound for $q$ is equal to its true value. $q$ is the probability of reporting a success as a failure, so the true conditional density $f_0^*$ is also on the boundary of the identified set.
Figure 2.2: $Y|X^*$ is Normal with a scale change
2.2.1 Regression of \( Y \) on \( X^* \)

We can use the bounds on \( t_0 \) and \( t_1 \) to obtain bounds on the regression of \( Y \) on \( X^* \) to get

\[
E[Y|X^* = 1] - E[Y|X^* = 0] = \int y(t_1 f_1(y) + (1 - t_1)f_0(y))dy - \int y((1 - t_0)f_1(y) + t_0 f_0(y))dy \tag{2.2.22}
\]

where \( t_1 \in [1, t_1] \) and \( t_0 \in [1, t_0] \).

The lower bound for \( t_1 + t_0 - 1 \) is 1 and this is, of course, the bound derived in Aigner (1973). The upper bound is smaller than the one in Bollinger (1996) as it uses more than just the information contained in the first and second moments of the observed distribution.

To see this, consider Model I from this paper.

\[
Y = \alpha + \beta X^* + u, \quad E[u|X^*] = 0, \quad \beta > 0,
\]

\[
P(X = 1|X^* = 2) + P(X = 2|X^* = 1) < 1.
\]

Let \( b \) be the slope from the least squares projection of \( Y \) on \( X \), \( d \) be the inverse of the slope from the least squares projection of \( X \) on \( Y \), and \( P_X \) be the marginal probability of observing \( X = 1 \). The bounds obtained in Bollinger (1996) are

\[
b \leq \beta \leq \max\{dP_X + b(1 - P_X), bP_X + d(1 - P_X)\}.
\]

Theorem 1 of this paper says that these bounds are tight relative to the information contained in the first and second moments of the data.

Suppose that in fact there is no misclassification. Then, if \( u \) is bounded and independent of \( X^* \), the support of \( Y \) changes with \( X^* \) (and \( X \)) in such a way that \( t_1 \) and \( t_0 \) are both 1.
In this case, $\beta$ is identified and equal to $b$. However, the upper bound from Bollinger (1996) is strictly greater because $b < d$.

In the following example, we consider cases where $u$ is normally distributed (therefore unbounded) and heteroscedastic in $X^*$.

Example 2.2.3. Let $Y|X^* = 0 \sim N(1,1)$ and $Y|X^* = 1 \sim N(1 + \beta, \sigma^2)$. As before, let the misclassification probabilities be $p = P(X = 1|X^* = 0)$ and $q = P(X = 0|X^* = 1)$, and the marginal probability of $X^* = 1$ be $P^*$. This gives us a complete description of the joint distribution of $Y, X$ and $X^*$.

The true value of the marginal effect of $X^*$ on $Y$ is $\beta$. In this example, we describe how the identified set for $\beta$ changes as we change the underlying parameters. The lower bound for $\beta$ is well-known to be the slope of the projection of $Y$ on $X$ and is the same as the ones derived in Aigner (1973) and Bollinger (1996). We show that it changes with the true value of $\beta$, with the marginal distribution of $X^*$ and with the misclassification probabilities. It does not change when we change $\sigma^2$, which affects the variance of $Y$ but not the covariance between $X$ and $Y$.

In addition to $b$, the upper bound for $\beta$ derived in this paper depends on the ratio of the densities of $Y$ conditional on the two values of $X$, at its minimum and maximum. We find that the upper bound is only sensitive to changes in parameters that affect the conditional distribution of $Y$ given $X^*$, namely $\beta$ and $\sigma$. The upper bound does not change with any changes to the joint distribution of $X$ and $X^*$ that leave $f(Y|X^*)$ unchanged.

We also calculate the upper bound derived by Bollinger (1996) which depends on the
information contained in the first two moments of the data. We find that it is much larger than the upper bound described above and is sensitive to all parameter values.

Figure 2.3 shows how the identified set for $\beta$ changes as a function of its true value. The other parameters are set at $\sigma = 2$, $P_{X^*} = 0.3$, $p = 0.1$ and $q = 0.2$. Note that the variance of $Y$ changes with $X^*$ and we have heteroscedasticity.

The first three rows of the top panel give the marginal and conditional densities of $Y$ for $\beta = 2, 4$ and 6. As $\beta$ increases, $f(y)$ and $f(y|X)$ appear more bimodal. The last row gives the function of the ratio $f(y|1)/f(y|2)$ that we use to calculate $\hat{t}_1$ and $\hat{t}_0$.

In the bottom panel, we see that both the upper and the lower bounds increase as $\beta$ increases. Although it is not immediately clear from the figure, the width of the identified set decreases relative to $\beta$. Moreover, when $\beta = 0$, the upper bound is not defined. For large values of $\beta$, the upper bound coincides with the true value. This is a peculiarity of the normal distribution and was also the case to a lesser extent when we considered the t-distribution. In this case, the true value of $\beta$ as well as the other parameters lies on the boundary of the identified set.

Figure 2.4 shows how the identified set for $\beta$ changes as a function of $\sigma$. The other parameters are set at $\beta = 2$, $P_{X^*} = 0.3$, $p = 0.1$ and $q = 0.2$. As before, the top panel shows the various densities of $Y$ and also a function of the observed density ratio. The bottom panel gives the upper and lower bounds for $\beta$ when $\sigma$ is in $\{1, 2, 3, 4, 5\}$.

When $\sigma = 1$, the variance of $Y$ is constant in $X^*$ but changes with $X$. In this case, our upper bound coincides with the true value of $\beta$. Again, this is a peculiarity of the
normal distribution and was not observed in the t-distribution case. As heteroscedasticity ‘increases’, we find that the upper bound initially increases and then decreases, while the upper bound from Bollinger (1996) increases sharply. On the other hand, the lower bound does not change with $\sigma$, as the latter does not affect the covariance of $Y$ and $X$.

Figures 2.5 and 2.6 show how the identified set for $\beta$ changes as we change the distribution of $X$ and $X^*$. The upper bound is unaffected by changes in either the marginal distribution of $X^*$ or the misclassification probabilities. The same is true when we use the t-distribution. The upper bound of Bollinger (1996) and the lower bound perform better when the two values of $X^*$ are close to equiprobable, and when the misclassification is not severe. The former gets rapidly worse as one value of $X^*$ becomes less likely or as $X$ and $X^*$ become close to independent.

**Remark 2.2.2.**

i. The lower bound from Bollinger and from this paper coincide, and are the same as the one first derived by Aigner (1973).

ii. Bollinger’s upper bound is based on the information contained in the first two moments of the observed distribution, while the upper bound in this paper draws from the entire distribution. Mixture distributions typically have higher moments. This accounts for the large difference in the two upper bounds. In the above example, the observed distributions are mixtures of normal distributions and so have higher moments.
Figure 2.3: Identified set for $\beta$ as a function of its true value
(a) Densities of \( Y \) for \( \sigma \in \{1, 2, 4\} \)

<table>
<thead>
<tr>
<th>( \sigma )</th>
<th>&quot;Lower bound&quot;</th>
<th>&quot;Upper bound&quot;</th>
<th>&quot;Bollinger (1996)&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.21538</td>
<td>2.</td>
<td>4.19082</td>
</tr>
<tr>
<td>2</td>
<td>1.21538</td>
<td>2.69073</td>
<td>8.40511</td>
</tr>
<tr>
<td>3</td>
<td>1.21538</td>
<td>2.70124</td>
<td>15.4289</td>
</tr>
<tr>
<td>4</td>
<td>1.21538</td>
<td>2.56014</td>
<td>25.2623</td>
</tr>
<tr>
<td>5</td>
<td>1.21538</td>
<td>2.45101</td>
<td>37.9051</td>
</tr>
</tbody>
</table>

(b) Identified set for \( \beta \) with \( \sigma \in \{1, 2, 3, 4, 5\} \)

Figure 2.4: Identified set for \( \beta \) as a function of \( \sigma \)
Figure 2.5: Identified set for $\beta$ as a function of the marginal distribution of $X^*$
Figure 2.6: Identified set for $\beta$ as a function of the misclassification probabilities
2.3 Estimation

In this section, we suggest simple estimators for $r$ and $\tau$. We separately obtain kernel density estimates for $f_1(y)$ and $f_0(y)$, and take the infimum and supremum of the ratio of these estimates on a subset of $\mathcal{R}$ where they are bounded away from zero.

Let $(y_i, X_i)$ be an iid sample for $i = 1, \ldots, n$. Let $n_1$ be the number of observations with $X_i = 1$, and $n_0 = n - n_1$. The kernel density estimates of $f_1(y)$ and $f_0(y)$ are

$$
\hat{f}_1(y) = \left( n_1 h_1 \right)^{-1} \sum_{i=1}^{n_1} K \left( h_1^{-1} (y - y_i) \right)
$$

(2.3.1)

$$
\hat{f}_0(y) = \left( n_0 h_0 \right)^{-1} \sum_{i=1}^{n_0} K \left( h_0^{-1} (y - y_i) \right)
$$

(2.3.2)

Here $K(\cdot)$ is a suitable kernel function and $h_0$ and $h_1$ are smoothing parameters that go to 0 as $n_1$ and $n_0$ respectively go to infinity.

Let

$$
\hat{r}(y) = \frac{\hat{f}_1(y)}{\hat{f}_0(y)} \mathbf{1} \left[ \hat{f}_1(y) > b, \hat{f}_0(y) > b \right]
$$

(2.3.3)

We propose the estimator

$$
\hat{\tau} = \sup_{\{y: \hat{f}_1(y) > b, \hat{f}_0(y) > b\}} \hat{r}(y)
$$

(2.3.4)

$$
\hat{\tau} = \inf_{\{y: \hat{f}_1(y) > b, \hat{f}_0(y) > b\}} \hat{r}(y)
$$

(2.3.5)

where the infimum and supremum are taken over a “trimming” set. $b$ is a trimming scalar that goes to 0 as $n$ goes to infinity. The use of the same trimming scalar for both $\hat{f}_1(y)$ and $\hat{f}_0(y)$ is without loss of generality. We need to trim both densities because the density ratio
is poorly estimated when either density is small, and because we are ultimately interested in the supremum and infimum of this ratio.

Let \( \hat{B} \) and \( B \) be the trimming sets based on the kernel density estimates and the true densities, respectively.

\[
\hat{B} := \{ y : \hat{f}_1(y) > b, \hat{f}_0(y) > b \} \tag{2.3.6}
\]

\[
B := \{ y : f_1(y) > b, f_0(y) > b \} \tag{2.3.7}
\]

Then \( \hat{r} \) and \( \check{r} \) are the supremum and infimum of the ratio of density estimates over \( \hat{B} \).

Let

\[
\bar{r} = \sup_{y \in \hat{B}} \hat{r}(y) \tag{2.3.8}
\]

\[
\check{r} = \sup_{y \in B} r(y) \tag{2.3.9}
\]

and similarly define \( \check{r} \) and \( \bar{r} \).

We will show that \( \hat{r}(y) \) is a uniformly consistent estimator of \( r(y) \) when we only consider the trimming set \( B \). The technique used here follows Kelsall and Diggle (1995), who show consistency of the log of the density ratio.

By construction, \( f_1 \) and \( f_0 \) are bounded away from 0 on the trimming set \( B \). Define the relative error in \( \hat{f}_1 \) as

\[
\nu_1(y) = \frac{\hat{f}_1(y) - f_1(y)}{f_1(y)} \implies \hat{f}_1(y) = f_1(y)(1 + \nu_1(y)) \tag{2.3.10}
\]

for \( y \in \hat{B} \). Similarly define \( \nu_0 \) as the relative error in \( \hat{f}_0 \).

\[
\log (\hat{r}(y)) = \log (f_1(y)) + \log (1 + \nu_1(y)) - [\log (f_0(y)) + \log (1 + \nu_0(y))] \tag{2.3.11}
\]

\[
= \log (r(y)) + \nu_1(y) - \nu_0(y) + O(\nu^2) \tag{2.3.12}
\]
where the last equality follows from a first-order Taylor expansion, and \( \nu \) denotes either \( \nu_1 \) or \( \nu_0 \). It follows that

\[
\sup_{y \in B} |\log(\hat{r}(y)) - \log(r(y))| \leq \sup_{y \in B} |\nu_1(y)| + \sup_{y \in B} |\nu_0(y)| + O(\nu^2) \tag{2.3.13}
\]

With appropriate choice of kernel function and bandwidth parameters, and under the assumption that \( f_1(y) \) and \( f_0(y) \) are uniformly continuous, we have uniform consistency of \( \hat{f}_1(y) \) and \( \hat{f}_0(y) \). Then the relative errors in the density estimates are uniformly small and \( \log(\hat{r}) \) is uniformly consistent on \( B \). This in turn implies that \( \hat{r}(y) \) is uniformly consistent on \( B \).

This proves that the infeasible estimator \( \tilde{r} \) with trimming based on true densities is a consistent estimator for the trimmed parameter \( \tilde{\alpha} \), and similarly for \( \tilde{\beta} \) and \( \tilde{\gamma} \). This does not prove consistency of the proposed estimator. It remains to be shown that we have convergence to the trimmed parameter when trimming is based on kernel density estimates, and that the trimmed parameter itself converges to the supremum (or infimum) of the density ratio over the entire support of \( Y \).

### 2.4 Simulation Study

We perform a Monte Carlo analysis of the proposed estimator for the upper bound under the following specification.

\[
Y | X^* = 0 \sim N(1, 1), \quad Y | X^* = 1 \sim N(4, 2)
\]

\[
\Pr(X^* = 1) = 0.5, \quad p = q = 0.1
\]

\( X^* \) shifts the location and changes the scale of \( Y \), with the true value of \( \beta \) at 3. The two values of \( X^* \) are equally likely, which is not restrictive because the upper bound does not
depend on the marginal distribution of $X^*$. Misclassification is moderate at 10% in each
direction. The true upper bounds for $t_0$, $t_1$ and $\beta$ are

$$
\overline{t}_0 = 1.06, \quad \overline{t}_1 = 1.44, \quad \overline{\beta} = 3.38
$$

The density ratio is minimized close to 0 and maximized for very large $y$. So we expect
that $\overline{t}_1$ will be very well estimated, while $\overline{t}_0$ will be poorly estimated.

We generate 1000 samples with sample size $n = 10000$. We estimate the conditional
density of $Y$ given $X = 1$ and $X = 2$ separately in the two subsamples. As proposed, we
use the kernel density estimator with a Gaussian kernel and an oversmoothing bandwidth
parameter. The ratio of the estimated densities is evaluated for $y \in \{-5, 10\}$ with a step
size of 0.001. We then calculate upper bounds for $t_0$, $t_1$ and $\beta$ for several values of the
trimming parameter $b$, by discarding the grid points with estimated density less than $b$.

Table 2.1 contains the sampling statistics and error magnitudes for the three parameters
under different levels of trimming. The first block of results corresponds to no trimming.
The estimator of $\overline{t}_0$ has essentially no variance in this case. It is always 1, which would
correspond to the case of no misclassification when $X^* = 1$. In the absence of trimming, the
estimated density ratio attains its maximum value when $\hat{f}_0$ is 0 and $\hat{f}_1$ is some small positive
number. This illustrates the importance of trimming when the ratio $f_1/f_0$ approaches its
sup or inf in the tails of the observed distribution.

If we consider only the error magnitudes, the estimator for $\overline{t}_0$ seems to perform best
with the trimming parameter set to 0.01 in the current specification. As we discard more
of the tails of the estimated densities, it is biased upwards. This also leads to wider bounds
Table 2.1: Simulation Results for Location-Scale Mixture

<table>
<thead>
<tr>
<th>Trimming</th>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>Q₁</th>
<th>Median</th>
<th>Q₃</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>b = 0</td>
<td>$\overline{t}_0$</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\overline{t}_1$</td>
<td>1.39</td>
<td>0.10</td>
<td>1.44</td>
<td>1.42</td>
<td>0.11</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>3.13</td>
<td>0.22</td>
<td>3.14</td>
<td>3.19</td>
<td>3.24</td>
<td>0.33</td>
<td>0.24</td>
</tr>
<tr>
<td>b = 0.01</td>
<td>$\overline{t}_0$</td>
<td>1.07</td>
<td>0.00</td>
<td>1.07</td>
<td>1.07</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\overline{t}_1$</td>
<td>1.43</td>
<td>0.03</td>
<td>1.45</td>
<td>1.42</td>
<td>1.00</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>3.36</td>
<td>0.06</td>
<td>3.32</td>
<td>3.36</td>
<td>3.40</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>b = 0.02</td>
<td>$\overline{t}_0$</td>
<td>1.15</td>
<td>0.01</td>
<td>1.14</td>
<td>1.15</td>
<td>1.15</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$\overline{t}_1$</td>
<td>1.43</td>
<td>0.03</td>
<td>1.45</td>
<td>1.43</td>
<td>1.00</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>3.53</td>
<td>0.07</td>
<td>3.48</td>
<td>3.53</td>
<td>3.58</td>
<td>0.17</td>
<td>0.15</td>
</tr>
<tr>
<td>b = 0.05</td>
<td>$\overline{t}_0$</td>
<td>1.49</td>
<td>0.03</td>
<td>1.47</td>
<td>1.49</td>
<td>1.50</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>$\overline{t}_1$</td>
<td>1.43</td>
<td>0.03</td>
<td>1.45</td>
<td>1.43</td>
<td>1.00</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>4.30</td>
<td>0.10</td>
<td>4.24</td>
<td>4.30</td>
<td>4.38</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>b = 0.1</td>
<td>$\overline{t}_0$</td>
<td>3.08</td>
<td>0.22</td>
<td>2.92</td>
<td>3.05</td>
<td>3.22</td>
<td>2.03</td>
<td>2.02</td>
</tr>
<tr>
<td></td>
<td>$\overline{t}_1$</td>
<td>1.45</td>
<td>0.03</td>
<td>1.46</td>
<td>1.45</td>
<td>1.00</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>7.93</td>
<td>0.55</td>
<td>7.54</td>
<td>7.88</td>
<td>8.29</td>
<td>4.59</td>
<td>4.55</td>
</tr>
</tbody>
</table>

for the regression coefficient $\beta$. But we should keep in mind that the trimming is binding in this case, i.e., the ratio attains its maximum value when $\hat{f}_0$ is exactly $b$ and any sampling variation is coming only from the estimation of $\hat{f}_0$.

By contrast, the estimator of $\overline{t}_1$ performs quite well for all positive levels of trimming. As noted earlier, $\overline{t}_1$ corresponds to the minimum of the density ratio which is attained at about 0, only a standard deviation away from the mean of $N(1,1)$. 63
Chapter 3

Measurement Error in Finite-Valued Explanatory Variables: Identification with Additional Information

3.1 Introduction

In this short chapter, we consider the identifying power of additional information in the form of a third observed discrete random variable $V$. All observed variables are assumed to be mutually independent conditional on the unobserved variable $X^*$. Recently, Mahajan (2006), Lewbel (2007), and Hu (2008) obtained point identification of the parameters of interest in models with a misclassified discrete regressor. All three papers use the existence of an additional random variable that acts as an ‘instrument.’ See the survey by Chen, Hong, and Nekipelov (2011) for details. Our goal in this chapter is to gain some insight into the mechanism by which the availability of an instrument-like variable results in point identification.

Let us add to our set-up a discrete random variable $V$ that takes values in $\{1,\ldots,K\}$. 
Let us assume the following conditional independence structure:

\[ Y \perp (X, V) \mid X^* \quad \text{and} \quad X \perp V \mid X^* \quad (3.1.1) \]

These are simplified versions of the key assumptions used in Mahajan (2006) and Hu (2008), who allow \( Y \) to be continuous and also admit other regressors that are allowed to be related to the misclassification process.

It is important that \( V \) and \( X \) are independent conditional on \( X^* \). In the absence of this assumption, we are back in the setting of Chapter 1 with a new mismeasured regressor \( \tilde{X} = (X, V) \) taking values in the support of their joint distribution. As we saw before, \( \tilde{X} \) can contain additional information about the underlying distribution that will allow us to shrink the identified set, but this is typically not enough to obtain point identification.

The above conditional independence assumptions are equivalent to the following

\[
P(Y = i, X = j, V = k) = \sum_{r=1}^{R} P(Y = i \mid X^* = r) P(X = j \mid X^* = r) P(V = k, X^* = r)
\]

\[
\pi_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr}
\quad (3.1.2)
\]

If we think of \( \pi_{ijk} \) as an element of a three-dimensional array \( \Pi \), then we are interested in decomposing this array into three component matrices \( A, B \) and \( C \). In particular, we want to know the conditions under which such a decomposition is unique, because then we would have point identification of all unknown probabilities of interest.

Decomposition of higher-dimensional arrays or tensors have been studied in linear algebra for several decades, and have applications in signal processing, psychometrics, and chemometrics, to name just a few. See Kolda and Bader (2009) for a comprehensive survey.

The interesting fact is that while the matrix factorization considered in Chapter 1 is
seldom unique, tensor decomposition is unique under relatively mild conditions on the
distribution of each observed variable given the latent \( X^* \).

In the next section, we state some existing results from the tensor decomposition litera-
ture. In Section 3.3, we apply these results to probability arrays to get point identification
of the unobserved probabilities \( \text{Pr}(Y|X^*), \text{Pr}(X|X^*) \) and \( \text{Pr}(V|X^*) \).

### 3.2 Tensor Decomposition

A tensor of \( N \) dimensions or an \( N \)-way array, \( \Pi \), is an object with the typical element
\( \pi_{i_1 i_2 \ldots i_N} \in \mathcal{R} \) subscripted by \( N \) indices. For instance, a vector is a one-way array and a
matrix is a two-way array. We will be dealing with matrices and three-way arrays. As
before, we denote vectors by bold-face lower case letters (\( a, b, \ldots \)) and matrices by bold-
face upper-case letters (\( A, B, \Pi \ldots \)). We also write three-way arrays as bold-face upper-case
letters. Elements of arrays are denoted by the corresponding lower-case letters with the
appropriate indices. For instance, \( a_{ij} \) is the \((i,j)\)th element of the matrix \( A \).

It is well known that an \( I \times J \) matrix \( \Pi \) of rank \( R \) can be factorized as

\[
\Pi = AB^T
\]  

(3.2.1)

where \( A \) and \( B \) are respectively \( I \times R \) and \( J \times R \) matrices with full column rank. Let \( a_r \)
be the \( r \)-th column of \( A \) and \( b_r \) be the \( r \)-th column of \( B \). Then

\[
\Pi = \sum_{r=1}^{R} a_r b_r^T = \sum_{r=1}^{R} a_r \otimes b_r
\]  

(3.2.2)

where \( \otimes \) means the outer product or tensor product of two arrays.\(^1\) Note that each term in

\(^1\)Let \( A \) be an \( N \)-way array of dimension \( I_1 \times I_2 \times \ldots \times I_N \) and \( B \) be an \( M \)-way array of dimension
the summation is an $I \times J$ matrix of rank one given by the outer product of corresponding columns of the matrices $A$ and $B$.

We would like to extend the concept of matrix factorization to a three-way array. For this, it is useful to express the three-way array as the sum of rank-one three-way arrays defined below.

**Definition 3.2.1.** Let $U$ be a three-way array of dimensions $I \times J \times K$. We say that $U$ is a **rank-one array** if it can be written as the outer product of three vectors $a, b$ and $c$.

**Example 3.2.1.** The joint distribution of three mutually independent binary random variables can be represented by a rank-one array of dimensions $2 \times 2 \times 2$.

**Definition 3.2.2.** The **rank** of a three-way array $\Pi$ of dimensions $I \times J \times K$ is the minimal $R$ for which

$$
\Pi = \sum_{r=1}^{R} U_r = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r
$$

(3.2.3)

where the $U_r$ are rank-one $I \times J \times K$ arrays and $a_r \in \mathcal{R}^I$, $b_r \in \mathcal{R}^J$ and $c_r \in \mathcal{R}^K$

The model in (3.2.3) was introduced independently in Carroll and Chang (1970) who called it the Canonical Decomposition model and in Harshman (1970) who called it Parallel Factor analysis. It is known in the literature and shall be referred to as the CP model.

Note that the component matrices must have the same number of columns $R$, just as the factoring matrices in the ordinary matrix product (3.2.1) have the same number of columns.

Note that the CP decomposition always exists for real-valued tensors. We can always write $\Pi$ as the sum of $R = I \times J \times K$ rank-one arrays with only one non-zero entry equal $J_1 \times J_2 \times \ldots \times J_N$. Then $A \otimes B$ is the $(N + M)$-way array obtained by multiplying every element of $A$ by $B$. It has dimension $I_1 \times I_2 \times \ldots \times I_N \times J_1 \times J_2 \times \ldots \times J_N$. 

67
to the corresponding element of \( \Pi \).

However, we are more interested in the uniqueness properties of the CP decomposition rather than its existence. So, for the purposes of this paper, we will assume that for a given three-way array \( \Pi \) the rank \( R \) is known or equivalently there exists a CP decomposition \((A, B, C)\) involving \( R \) columns.

Observe that if we apply a common permutation to the columns of the component matrices \( A, B, \) and \( C \), the CP decomposition (3.2.3) still holds for the new matrices. This is also true if we obtain the new matrices by rescaling the columns in such a way that their outer product remains unchanged.

**Definition 3.2.3.** We say that \((A, B, C)\) is equivalent to \((\bar{A}, \bar{B}, \bar{C})\) if there exist a unique permutation matrix \( P \) and non-singular diagonal matrices \( \Lambda, M \) and \( N \) with \( \Lambda M N = I_R \) such that

\[
\bar{A} = AP\Lambda, \quad \bar{B} = APM \quad \bar{C} = CPN
\]

As a first step to uniqueness, we would like to know when two decompositions of a given three-way array are equivalent. Kruskal (1977) provides several sufficient conditions for this equivalence. In much later work, Sidiropoulos and Bro (2000) extend this result to higher-order arrays.

**Definition 3.2.4.** The k-rank (named after Joseph Kruskal) \( k_A \) of an \( I \times R \) matrix \( A \) is the largest integer such that every collection of \( k_A \) columns drawn from \( A \) is linearly independent.

Compare this object to the rank of a matrix \( r_A \) which is the largest integer such that \( A \)....
contains at least one collection of \( r_A \) linearly independent columns. Note that \( k_A \leq r_A \leq \min(I, R) \).

**Lemma 3.2.1.** (a)(Kruskal 1977) Let \( \Pi \) be an \( I \times J \times K \) three-way array of rank \( R \). Suppose there exist matrices \((A, B, C)\) and \((\bar{A}, \bar{B}, \bar{C})\) such that

\[
\pi_{ijk} = \sum_{r=1}^{R} a_{ir}b_{ir}c_{ir} = \sum_{r=1}^{R} \bar{a}_{ir}\bar{b}_{ir}\bar{c}_{ir}
\]

Suppose the following condition holds

\[
k_A + k_B + k_C \geq 2R + 2.
\] (3.2.4)

Then \((A, B, C)\) is equivalent to \((\bar{A}, \bar{B}, \bar{C})\).

(b)(Ten Berge and Sidiropoulos 2002) Further, for \( R = 2 \) and \( R = 3 \), the sufficient condition (3.2.4) is also necessary.

The \( k \)-rank condition (3.2.4) is particularly easy to interpret when \( R = 2 \). The only way that it can hold is if \( k_A = r_A = 2 \) and similarly for \( B \) and \( C \). In other words, when \( R = 2 \) the CP decomposition \((A, B, C)\) is unique up to common permutation and scaling of columns if and only if each component matrix has full column rank.

### 3.3 Implications of Uniqueness of Tensor Decomposition for Probability Arrays

Recall that we use a three-way array \( \Pi \) of dimensions \( \dim(Y) \times \dim(X) \times \dim(V) \) to represent the distribution of \((Y, X, V)\). If our assumptions of conditional independence are true then there exists a CP decomposition \((A, B, C)\) of \( \Pi \) involving \( R \) columns where
\( R = \dim(X^*) \) with typical elements

\[
a_{ir} = P(Y = i|X^* = r), \quad b_{jr} = P(X = j|X^* = r), \quad c_{kr} = P(V = k, X^* = r). \tag{3.3.1}
\]

Then \( A \) and \( B \) are stochastic matrices whose columns add up to 1. We will refer to the CP decomposition of a three-way array of probabilities into three probability matrices as stochastic factorization.

Let the columns of the component matrices be denoted by the corresponding letter with the appropriate subscript. For instance, \( a_r \) is the \( r \)-th column of \( A \).

**Theorem 3.3.1.** Consider an \( I \times J \times K \) array \( \Pi \) such that \( \sum_i \sum_j \sum_k \pi_{ijk} = 1 \). Suppose there is a stochastic factorization \((A, B, C)\) of \( \Pi \), i.e. we can write

\[
\pi_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{ir}
\]

where for every \( r \) we have \( \sum_i a_{ir} = \sum_j b_{jr} = 1 \).

Suppose the \( k \)-rank condition (3.2.4) holds. Then \((A, B, C)\) is the unique stochastic factorization of \( \Pi \) up to a common permutation of columns.

**Proof.** Let \((\tilde{A}, \tilde{B}, \tilde{C})\) be another stochastic factorization of \( \Pi \). From Lemma 3.2.1 we know that \((A, B, C)\) and \((\tilde{A}, \tilde{B}, \tilde{C})\) are equivalent, i.e. the \( r \)-th column of \( \tilde{A} \) is the multiple of some column of \( A \)

\[
\tilde{a}_r = \lambda_r a_{p(r)}.
\]

Since both columns sum up to 1, we must have \( \lambda_r = 1 \) for every \( r \) and \( \Lambda = I_R \). Similarly \( M = I_R \). Finally \( \Lambda M N = I_R \) gives us \( N = I_R \). \( \square \)
Therefore under the conditions of Theorem 3.3.1 the parameters of interest in (3.3.1) are identified up to a ‘switching’ of labels. Unlike the matrix case, the identified set here is finite, and we can get point identification by adding an assumption like monotonicity or information about the sign of the relationship between \( Y \) and \( X^* \).

We should note that the sufficiency of Lemma 3.2.1 when \( R = 2, 3 \) does not translate to stochastic factorizations. Uniqueness of stochastic factorization does not imply uniqueness of CP decomposition. Given a unique stochastic factorization \((A, B, C)\), we could find \((\bar{A}, \bar{B}, \bar{C})\) that are not probability arrays and are not equivalent to \((A, B, C)\). The author’s conjecture is that this would occur for \( \Pi \) that are on the boundary of the set of probability arrays.

Theorem 3.3.1 has useful interpretations when some or all of the variables are binary.

**Corollary 3.3.1.** Suppose \( R = 2 \) i.e., \( X^* \) is binary and takes values in \( \{1, 2\} \). Then \((A, B, C)\) are unique up to permutation of columns if

i. \( Y \) is not independent of \( X^* \), i.e., \( P(Y = i|X^* = 1) \neq P(Y = i|X^* = 2) \) for some \( i \),

ii. \( X \) is not independent of \( X^* \), i.e., \( P(X = j|X^* = 1) \neq P(X = j|X^* = 2) \) for some \( i \),

iii. \( V \) is not independent of \( X^* \), i.e., \( P(V = k, X^* = 1) \neq P(V = k', X^* = 1) \) for some \( k, k' \).

**Proof.** A has full column rank if and only if \( a_1 \) and \( a_2 \) are linearly independent. Since \( a_1 \) and \( a_2 \) both add up to one, they both lie in the \( I - 1 \) dimensional simplex. The only way they can be linearly dependent is if they are equal. A similar argument proves ii and iii. \( \square \)
The conditions obtained from Corollary 3.3.1 match up nicely with the assumptions used by Mahajan (2006) to get point identification in the regression model with a binary misclassified regressor. Condition (i) above is required for the identification of the misclassification probabilities, but not for the conditional distribution of $Y$ given $X$. If it does not hold, then $P(Y|X^*) = P(Y|X) = P(Y)$. Condition (ii) can be strengthened to obtain point identification. For instance, when $X$ is binary, we can assume that the misclassification is not bad enough to reverse labels, i.e., $P(X = 1|X^* = 1) > P(X = 2|X^* = 2)$. Condition (iii) requires that the distribution of $X^*$ should vary with the additional variable $V$, i.e., $V$ contains information about $X^*$.

The k-rank condition is harder to interpret for probability matrices with $R \geq 3$. However we can still gain some insights from it. An immediate implication is that each of $k_A, k_B$ and $k_C$ must be greater than or equal to 2. This means that the rank of each of these matrices is at least 2, or that none of the observed variables can be independent of $X^*$.

Like the rank, the k-rank is a measure of the dependence between the observed variable and $X^*$. The k-rank condition puts a lower bound on this dependence. It is additive in the individual k-ranks, which means that we can “trade” the dependence on $X^*$ among the observed variables. This becomes noticeable as $R$ increases and the full-rank assumptions we needed to make on the component matrices can be relaxed. For instance, when $R = 3$ and $I, J, K \geq 3$, the lower bound on the sum of k-ranks is less than 9, so that one of the matrices need not have full column rank. One implication of this is that the observed regressor $X$, could potentially take fewer values than the true regressor $X^*$. This relaxes
the assumptions in Hu (2008) to some extent.

In summary, the k-rank condition requires that each observed variable should have relevant information about $X^*$ that is independent of the information contained in the other two variables.

A special case of the k-rank condition is when $Y$ is binary, $X$ and $V$ each take one of $R$ values, and the probability matrices $A, B$ and $C$ have full rank. This is essentially the binary choice model with a misclassified regressor. We can allow for other regressors $Z$ in this model by requiring stochastic factorization of the arrays $\Pi_Z = \Pr(Y = i, X = j, V = k|Z)$ obtained by conditioning on $Z$. As noted in Mahajan (2003), this does allow for the misclassification process to be correlated with the additional regressors $Z$.

3.4 Conclusion.

In this chapter, we continued with our formulation of nondifferential measurement error in a discrete latent variable as a problem of factorizing the conditional probability distribution of the observed variables. Adding an additional measurement or instrumental variable to this model leads to a problem of factorizing a three-way array.

We have shown that the point identification obtained in this case is related to a remarkable property of higher-dimensional arrays - that tensor decomposition is essentially unique under a relatively mild condition on the component matrices. This condition can be interpreted in terms of restrictions on the distributions of the observed variables conditional on the latent variable. Essentially, it requires that each observed variable should contain relevant information about the latent variable.
Chapter 4

Identifying the Number of Components in a Finite Mixture Model

4.1 Introduction

Mixture models are common in statistics. The simplest mixture models arise when data are sampled from a population that consists of a finite number of subpopulations, known as components, and membership in the components is not observed. The observed distribution of the data is then a mixing of the distributions within the components.

For a multivariate random variable \( X \) we have

\[
F(x) = \sum_{r=1}^{R} \pi_r F_r(x), \quad x \in \mathcal{X} \subseteq \mathbb{R}^K; \quad \pi_r \geq 0, \quad \sum_{r=1}^{R} \pi_r = 1 \tag{4.1.1}
\]

where \( F \) is the observed distribution, \( F_r \) is the distribution of \( X \) in the \( r \)-th component and \( \pi_r \) is the proportion of the population that belongs to the \( r \)-th component. We can think of \( \pi \) as the distribution of a discrete latent variable \( X^* \) that determines component membership. \( F_r \) is then the distribution of \( X \) conditional on \( X^* = r \). When the number of components \( R \) is known, the model in (4.1.1) is called an \( R \)-component mixture model.
with parameters $\{\pi_r, F_r\}, r = 1, \ldots R)$. When $R$ is unknown, an additional parameter of interest is the minimal $R$ for which (4.1.1) is true.

Several statistical models including the random effects model, the latent class and trait models, and measurement error models can be written as mixtures. Lindsay (1995) contains a useful discussion on the universality of mixture models.

The number of components $R$ may sometimes be dictated by theory. In a measurement error model where $X$ consists of $K$ repeated measurements of a finite-valued variable of interest $X^*$, we can think of each $X_k$ and $X^*$ as taking the same number of values. On the other hand, in a random effects model of unobserved heterogeneity, we think of the values of $X^*$ as the type of the individual. In this case, $R$ is the possible number of types and may be unknown.

The component distributions $F_r$ are often specified as parametric distributions. This is sometimes plausible, as in the simple example of Lindsay (1995) where $X$, the length of an animal, is modeled as a mixture of two normals with components corresponding to male and female, respectively. In other applications, it is desirable to leave the form of $F_r$ unspecified. Cruz-Medina, Hettmansperger, and Thomas (2004) model the time required to solve a cognitive task as a finite mixture with components corresponding to different possible solution strategies. They note that theory does not suggest a particular shape for the distribution of reaction time within a component. Moreover, when the number of components is unknown, Kasahara and Shimotsu (2010) point out that different specifications for the component distributions lead to different estimates of the number of components.
When the component distributions are left unspecified, it is commonly assumed that they are product distributions. Partition $X$ into $K_z \leq K$ groups given by $Z_1, \ldots, Z_{K_z}$. Let each $Z_k$ be a $G_k$-variate random variable.

$$F_r(x) = \prod_{k=1}^{K} F_r^k(z_k), \quad z_k \in Z_k \subseteq \mathbb{R}^{G_k}$$  \hfill (4.1.2)

Equation (4.1.3) says that $Z_1, \ldots, Z_{K_z}$ are independent conditional on belonging to a component. When $X_1, \ldots, X_{K}$ are independent conditional on component membership, we have

$$F_r(x) = \prod_{k=1}^{K} F_r^k(x_k), \quad x_k \in X_k \subseteq \mathbb{R}$$  \hfill (4.1.3)

Hettmansperger and Thomas (2000) use this assumption in the context of repeated measurements and refer to it as *within-individual independence*. Such an assumption is also common in measurement error models where it is called *non-differential measurement error*. (Carroll, Ruppert, and Stefanski 1995, Mahajan 2006, Hu 2008). It is used in a random effects setting by Hall and Zhou (2003).

The conditional independence assumption, although strong, is common in practice, and the list above is by no means comprehensive. The popularity of this assumption is due to the fact that it leads to nonparametric point identification of the marginal distributions $F_r^k$ and the mixing probabilities $\pi_r$ under relatively mild sufficient conditions.

Hall and Zhou (2003) discuss its validity in the following scenario. Let $X_1, \ldots, X_{K}$ be the results from a series of tests applied to an individual. Consider a causal framework in which these responses are determined by a latent variable $\Lambda$ which contains comprehensive information about the health of the individual. In this case, it is reasonable to assume
that the $X_k$’s are independent given a particular value of $\Lambda$. If $\Lambda$ is finite, we can set the component membership indicator $X^* = \Lambda$ and we have a finite mixture of products of $K$ distributions. Thus the equations (4.1.1) and (4.1.3) are a statement of our belief in this latent structure - the existence of a latent variable and its finiteness.

In the econometrics literature, Henry, Kitamura, and Salanie (2010) consider the identification of finite mixtures of products and give three classes of econometric models that can be interpreted in this manner. First, they consider models with a misclassified binary regressor. They note that the conditional independence assumption is used in much of the literature,\footnote{Indeed, this model is extensively discussed in the preceding chapters of this dissertation.} and is a reasonable assumption when “errors in regressor values are due to clerical mistakes”. Second, they consider a Markov switching model in which $X_t$ are indexed by time and are iid conditional on a binary state variable. Finally, they consider a model with unobserved heterogeneity, where agents can be one of two latent types.

In this paper, we address the problem of identifying the minimal number of components that is consistent with the distribution of the data. We are concerned then with the existence of an $R$-component mixture model rather than with its uniqueness.

A motivating example is a model with unobserved heterogeneity and an unknown number of types. Let

$$Y = G(X, W, \alpha, \epsilon)$$

where $Y \in \mathcal{R}$ and $(X, W) \in \mathcal{K}_X \times \mathcal{K}_W$ are respectively endogenous and exogenous observed variables. The unobserved variable $\alpha \in \{1, \ldots, R\}$ denotes the type of the agent, and $\epsilon$ is the structural error. Suppose that $G$ does not depend on $X$ and that $\epsilon$ and $X$ are independent
conditional on $W$ and $\alpha$. Then $Y$ and $X$ are independent conditional on $(W, \alpha)$. We can write this as a form of (4.1.1) and (4.1.2) with conditioning on $W$.

$$F(y, x|w) = \sum_{r=1}^{R} \Pr(\alpha = r)F_{W}(x|w, \alpha = r)F_{Y}(y|w, \alpha = r)$$

In the rest of this chapter, we consider a discretized version of each $X_k$ obtained by partitioning its support $X_k$. There is some loss of information here but it is limited as the discretized variable can take any finite number of values. The joint distribution of two finite-valued random variables can be represented by a matrix. Similarly, the joint distribution of $K$ finite-valued random variables can be represented by a “$K$-dimensional” matrix or an order-$K$ tensor. The mixture model together with the conditional independence assumption implies that this tensor can be factorized in a certain way. We show that the minimum number of components in this factorization is equal to the nonnegative rank of the tensor.

In practice, the nonnegative rank of a tensor, or even a matrix, is hard to calculate. We obtain a lower bound that is determined by the ranks of all possible matrix representations of the observed distribution. Thus it is the same as the lower bound obtained by Kasahara and Shimotsu (2010), which is determined by the ranks of the probability matrices of all possible partitions of $\{X_1, \ldots, X_K\}$. In addition, we construct examples to show that this bound is not sharp. In the bivariate case, this means that given a probability matrix of rank $R$, there may not exist an $R$-component mixture of product distributions that is consistent with the observed data.
4.2 The Mixture Model and Stochastic Factorization

Suppose that each $X_k = \{1, \ldots, I_k\}$ is finite. When $X_k$ is continuous, we can partition its support into $I_k$ disjoint sets to get a finite-valued version of $X_k$. Define

$$p_{i_1i_2\ldots i_K} = \Pr(X_1 = i_1, X_2 = i_2, \ldots, X_K = i_K) \quad (4.2.1)$$

$$a_{i_k r}^{(k)} = \Pr(X_k = i_k | X^* = r) \quad (4.2.2)$$

$$\pi_r = \Pr(X^* = r) \quad (4.2.3)$$

for $i_k = 1, \ldots, I_k, k = 1, \ldots, K$ and $r = 1, \ldots, R$. Here $p_{i_1i_2\ldots i_K}$ are the observed probabilities, $a_{i_k r}^{(k)}$ are the unknown marginals within the $r$-th component, and as before $\pi_r$ are the mixing probabilities.

The probabilities defined in (4.2.1), (4.2.2) and (4.2.3) can be naturally arranged into vectors, matrices and higher-order arrays. Let $P$ be the $K$-way array of dimensions $I_1 \times \cdots \times I_K$ with typical element $p_{i_1i_2\ldots i_K}$. $P$ is the observed distribution of the data. Its elements are positive and add up to one. Let $A^{(k)}$ be the $I_k \times R$ matrix with typical element $a_{i_k r}^{(k)}$. The $r$-th column $a_{i_k r}^{(k)}$ of $A^{(k)}$ is the distribution of $X_k$ conditional on belonging to the $r$-th component. The elements of every column of $A^{(k)}$ are positive and sum to one. Let $\Delta^{I-1} = \{ p \in \mathbb{R}^I | p_i \geq 0, \sum_i p_i = 1 \}$ be the standard $(I - 1)$-simplex. Then each $a_{i_k r}^{(k)}$ is in $\Delta^{I_k-1}$. Finally, let $\pi$ in $\Delta^{R-1}$ represent the mixing distribution.

The mixture model in equations (4.1.1) and (4.1.3) implies that we can somehow decompose the $K$-way array $P$ into the $K$ matrices $A^{(k)}$.

**Definition 4.2.1.** Let $P$ be a $K$-way array with $p_{i_1i_2\ldots i_K} \geq 0$ and $\sum_{k=1}^{K} \sum_{i_k=1}^{I_k} p_{i_1i_2\ldots i_K} = 1$. 79
We say that \((\pi, \{A^{(k)}\}_{k=1}^{K})\) is an \(R\)-component stochastic factorization of \(P\) if for every \(i_k = 1, \ldots, I_k, k = 1, \ldots, K\) and \(r = 1, \ldots, R\)

\[
p_{i_1i_2\ldots i_K} = \sum_{r=1}^{R} \pi_r \prod_{k=1}^{K} a_{i_k r}^{(k)} \tag{4.2.4}
\]

\[
a_{i_k r}^{(k)} \geq 0 \quad \sum_{i_k=1}^{I_k} a_{i_k r}^{(k)} = 1 \tag{4.2.5}
\]

\[
\pi_r \geq 0 \quad \sum_{r=1}^{R} \pi_r = 1 \tag{4.2.6}
\]

Example 4.2.1. Suppose \(X_1, \ldots, X_K\) are mutually independent. Then there exists a 1-component stochastic factorization of \(P\) with \(A^{(k)} = (\Pr(X_k = 1) \cdots \Pr(X_k = I_k))^T\) and \(\pi = (1)\).

Equation (4.2.4) is implied by the mixture model equations (4.1.1) and (4.1.3). When every \(X_k\) is finite-valued, then (4.2.4) is equivalent to a mixture model of product distributions. The problem of existence of an \(R\)-component mixture model for a given distribution of the data can now be expressed as follows. Given a probability array \(P\) and a positive integer \(R\), is there an \(R\)-component stochastic factorization of \(P\)?

Definition 4.2.2. The outer product \(u \otimes v\) of two vectors \(u \in \mathbb{R}^M\) and \(v \in \mathbb{R}^N\) is the \(M \times N\) matrix obtained by multiplying every element of \(u\) by every element of \(v\). The outer product of \(K\) vectors is a \(K\)-way array.

Equation (4.2.4) can be written as

\[
P = \pi_1 a_1^{(1)} \otimes \cdots \otimes a_1^{(K)} + \cdots + \pi_R a_R^{(1)} \otimes \cdots \otimes a_R^{(K)} \tag{4.2.7}
\]
It is immediately obvious that a decomposition of the form (4.2.7) always exists with \(I_1 I_2 \cdots I_K\) components, \(\pi_r\)'s given by the elements of \(P\) and \(a_r^{(k)}\)'s given by the appropriate standard basis vector for \(\mathbb{R}^{I_k}\). In this trivial case, each cell of \(P\) is treated as a separate component with a degenerate distribution.

A more interesting object is the minimal \(R\) for which (4.2.7) is true. This would give us sharp lower bound on the number of components for a given probability array \(P\).

*Example 4.2.2 (Bivariate Mixture of Products).* Suppose \(K = 2\) so that \(P\) is an \(I_1 \times I_2\) matrix. Let \(P\) come from an \(R\)-component mixture of products. Then there exist \(A\) with columns in \(\Delta^{I_1-1}\), \(B\) with columns in \(\Delta^{I_2-1}\) and \(\pi\) in \(\Delta^{R-1}\) such that

\[
P = A \text{diag}(\pi)B^T = (a_1 \cdots a_R) \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \pi_R \end{pmatrix} \begin{pmatrix} b_1^T \\ \vdots \\ b_R^T \end{pmatrix}
\]  

(4.2.8)

We can compare this example to the singular value decomposition. In an SVD, we require the columns of \(A\) and \(B\) to be orthogonal, but do not restrict them to lie in the simplex. The \(\pi_r\)'s are the non-zero singular values of \(P\) and \(R\) is equal to the rank of \(P\).

In the stochastic factorization (4.2.8), we do not require the columns of \(A\) and \(B\) to be orthogonal, so that this decomposition is not unique. Moreover, \(R\) may be bigger than the rank of \(P\) because we require \(A\) and \(B\) to be nonnegative. So the rank of \(P\) is a lower bound for the minimal number of components, but it is not sharp.

In the following section, we will define an SVD-like decomposition and a concept of rank for higher-order arrays. We will obtain sufficient conditions under which the rank is equal to the minimal number of components. We will also construct examples where this fails to be true.
4.3 Rank, Nonnegative Rank and the Number of Components

For our purposes, an order-$K$ tensor $\mathbf{P}$ of size $I_1 \times \ldots \times I_K$ is an element of $\mathbb{R}^{I_1 \times \ldots \times I_K}$ with typical entry $p_{i_1i_2...i_K}$. Thus an order-1 tensor is a vector, and an order-2 tensor is a matrix. Throughout, we will denote vectors by bold-face lower case letters ($\mathbf{a}, \mathbf{p}, \ldots$), and matrices and higher-order tensors by bold-face upper-case letters ($\mathbf{A}, \mathbf{P}, \ldots$). Elements of arrays will be denoted by the corresponding lower-case letters with the appropriate indices.

For instance, $a_{ij}$ will be the $(i, j)$-th element of the matrix $\mathbf{A}$, $a_{kj}$ will be its $j$-th row etc.

Let $K = \{1, 2, \ldots, K\}$ be the set of modes or directions, and $I_k = \{1, 2, \ldots, I_k\}$ be the set of indices in the $k$-th mode.

Definition 4.3.1. Let $\mathbf{U}$ be an order-$K$ tensor of size $I_1 \times \ldots \times I_K$. We say that $\mathbf{U}$ is a rank-1 tensor if it can be written as the outer product of $K$ vectors $a^1 \in \mathbb{R}^{I_1}, \ldots, a^K \in \mathbb{R}^{I_K}$.

We write $\mathbf{U} = a^1 \otimes \cdots \otimes a^K$ with typical element $u_{i_1\ldots i_K} = a^1_{i_1} \ldots a^K_{i_K}$.

Example 4.3.1. The distribution of two independent binary random variables can be represented by a rank-1 $2 \times 2$ matrix. Similarly, the joint distribution of three mutually independent binary random variables can be represented by an order-3 rank-1 tensor of size $2 \times 2 \times 2$.

Definition 4.3.2. Any order-$K$ tensor can be written as the sum of rank-1 order-$K$ tensors.

This is known as the canonical decomposition of a tensor.

$$\mathbf{P} = \sum_{r=1}^{R} \mathbf{U}_r = \sum_{r=1}^{R} a^{(1)}_r \otimes \cdots \otimes a^K_r$$ (4.3.1)
The decomposition in (4.3.1) was introduced by Hitchcock (1927a) and rediscovered independently by Carroll and Chang (1970) and Harshman (1970) with applications to psychometrics and phonetics respectively. Hitchcock (1927a) was interested in the complexity of tensors and defined several notions of tensor rank analogous to the rank of a matrix.

**Definition 4.3.3.** The outer product rank $\text{rank}_\otimes(P)$ or simply the rank of an order-$K$ tensor $P \in \mathbb{R}^{I_1 \times \ldots \times I_K}$ is the minimal $R$ for which (4.3.1) holds.

We are interested in nonnegative canonical decompositions and so we define the concept of nonnegative rank.

**Definition 4.3.4.** The nonnegative rank $\text{rank}_{\geq 0}(P)$ of an order-$K$ tensor $P \in \mathbb{R}^{I_1 \times \ldots \times I_K}$ is the minimal $R$ for which (4.3.1) holds with $a_{r}^{(k)}$ nonnegative for each $k$ and $r$.

When $K = 2$, the outer product rank and nonnegative rank correspond to the rank and nonnegative rank of a matrix. The nonnegative rank of a matrix was studied extensively in Cohen and Rothblum (1993).

**Proposition 4.3.1.** Let $P$ be a nonnegative order-$K$ tensor whose entries add up to one.

Then $\text{rank}_{\geq 0}(P)$ is a sharp lower bound on the number of components in a stochastic factorization of $P$.

**Proof.** The nonnegative rank is clearly a lower bound for the number of components. We will show that it is sharp. Let $\text{rank}_{\geq 0}(P) = R$. Then

$$P = \sum_{r=1}^{R} a_{r}^{(1)} \otimes \ldots \otimes a_{r}^{K}.$$
with $a_r^{(k)} \geq 0$ and $a_r^{(k)} \neq 0$. Let

$$\tilde{a}_r^{(k)} = \frac{a_r^{(k)}}{\sum_{i_k} a_{i_k r}^{(k)}} \geq 0 \quad \text{and} \quad \pi_r = \sum_{i_1} \ldots \sum_{i_K} a_{i_1 r}^{(1)} \ldots a_{i_K r}^{(K)}.$$  

We have $\sum_r \pi_r = \sum_{k=1}^{K} \sum_{i_k=1}^{I_k} \pi_{i_1 \ldots i_K} = 1$. \hfill \Box

Proposition 4.3.1 tells us that the minimal number of components in a mixture model is the nonnegative rank, and not the rank, of the observed tensor. As in the matrix case, $\text{rank}_{\otimes}(P) \leq \text{rank}_{\geq}(P)$, the rank is a lower bound for nonnegative rank. We would like to know when the rank and nonnegative rank of a tensor coincide for two reasons. First, in the bivariate case, the rank of a matrix is easy to calculate compared to the nonnegative rank. Second, the identification obtained when $K \geq 3$ comes from the uniqueness of the $\text{rank}_{\otimes}(P)$-component canonical decomposition of a higher-order array $P$ under certain sufficient conditions. When $\text{rank}_{\otimes}(P) < \text{rank}_{\geq}(P)$, the stochastic factorization of $P$ may not be unique, even though there exists a unique canonical decomposition.

We first obtain an alternative characterization of the rank and nonnegative rank of tensors.

**Definition 4.3.5.** A $k$-slice of an order-$K$ tensor is the order-$(K - 1)$ tensor obtained by fixing the $k$-th index while varying the remaining $K - 1$ indices.

The rows of a matrix are its 1-slices and the columns of a matrix are its 2-slices. For a nonnegative order-$K$ tensor, we consider the scaled $K$-slices

$$P_{i_K} = \frac{P_{\bullet \ldots \bullet i_K}}{\sum_{i_k=1}^{I_k} \sum_{i_{k-1}=1}^{I_{k-1}} P_{i_1 \ldots i_{K-1} i_K}}.$$  \hfill (4.3.2)

The $P_{i_K}$’s are probability tensors with nonnegative entries that add up to one.
Lemma 4.3.1. Let $\mathbf{P}$ be a nonnegative order-$K$ tensor.

i. $\text{rank}_{\otimes}(\mathbf{P}) = R$ if and only if there exist $R$ and no fewer rank-1 order-$(K - 1)$ tensors $\mathbf{u}^{(1)}_r \otimes \ldots \otimes \mathbf{u}^{(K-1)}_r$ and non-zero $\lambda_{i_Kr}$ such that

$$
\mathbf{P}_{i_K} = \sum_{r=1}^{R} \lambda_{i_Kr} \mathbf{u}^{(1)}_r \otimes \ldots \otimes \mathbf{u}^{(K-1)}_r, \quad i_K = 1, \ldots, I_K \\
\sum_{r=1}^{R} \lambda_{i_Kr} = 1, \quad i_K = 1, \ldots, I_K \\
\sum_{i_k=1}^{I_k} \mathbf{u}^{(k)}_{iKr} = 1, \quad k = 1, \ldots, K - 1, r = 1, \ldots, R.
$$

ii. The above statement is true for $\text{rank}_{\geq 0}(\mathbf{P})$ with $\mathbf{u}^{(k)}_{iKr} \in \Delta_{I_k-1}$ and positive $\lambda_{i_Kr}$.

Part 2 of lemma 4.3.1 is the generalization to higher-order arrays of Theorem 3.2 of Cohen and Rothblum (1993). The above result has a useful geometric interpretation. The rank of a tensor $\mathbf{P}$ is the smallest number of rank-1 order-$(K - 1)$ tensors such that every scaled $K$-slice of $\mathbf{P}$ is an affine combination of them. The nonnegative rank of $\mathbf{P}$ is the smallest number of rank-1 order-$(K - 1)$ probability tensors such that every $K$-slice of $\mathbf{P}$ is a convex combination of them.

When $K = 2$, the second observation leads to the complete characterization of matrices with nonnegative rank equal to rank obtained first by Thomas (1974) and most recently by Mond, Smith, and van Straten (2003). To the best of our knowledge, such a characterization does not exist for higher-order tensors.

In Proposition 4.3.2, we obtain some sufficient conditions for equality of rank and nonnegative rank. In the subsequent examples, we construct tensors for which rank is strictly smaller than nonnegative rank.
Proposition 4.3.2. Let $P$ be a nonnegative order-$K$ tensor. $\text{rank}_\otimes(P) = \text{rank}_{\geq 0}(P)$ if

i. $\text{rank}_\otimes(P) = 1$

ii. $K = 2$ and $\text{rank}_\otimes(P) = \min(I_1, I_2)$

iii. $K = 2$ and $\text{rank}_\otimes(P) = 2$.

Proof. 1. Suppose $\text{rank}_\otimes(P) = 1$. Then

$$P = a^{(1)} \otimes \cdots \otimes a^{(K)} = |a^{(1)}| \otimes \cdots \otimes |a^{(K)}|$$

where $|a|$ for a vector is its component-wise absolute value. The second equality follows from the fact that $P$ is nonnegative.

2. Suppose $K = 2$ and $\text{rank}_\otimes(P) = \min(I_1, I_2) = I_2$. Then a trivial nonnegative decomposition is

$$P = P_{1} \otimes e_1 + \cdots + P_{I_2} \otimes e_{I_2}.$$ 

3. Suppose $K = 2$ and $\text{rank}_\otimes(P) = 2$. Then by part 1 of Lemma 4.3.1, the scaled columns of $P$ are

$$P_{i_2} = \lambda_{i_2} u_1 + (1 - \lambda_{i_2}) u_2, \quad i_2 = 1, \ldots, I_2.$$ 

This is the equation of a line in $\mathbb{R}^{I_1}$. The columns of $P$ lie on a line segment. Let $P_1$ and $P_2$ be the endpoints of this line segment. Then there exist $\mu_{i_2} \in [0, 1]$ such that

$$P_{i_2} = \mu_{i_2} P_1 + (1 - \mu_{i_2}) P_2, \quad i_2 = 1, \ldots, I_2.$$ 

\qed
4.3.1 An example with a rank-3 $4 \times 4$ matrix

By parts 2 and 3 of Prop 4.3.2, the simplest matrix for which rank and nonnegative rank can be different is a $4 \times 4$ matrix of rank 3. The (scaled) columns of a $4 \times 4$ probability matrix $P$ are four points in the 3-simplex. By lemma 4.3.1, when $\text{rank}_\otimes(P) = 3$, these points lie on a plane in the simplex. Figure 4.1 shows two examples of rank-3 $4 \times 4$ probability matrices.

For $\text{rank}_{\geq 0}(P) = 3$, we require the four points to lie in the convex hull of three vectors in the 3-simplex. This is the case in the top right panel. On the other hand, in the bottom right panel, we cannot find a triangle inside the 3-simplex that contains the four points in its convex hull. Here the nonnegative rank is 4.
4.3.2 An example with $2 \times 2 \times 2$ tensors

Consider probability tensors of the following type

$$P = \begin{pmatrix}
    p_{111} & p_{121} & \vline & p_{112} & p_{122} \\
    p_{211} & 0 & \vline & p_{212} & 0
\end{pmatrix} \quad (4.3.3)$$

We can transform $P$ so that each of its 3-slices is a probability matrix.

$$P_i = \frac{P_{\bullet \bullet}}{p_{11i} + p_{12i} + p_{21i} + p_{22i}} = \begin{pmatrix}
    p_{i1} & p_{i2} \\
    1 - p_{i1} - p_{i2} & 0
\end{pmatrix} \quad i = 1, 2. \quad (4.3.4)$$

Any $2 \times 2$ nonnegative matrix is a point in $\mathbb{R}^4_+$. A nonnegative matrix with entries summing up to one is a point in the 3-dimensional simplex. We set corresponding elements of the two slices of $P$ to zero. So each of the matrices $P_1$ and $P_2$ is a point on the same face of the 3-simplex.

In Figure 4.2, we are in the 2-dimensional subspace of $\mathbb{R}^4$ given by $\{p \in \mathbb{R}^4 \mid p_1 + p_2 + p_3 = 1, p_4 = 0\}$. The triangle given by the convex hull of the points $(1, 0, 0, 0), (0, 1, 0, 0)$ and $(0, 0, 1, 0)$ is a face of the 3-simplex. The double-dashed lines are rank-1 matrices i.e., the matrices that can be written as $u \otimes v$ for $u, v \in \mathbb{R}^2$. Clearly, we must have either $u_2 = 0$ or $v_2 = 0$. 

Figure 4.2: Face of the 3-simplex (Double-dashed lines are the rank-1 matrices)
From Lemma 4.3.1, we have:

i. The rank of a non-trivial $2 \times 2 \times 2$ tensor is the smallest number of rank-1 $2 \times 2$ matrices that contain the slices of the tensor in their affine hull.

ii. The nonnegative rank of a non-trivial $2 \times 2 \times 2$ tensor is the smallest number of rank-1 $2 \times 2$ probability matrices that contain the slices of the tensor in their convex hull.

It is immediately obvious that every $P$ of type (4.3.3) has rank and nonnegative rank at most 3. Rank-1 tensors correspond to a single point on the set of rank-1 probability matrices. E.g. in Figure 4.2, $P_1 = P_2 = u \otimes v$.

The first panel of Figure 4.3 shows a rank-2 tensor that has nonnegative rank 2. There exists a unique 2-component nonnegative decomposition that is consistent with the observed tensor. Thus a 2-component mixture of products exists and is uniquely identified. On the other hand, the second panel shows a rank-2 tensor with nonnegative rank 3.
3. There is a unique 2-component decomposition that is not nonnegative. There exists no 2-component mixture of products that is consistent with the observed tensor. Moreover, while 3-component mixtures exist, they are not uniquely identified.

Finally, Figure 4.4 shows two rank-3 tensors whose slices cannot be written as the affine combination of two rank-1 matrices.

### 4.4 Lower Bounds for Tensor Rank

The rank of a matrix can be calculated in polynomial time by, say, Gaussian elimination. In contrast, the nonnegative rank of a matrix and the rank of a higher order tensor are calculated by iteratively estimating the appropriate decompositions with different numbers of components. Nonnegative matrix rank and tensor rank have been shown to be NP-hard (Vavasis 2007, Håstad 1990). In this section, we obtain lower bounds for tensor rank (and hence for nonnegative rank) that can be calculated as the rank of certain matrices.
The following definitions due to Hitchcock (1927b) generalize to higher-order tensors the notions of column rank and row rank of a matrix.

Definition 4.4.1. A mode-$k$ vector is the vector $p_{i_1...i_{k-1}i_{k+1}...i_K}$ obtained by varying the $k$-th index of $P$ while keeping the remaining $K-1$ indices fixed.

Definition 4.4.2. The mode-$k$ rank of a tensor $P$ is the dimension of the linear span of its mode-$k$ vectors. Let $P^{\{k\},K\{k\}}$ be the matrix of mode-$k$ vectors of $P$. Then the rank of $P^{\{k\},K\{k\}}$ is the mode-$k$ rank of $P$. It is denoted by $\text{rank}_k(P)$.

For a matrix $P$, the mode-1 vectors are its columns and the mode-2 vectors are its rows. The column rank and row rank are the mode-1 and mode-2 ranks respectively, and they coincide with the outer product rank. This is not in general true for higher order tensors.

The purpose of this section is to obtain bounds on the outer product rank of a tensor.

Note that the matrix $P^{\{k\},K\{k\}}$ is a flattened version of $P$. We can unfold or flatten an order-$K$ array in a number of different ways. Kolda (2006) refers to this operation as matricization. Consider a partition of the set of modes $\mathcal{K}$, $\mathcal{R} = \{r_1, \ldots, r_l\}$ and $\mathcal{C} = \{c_1, \ldots, c_{K-l}\}$. Let $P^{\mathcal{R},\mathcal{C}}$ be the $(I_{r_1}I_{r_2}...I_{r_l}) \times (I_{c_1}I_{c_2}...I_{c_{K-l}})$ matrix obtained by treating the indices corresponding to the modes in $\mathcal{R}$ as row indices, and the indices corresponding to the modes in $\mathcal{C}$ as column indices. Thus the matrix of mode-1 vectors can be obtained by setting $\mathcal{R} = \{1\}$ and $\mathcal{C} = \{2, \ldots, K\}$. The number of possible partitions of $\mathcal{K}$ into two
sets is given by

\[
\sum_{l=1}^{K-1} \binom{K}{l} \quad \text{for } K \text{ odd}
\]

\[
\sum_{l=1}^{K-2} \binom{K}{l} + \frac{1}{2} \binom{K}{2} \quad \text{for } K \text{ even}
\]

These are the number of ways in which an order-\(K\) tensor can be matricized. For example, an order-3 tensor can be unfolded in three ways into the matrices of mode-1, mode-2 and mode-3 vectors respectively.

**Proposition 4.4.1.** Let \(\text{rank}_{R,C}(P)\) be the rank of the matrix \(P^{R,C}\). Then \(\max(\text{rank}_{R,C}(P)) \leq \text{rank}_{\otimes}(P)\).

**Proof.** First consider the matrix of mode-1 vectors. From (4.3.1) we have

\[
P^{i_1 \ldots i_K} = \sum_{r=1}^{R} a_{2r}^i a_{K+1}^i a_r^1
\]

The columns of \(P^{(1),K\{1\}}\) lie in the span of \(\{a_1, \ldots, a_R^1\}\). Therefore \(\text{rank}_{1}(P) \leq \text{rank}_{\otimes}(P)\).

For arbitrary \(R\) and \(C\), each column of \(P^{R,C}\) can be written as

\[
\sum_{r=1}^{R} a_{i_1 r}^{r_1} \cdots a_{i_l r}^{r_l} \text{vec}(a_r^{c_1} \otimes \cdots \otimes a_r^{c_{K-l}})
\]

The columns of \(P^{R,C}\) lie in the span of \(\{\text{vec}(a_r^{c_1} \otimes \cdots \otimes a_r^{c_{K-l}})\}_{i=1}^{R}\). Therefore \(\text{rank}_{R,C}(P) \leq \text{rank}_{\otimes}(P)\). \(\square\)

The ranks of matricized versions of a tensor give us a lower bound for its rank. In general, this bound may not be achieved. Kruskal (1989) shows that \(2 \times 2 \times 2\) tensors have
typical ranks 2 and 3 over the real field. The following is an example from de Silva and Lim (2008) of a $2 \times 2 \times 2$ tensor with rank 3.

**Example 4.4.1.** Consider the $2 \times 2 \times 2$ tensor $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

\[
P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}
\]

\[
= e_1 \otimes e_2 \otimes e_1 + e_2 \otimes e_1 \otimes e_1 + e_1 \otimes e_1 \otimes e_2
\]

where $e_1$ and $e_2$ are the standard basis vectors for $\mathbb{R}^2$. So $\text{rank}_\otimes(P) \leq 3$ and $\text{rank}_k(P) = 2$ for every $k$. Suppose that $\text{rank}_\otimes(P) = 2$. Then we can find non-zero $u_r, v_r, w_r$ in $\mathbb{R}^2$ such that

\[
P = u_1 \otimes v_1 \otimes w_1 + u_2 \otimes v_2 \otimes w_2
\]

$w_1$ and $w_2$ are linearly independent as they span $e_1$ and $e_2$.

We can slice $P$ into two $2 \times 2$ matrices by fixing the third index. Thus

\[
P^1 = w_{11}(u_1 \otimes v_1) + w_{21}(u_2 \otimes v_2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}
\]

\[
P^2 = w_{12}(u_1 \otimes v_1) + w_{22}(u_2 \otimes v_2) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}
\]

Let $w_1^\perp$ be orthogonal to $w_1$. Then the matrix

\[
\begin{pmatrix} w_{12}^\perp & w_{11}^\perp \\ w_{11}^\perp & 0 \end{pmatrix} = w_{11}^\perp P^1 + w_{12}^\perp P^2 = (w_{11}^\perp w_{21} + w_{12}^\perp w_{22})(u_2 \otimes v_2)
\]

has rank 1. This is possible only when $w_{11}^\perp = 0$, which means that $w_{12} = 0$. By a similar argument we can show that $w_{22} = 0$, which gives us the required contradiction.
The above example corresponds to the first panel of Figure 4.4. We can obtain a similar proof for the tensor in the second panel. In both these case, the outer product rank was bigger than the mode-k ranks as well as the dimension along each mode of the tensor. Although there exist loose upper bounds for tensor rank, the maximum possible rank of a tensor of arbitrary size is not known.

4.5 Conclusion

The main contribution of this paper is the observation that a finite mixture of products can be represented as the stochastic factorization of the probability tensor of the observed distribution. Lower bounds on the minimal number of components immediately follow from this representation.

We obtain certain sufficient conditions under which these lower bounds are sharp i.e., when the rank and nonnegative rank of the tensor coincide. However, it would be desirable to have a complete characterization of tensors for which this property holds. The examples we construct suggest that such a characterization would involve conditions on the position of slices of the tensor, relative to each other and to the space of rank-1 tensors.

In general, the lower bounds obtained here are not sharp. However, unlike the nonnegative rank, they are easily calculated and can be used as initial values in model fitting.
Appendix

Proof of Lemma 4.3.1.

Proof. 1. ‘If’ part: Suppose \( \text{rank}_\otimes(P) = R \). Then there exist \( R \) and no fewer non-zero \( a_r^{(k)} \in \mathbb{R}^{I_k} \) such that \( P = \sum_{r=1}^{R} a_r^{(1)} \otimes \cdots \otimes a_r^{(K)} \). From equation (4.3.2)
\[
P_{i_K} = \sum_{r=1}^{R} \sum_{k=1}^{K-1} \sum_{i_k=1}^{I_k} p_{i_1 \ldots i_{K-1} i_K} a_r^{(1)} \otimes \cdots \otimes a_r^{(K-1)}
\]

Let
\[
u_r^{(k)} = \frac{a_r^{(k)}}{\sum_{i_k=1}^{I_k} a_{i_k r}}, \quad \lambda_{i_K r} = \frac{a_r^{K} \left( \sum_{i_1}^{I_1} a_{i_1 r}^{(1)} \right) \cdots \left( \sum_{i_{K-1} = 1}^{I_{K-1}} a_{i_{K-1} r}^{(K-1)} \right)}{\sum_{k=1}^{K-1} \sum_{i_k=1}^{I_k} p_{i_1 \ldots i_{K-1} i_K}},
\]

and note that
\[
\sum_{r=1}^{R} \left( a_r^{K} \left( \sum_{i_1}^{I_1} a_{i_1 r}^{(1)} \right) \cdots \left( \sum_{i_{K-1} = 1}^{I_{K-1}} a_{i_{K-1} r}^{(K-1)} \right) \right) = \sum_{k=1}^{K-1} \sum_{i_k=1}^{I_k} p_{i_1 \ldots i_{K-1} i_K}.
\]

‘Only if’ part: Let \( u_r^{(K)} = \left( \lambda_{1 r} \left( \sum_{k=1}^{K-1} \sum_{i_k=1}^{I_k} p_{i_1 \ldots i_{K-1} i_K} \right) \right)^T \). Then \( P = \sum_{r=1}^{R} u_r^{(1)} \otimes \cdots \otimes u_r^{(K-1)} \otimes u_r^{(K)} \) and \( \text{rank}_\otimes(P) \leq R \). Equality follows from the ‘if’ part. \( \square \)


