The Econometrics Journal

Econometrics Journal (2016), volume **19**, pp. C95–C127.
doi: 10.1111/ectj.12068

Using mixtures in econometric models: a brief review and some new results

GIOVANNI COMPIANI[†] AND YUICHI KITAMURA[‡]

[†]*Department of Economics, Yale University, New Haven, CT-06520, USA*.
E-mail: giovanni.compiani@yale.edu

[‡]*Cowles Foundation for Research in Economics, Yale University, New Haven, CT-06520, USA*.
E-mail: yuichi.kitamura@yale.edu

First version received: June 2015; final version accepted: June 2016

Summary This paper is concerned with applications of mixture models in econometrics. Focused attention is given to semiparametric and nonparametric models that incorporate mixture distributions, where important issues about model specifications arise. For example, there is a significant difference between a finite mixture and a continuous mixture in terms of model identifiability. Likewise, the dimension of the latent mixing variables is a critical issue, in particular when a continuous mixture is used. We present applications of mixture models to address various problems in econometrics, such as unobserved heterogeneity and multiple equilibria. New nonparametric identification results are developed for finite mixture models with testable exclusion restrictions without relying on an identification-at-infinity assumption on covariates. The results apply to mixtures with both continuous and discrete covariates, delivering point identification under weak conditions.

Keywords: *Continuous mixture models, Finite mixture models, Multiple equilibria, Nonparametric identification, Unobserved heterogeneity.*

1. INTRODUCTION

Mixture models are widely used in economic applications. First of all, mixtures are commonly employed as a modelling device to account for unobserved heterogeneity, with applications ranging from the labour to the industrial organization literature; references include Berry et al. (2006), Keane and Wolpin (1997) and Cameron and Heckman (1998). Mixtures also arise when the unobserved element that varies across individuals is the form of their utility functions, as in random utility models (RUMs). Another application in microeconometrics is the treatment of multiple equilibria in discrete games; see, e.g. Cooper (2002), Berry and Tamer (2006) and Ciliberto and Tamer (2009). Measurement error models can be reformulated in terms of mixtures, as in Horowitz and Manski (1995), Manski (2003), Chen et al. (2011) and Schennach (2013). Related applied papers include Hu et al. (2013), Bonhomme and Robin (2014) and Arellano et al. (2014). Heckman and Singer (1984a, b) and Abbring and van den Berg (2003) apply mixtures to the analysis of duration models. Finite mixtures have also been employed in regime switching models; see, e.g. Cho and White (2007). The above is an incomplete list of the many applied settings in which mixture models arise. Some of these applications are analysed in subsequent sections in greater detail.

The tremendous importance of mixtures in economics, as well as other fields, has motivated a large body of theoretical research, in both the statistics and the econometrics literature. Given the vastness of this research area, a comprehensive account is far beyond the scope of this paper. Instead, we focus on certain recent advances concerning semiparametric and nonparametric treatments of mixtures. These developments increase the flexibility of the models and reduce their reliance on restrictive parametric assumptions. This is important because consistency of estimators for the parameter of interest usually hinges on the correct specification of the unobserved heterogeneity distribution. Cases where this issue is highly relevant include proportional hazard models, as in Heckman and Singer (1984b), discrete choice models, as in, e.g. Matzkin (1992), and switching regression models, as in, e.g. Kitamura (2003).

The greater flexibility allowed by the semiparametric and nonparametric approaches often comes at the cost of reduced identifying power. In other words, relaxing the traditional parametric assumptions often does not allow the researcher to uniquely identify the parameters of interest and only partial identification results can be obtained. A variety of examples in the paper illustrate this trade-off.

We now lay out the general model that is used throughout the paper. Consider the following:

$$F(y) = \int F(y|\alpha)\,\mathrm{d}G(\alpha). \tag{1.1}$$

In (1.1), F is a cumulative distribution function (CDF) of an observed random variable Y, $F(\cdot|\alpha)_{\alpha\in\mathcal{A}}$ is a collection of CDFs indexed by a random variable $\alpha \in \mathcal{A}$, where $\mathcal{A}$ is a (possibly infinite) set, and G is a CDF on the space $\mathcal{A}$. Throughout the paper, we employ the following terminology. The CDF F is referred to as the mixture distribution, the CDFs $F(\cdot|\alpha)_{\alpha\in\mathcal{A}}$ are called the component distributions, G is the mixing distribution and α is the unobserved latent (or mixing) variable. In words, (1.1) says that the distribution F is a mixture with components $F(\cdot|\alpha)_{\alpha\in\mathcal{A}}$ and mixing distribution G. In the case where α takes on a finite number of values, the probability mass attached by the distribution G to each α_i is often called the weight of component i.

In a typical setting, the econometrician observes repeated draws $(y_i)_{i=1}^n$ from the mixture distribution F and wishes to estimate the unknown parameters on the right-hand side of (1.1), namely the component distributions and the mixing distribution.

The basic framework of (1.1) can be specialized and extended in several directions. First, the latent variable α can take a finite number of values or an infinite (typically uncountable) number of values. The case of a finite mixture is dealt with in Section 2, while general mixtures are addressed in Section 3. Second, the econometrician needs to choose whether to specify a parametric form for the component distributions or to pursue a nonparametric approach. A similar decision applies to the mixing distribution. Third, the dimension of the latent variable α can play an important role, particularly in the case of mixtures with infinitely many components. Finally, one may want to introduce covariates in the basic model of (1.1). As discussed in the following sections, the assumptions about the way in which the component distributions and/or the mixing distribution depend on the covariates are key to obtaining identification of the parameters of interest.

This paper is organized as follows. In Section 2, we discuss finite mixtures and present some new identification results. In Section 3, we deal with possibly infinite mixtures; in particular, panel data models, random coefficient models and random utility models are considered. We conclude in Section 4. All proofs are given in the Appendix.

2. FINITE MIXTURES

Finite mixtures have been analysed extensively in the statistical literature. An exhaustive account of the statistical literature on mixtures (including infinite mixtures) is provided by Lindsay (1988). Chapter 8 of Prakasa Rao (1983) provides excellent discussions of identifiability of mixtures, including finite mixtures. See also McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

Most of the literature has taken a parametric approach to finite mixture models, which means that the component distributions are assumed to be known up to finite-dimensional parameters. Under such assumptions, identification of the parameters of interest is usually straightforward. However, these identification results are fully driven by the chosen functional form and model misspecification generally leads to inconsistent estimates and invalid inference. Therefore, a growing body of research, both in statistics and in econometrics, investigates how semiparametric and nonparametric methods can be applied to mixture models. In what follows, we consider some of these recent developments. Identification strategies under four different scenarios are discussed: multiple outcomes with independence properties are observed; covariates enter the mixing weights; covariates enter the component distributions; covariates enter both the mixing weights and the component CDFs. For the fourth case, new identification results are presented.

2.1. Multiple outcomes with independence property

An important contribution to the nonparametric analysis of mixtures in the statistical literature is provided by Hall and Zhou (2003). They consider two-component mixtures with an independence property. The model takes the form

$$F(\mathbf{y}) = \lambda \prod_{j=1}^{k} F_{j1}(y_j) + (1-\lambda) \prod_{j=1}^{k} F_{j2}(y_j), \tag{2.1}$$

where $\mathbf{y} = (y_1, \ldots, y_k)'$ and the CDF of each component factorizes by the independence assumption. The motivation for considering this model comes from the clinical trial literature. In particular, one can think of $\mathbf{y}$ as a vector of outcomes for k clinical tests, whose distribution is allowed to vary depending on whether a patient is affected by a given disease or not. The researcher observes $\mathbf{y}$, but does not observe the patient's disease status nor the proportion of people affected. Therefore, the problem is to nonparametrically identify and estimate the CDFs F_{ji}, for $j = 1, \ldots, k$ and $i = 1, 2$, and the mixture weight λ based on random draws from F. Clearly, if $k = 1$, identification is a hopeless task. For the case $k = 2$, Hall and Zhou (2003) show that the parameters of interest are not identified. In particular, given the estimable CDF F, (2.1) has a continuum of solutions indexed by two scalar parameters. However, when $k = 3$, the authors show that identification (up to switching of the two products on the right-hand side of (2.1)) is achieved under an irreducibility condition on the density f of the CDF F.[1]

The proof exploits the fact that every lower-dimensional submodel derived from (2.1) imposes a different restriction on the parameters of interest. The total number of restrictions is $2^k - 1$ while the number of parameters is $2k + 1$. Note that for $k = 3$, these two numbers

[1] The density f is said to be irreducible if none of its bivariate marginal densities factorizes into the product of the two corresponding univariate marginals.

coincide. Therefore, when $k < 3$ the model is not identified, when $k = 3$ we have just-identification and when $k > 3$ there are over-identifying restrictions.

Given identification (i.e. for $k = 3$), Hall and Zhou (2003) propose a fully nonparametric estimator for the component distributions and the mixing weight. The estimator is obtained by minimizing an integrated distance between an estimator of F, which does not take into account the structure of the model as specified in (2.1), and an expression (involving the unknowns with respect to which the minimization is carried out), which is instead based on (2.1). This estimation procedure is close in spirit to nonparametric maximum likelihood (ML) and empirical likelihood; see Laird (1978), Qin (1998), Qin (1999) and Zou et al. (2002) for details about how these frameworks can be applied to mixture models. The estimators proposed by Hall and Zhou (2003) for the component distributions and the mixing weight are strongly, uniformly consistent. Moreover, under regularity conditions, the estimator for λ achieves the root-n rate of convergence and the estimators for the components are root-n consistent in L^2.[2]

An attractive feature of the above results is that they rely on relatively mild regularity conditions and only require the number of observations for each individual in the sample to be larger than or equal to 3. In particular, this means that the results can be applied to settings in which only a few observations per individual unit are available, as in the case of a short panel.

The work by Hall and Zhou (2003) has been extended and complemented by several papers in both the statistics and the econometrics literature. Among the latest developments, Bonhomme et al. (2016b) provide sufficient conditions for nonparametric identification of the component distributions, the mixing weights and the number of components.[3] The conditions impose some restrictions on the component densities, namely absolute continuity, square integrability and lack of multicollinearity among the Fourier coefficients of an expansion of the component densities. The identification proof is quite different from that of Hall and Zhou (2003), which allows Bonhomme, Jochmans and Robin to obtain two novel results. First, they provide a consistent estimator for the number of mixture components. Second, their approach is more flexible as it allows for an arbitrary number of mixtures, while Hall and Zhou (2003) only deal with the two-component case. Finally, the estimators proposed by Bonhomme, Jochmans and Robin appear to be computationally less costly than the nonparametric ML estimators of Hall and Zhou (2003).

Further, Allman et al. (2009) show that Hall and Zhou (2003) can be viewed as a special case of a more general model in which identification comes from applying an algebraic result on three-way arrays by Kruskal (1977). This theorem can be used to show identification of a broad class of models, including discrete hidden Markov models and random graph models. However, the proof is typically not constructive and thus does not directly lead to an estimation procedure. This issue is explored by Bonhomme et al. (2016a), who suggest an estimator based on the joint approximate diagonalization of matrices.

Another relevant extension encompassing a variety of economic applications can be found in Kasahara and Shimotsu (2009). The focus is on dynamic discrete choice models with unobserved heterogeneity, where mixtures represent the different (latent) types that the agents belong to. Instead of the independence condition in (2.1), they consider Markovian structures. Moreover, following the identification strategy based on covariates in Kitamura (2003), they change the model structure in (2.1) by introducing covariates, and they employ techniques akin to those

[2] A sufficient condition for root-n convergence is that the component distributions be compactly supported. This condition can be relaxed by imposing constraints on the tail behaviour of the component distributions.

[3] As in Hall and Zhou (2003), the dimension of the outcome variable is assumed to be at least 3.

used by Hall and Zhou (2003).[4] Indeed, the main source of identifying power in this setting is exactly the fact that the outcome variable responds differently to changes in the covariates for different types. In other words, identification requires the covariates to affect the conditional choice probabilities of the agent in a sufficiently heterogeneous way across types. Importantly, the presence of covariates makes identification possible even with relatively short panels, which is the type of data set that is often available in applications.

The paper also provides sufficient conditions for the nonparametric identification of the number of components, i.e. the number of types in the dynamic discrete choice setting. Interestingly, the number of types can be nonparametrically identified even with just two-period panels. The proof relies on rank assumptions requiring the change in the covariates to have a sufficiently strong effect on the conditional choice probabilities. However, such conditions (which are also needed for identification of the other parameters of interest) cannot be tested using the data. A computationally attractive procedure for estimation of dynamic discrete models with unobserved heterogeneity is proposed by Kasahara and Shimotsu (2011).

2.2. *Covariates in the mixing weights*

As discussed above, the model is in general not identified if the outcome is less than three-dimensional. One way to improve on this negative result is to assume the data have covariates that satisfy certain exclusion restrictions. Moreover, even if point identification is not obtained, it may still be interesting to characterize the identified set. This is the approach taken by Henry et al. (2014), who consider the following model

$$F(y|x, w) = \sum_{j=1}^{J} \lambda_j(x, w) F_j(y|x), \tag{2.2}$$

where y is now scalar-valued. The paper shows that this model is only partially identified when the outcome is not required to be at least three-dimensional. However, the characterization of the identified set is constructive and can be used to extract useful information from the data, such as the nonparametric identification of the number of mixture components.

The main assumption on which the model relies is embedded in (2.2) and takes the form of an excluded covariate restriction. The assumption states that the covariate W affects the mixing weights, but does not change the component distributions. The entire analysis is conditional on values of X for which the excluded covariate assumption holds. As no other property is imposed on X, conditioning on this variable will be left implicit throughout. In accordance with instrumental variable models, it is also required that the dependence of the mixing weights on the covariate W be strong enough.

The exclusion restriction on W can be justified in a number of applications. First, consider Markov switching models; see, e.g. Cosslett and Lee (1985) and Hamilton (1989). Assume that the outcome variable Y is an mth-order autoregression conditionally on the value of a state variable that follows a Markov chain. The hidden Markov chain determines the distribution of the outcome variable; for instance, it could determine the expectation (as in mean switching models) or the variance (as in stochastic volatility models). This type of model can be expressed in the form of (2.2) by setting $X = (Y_{t-1}, \ldots, Y_{t-m})$ and W can be chosen from Y whose lag orders are

[4] We discuss details of this identification approach in Section 2.3.

$m+1$ or higher, e.g. $W = Y_{t-m-1}$ or $W = (Y_{t-m-1}, \ldots, Y_1)$. Then, each component distribution would not depend on W. Moreover, in general, the distribution of the unobserved state S will depend on lagged values of the outcome, i.e. $\lambda_j(x, w) = P(S_t = j|X = x, W = w)$ will depend on w, which corresponds to the relevance condition for instrumental variables.

Another example where the model (2.2) is relevant is the misclassification problem. In this set-up, the researcher observes an outcome variable Y and potentially flawed measurements T of an underlying categorical regressor T^*. The case of a discrete mismeasured regressor (often referred to as misclassification) is especially relevant, given that in this setting the classical assumption of independence between measurement error and true value is untenable. This implies that traditional methods based on deconvolution cannot be applied.

A common assumption in the misclassification literature is that, conditional on the true value of the regressor T^*, the outcome Y and the observed regressor T are independent. This assumption is sometimes referred to as nondifferential measurement error. Under this restriction, we can write

$$\begin{aligned} F_{y|T}(Y|T) &= \sum_{j=1}^{J} F_{Y|T,T^*}(y|T, T^* = t_j) P\{T^* = t_j|T\} \\ &= \sum_{j=1}^{J} F_{Y|T^*}(y|T^* = t_j) P\{T^* = t_j|T\}. \end{aligned} \tag{2.3}$$

This equation shows that misclassification models can be written in the form of model (2.2) by setting the excluded variable W equal to the observed regressor T and $\lambda_j = P\{T^* = t_j|T\}$. Here, the role of the latent variable giving rise to the mixture is played by the unobserved correctly-classified regressor T^*.

The nondifferential measurement error assumption is not innocuous. See, for example, Section 2.5 of Carroll et al. (2006) and Bound et al. (2001) for possible causes of differential measurement error. Further references in the misclassification literature include Mahajan (2006), Molinari (2008), Chen et al. (2011) and Hausman (2001).

A third setting where model (2.2) can be applied is microeconometric models with unobserved heterogeneity. For example, if the outcome variable is demand for a good, the researcher may be interested in allowing both for observed heterogeneity across buyers (given by the covariates for which data are available) and for unobserved heterogeneity (given by a finite number of types and, possibly, an idiosyncratic shock). In this case, the excluded instrument W could be a set of geographical variables that do not enter preferences or covariates – and thus do not affect demand for a given type – but do change the distribution of buyer types. Another example is an oligopoly model, where the outcome variable is again demand, the observed covariates are prices and (possibly mismeasured) costs, and the excluded instrument is (possibly mismeasured) profits. As long as profits do not belong to the buyer's information set, they do not affect demand given the covariates and the buyer type. Further, if profits do have an influence on the composition of demand, then the instrument relevance condition is satisfied as well. Additional details on this model can be found in Henry et al. (2014).

Finally, (2.2) can be used to account for multiple equilibria in economic models. In this context, each element of the mixture represents a different equilibrium and the mixing weights correspond to the probability distribution over equilibria given by some selection mechanism. The restrictions on the instrument require W not to affect the distribution of the outcome variable in any given equilibrium, but to have an impact on the equilibrium selection mechanism.

Policy interventions are often argued to satisfy these conditions. For instance, in the analysis of the airline market by Ciliberto and Tamer (2009), one may claim that anti-collusion policies do not affect a firm's entry decision on a given market, but do influence the equilibrium selection differentially across regional markets. See Lewbel and Tang (2015) for nonparametric identification results in an incomplete information game theoretic model that implies a closely related restriction. Henry et al. (2014) provide further examples of excluded instruments used in the development, macroeconomics and international finance literatures.

The previous paragraphs listed some of the applications in which the exclusion restriction on W may be plausibly justified. We now discuss the results that Henry et al. (2014) obtain based on model (2.2). The exclusion and relevance restrictions are in general not sufficient to obtain nonparametric point identification of the mixing weights and component distributions.[5] Henry et al. (2014) show that the parameters of interest can be expressed as functions of $J(J-1)$ scalar parameters, where J is the number of components and is, for now, assumed to be known. The identified set then is given by all the values in $R^{J(J-1)}$, which imply values of the mixing weights and of the component distributions that satisfy obvious restrictions. In particular, the implied mixing weights must be between 0 and 1 and the implied component CDFs must be nondecreasing, right-continuous and have the correct limits as their argument goes to $\pm\infty$. Moreover, the characterization of the identified set for the case of two-component mixtures provides insight into the role played by variation in the outcome and the instrument. Intuitively, the identified set shrinks as variation in the outcome conditional on the instrument increases and as the effect of the instrument on the distribution of the outcome becomes stronger.

While the model is in general not identified, it is possible to achieve point identification of certain quantities of interest. In the case of two-component mixtures, it is shown that any linear functional of $F_1(y|x) - F_0(y|x)$ is identified up to scale. This implies, for example, that the ratio

$$\frac{P_1\{y > a|x\} - P_0\{y > a|x\}}{E_1[y|x] - E_0[y|x]}$$

is point identified for all values of a, provided the denominator is nonzero.[6] In the context of randomized experiments with misclassified treatment, this quantity can be interpreted as a quantile treatment effect relative to the (nonzero) average treatment effect.

The results obtained by Henry et al. (2014) can be applied to a variety of situations as model (2.2) imposes relatively weak restrictions. In particular, both the outcome Y and the covariates X can be discrete or continuous. This is in contrast with other studies, where more structure is imposed on the model. For example, in the misclassification literature, Molinari (2008) assumes that the outcome takes a finite number of values, and Bollinger (2006) restricts attention to mismeasured binary regressors to derive bounds on $E[Y|X]$. Moreover, no constraints are put on the mixing weights, whereas Horowitz and Manski (1995) model contaminated data as a two-component mixture and impose an upper bound on the probability of contamination. Note that the results in Henry et al. (2014) rely on testable assumptions. In particular, for the $J = 2$ case, the identified set is defined in terms of estimable bounds. These identified quantities can be used to perform a specification test for the model. More precisely, the researcher can jointly test the exclusion restriction and the assumption that the mixture has

[5] Recall that in this section we do not impose any restrictions on the dimensionality of the outcome; in particular, it is allowed to be one- or two-dimensional, in contrast to Section 2.1.

[6] Here, P_i and E_i denote probability and expectation under F_i, respectively.

two components. The same idea can be used to determine the number of components, to which end Henry et al. (2014) propose a simple iterative procedure. The discussion above is concerned with partial identification. The reader is referred to Henry et al. (2013) for the use of additional restrictions on relative tail behaviour of component distributions to achieve point identification and nonparametric estimation under such restrictions. In a recent paper, Hohmann and Holzmann (2013) discuss related results, and in particular, show that the Hall–Zhou model in Section 2.1 can be studied within the framework of the Henry–Kitamura–Salanié model in this section.

2.3. *Covariates in the component distributions*

Section 2.2 focused on the case where identification stems from the fact that there is a random variable affecting the mixing weights, which is excluded from the component distributions. Now we consider the opposite set-up, along the lines of Kitamura (2003). The key identifying assumptions will concern the way in which the component distributions depend on covariates, whereas the mixing weights will be assumed to be independent of covariates. The model is given by

$$F(y|x) = \sum_{j=1}^{J} \lambda_j F_\epsilon^j(y - m_j(x)), \tag{2.4}$$

where F_ϵ^j, $j = 1, \ldots, J$ are the CDFs of

$$\epsilon_j = y - m_j(x)$$

for $j = 1, \ldots, J$. We can interpret model (2.4) as a switching regression model, as follows

$$y = m_j(x) + \epsilon_j, \quad \epsilon_j \sim F_\epsilon^j, \quad \text{with probability } \lambda_j. \tag{2.5}$$

Models of the form (2.5) have traditionally been tackled in a parametric framework. Identification is achieved by assuming specific functional forms for the m_j functions and the conditional error distributions. Once the problem is reduced to a finite dimension, estimation can usually be carried out via ML. However, such parametric assumptions are very strong and, if incorrect, may lead to inconsistent ML estimates. This is in contrast with standard regression models (i.e. without switching) for which the researcher can estimate the regression function nonparametrically without making functional form assumptions or choosing a distributional form for the error. The risk of misspecification inherent in the parametric approach motivates the investigation of the nonparametric identification of model (2.5), as carried out in Kitamura (2003).

First, it is obvious that the standard conditional mean restriction on the errors is not sufficient to achieve identification. This is because mean independence allows multiple – in fact, infinitely many – ways to split the mixture distribution into distinct components. Thus, a stronger condition is required for point identification. Statistical independence between the errors and the covariates X turns out to be enough. This condition can be interpreted as a 'shift restriction', in the sense that the entire distribution of the ϵ_j needs to remain invariant with respect to covariate values. In addition, identification of the regression functions m_j requires these functions not to be parallel in some neighbourhood. Regularity conditions are also imposed on the tail behaviour of the

moment-generating functions of the error terms.[7] The assumptions listed above are sufficient to point identify all of the unknown parameters, namely the regression functions m_j, $j = 1, \ldots, J$, the error distributions F_{ϵ_j}, $j = 1, \ldots, J$ and the mixing weight λ_j, $j = 1, \ldots, J$. The result is fully nonparametric as it does not rely on any functional form assumptions.

Further, Kitamura (2003) uses the nonparametric identifiability results above to show identification of finite-dimensional parameters defined by semiparametric restrictions via instrumental variables when some of the regressors are endogenous. This model qualifies as semiparametric because the conditional distributions of the error terms are not specified and thus are treated as nuisance parameters. To our knowledge, a fully nonparametric treatment of endogeneity in mixture models is a largely unexplored area of research.

Finally, Kitamura (2003) proposes estimation procedures based on the constructive identification proof. The estimators are fully nonparametric. Moreover, they do not require numerical optimization to compute, making the method convenient in practice.

2.4. New results on point identification of finite mixtures

The previous three subsections have shown that, in general, strong conditions are needed to ensure identification of finite mixture models. In Section 2.1, the outcome was assumed to be at least three-dimensional with each component exhibiting independence. In Section 2.2, the identifying power came from an exclusion restriction on some of the covariates and, in general, one could obtain only partial identification in that set-up. Finally, the results in Section 2.3 imposed additional structure and full independence between the regressors and the error terms was required.

We now present some new point identification results, which apply to general finite mixtures. In particular, we show that under appropriate conditions on the way in which the covariates enter the model, we can obtain point identification of all unknown parameters. The set-up is similar to that of Section 2.2, where only partial identification was achieved in general. For clarity of exposition, we first focus on the case of two-component mixtures; we then extend the results to finite mixtures with an arbitrary number of components. The first model we consider is

$$K(y|\mathbf{z}, w, x) = \lambda(w, x)F_1(y|\mathbf{z}, x) + (1 - \lambda(w, x))F_2(y|\mathbf{z}, x), \tag{2.6}$$

where Y is a scalar or a vector outcome variable, and X, W and $\mathbf{Z} = (Z_1, Z_2)'$ are covariates. Note that the mixing weights are assumed to be a function of W, but not of $\mathbf{Z}$, a restriction to be relaxed in Section 2.4.4. Conversely, the component CDFs are assumed to depend on $\mathbf{Z}$, but not on W. Both the component CDFs and the mixing weights are allowed to be a function of X. Further, for simplicity, we focus on the case where W, Z_1 and Z_2 are all scalar-valued. If not, any element of (W, Z_1, Z_2) not used for identification can be subsumed into X, whose dimension is left unspecified. Suppose Y takes its values in R^p. We make the following assumptions.

ASSUMPTION 2.1. *The random variables Y, X, $\mathbf{Z}$ and W have support $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{W}$, respectively, and $\mathbf{Z}$ and W are continuously distributed.*

Note that $\mathcal{Y}$ can be a discrete set, so the result here applies to a mixture of discrete choice models, or models where the outcome vector has both continuous and discrete elements. For the

[7] Alternatively, Kitamura (2003) provides a set of assumptions on the tail behaviour of the characteristic functions of the error terms' distributions. These alternative assumptions include relevant cases, such as normal distributions with different variances and distributions for which the moment-generating functions do not exist.

rest of the paper, we use the convention that $\mathcal{XY}$ denotes the joint support of (X, Y), $\mathcal{XYZ}$ the joint support of $(Y, X, \mathbf{Z})$, and so on.

ASSUMPTION 2.2. *The functions $\lambda(\cdot, x)$ and $F_i(y|\cdot, x)$ for $i = 1, 2$, are differentiable for every $(x, y) \in \mathcal{X} \times R^p$.*

Assumptions 2.1 and 2.2 are not crucial and will be relaxed later. The following conditions, however, are key to our identification strategy.

ASSUMPTION 2.3. *For each $(x, \mathbf{z}) \in \mathcal{XZ}$ there exists $y^* \in R^p$ such that $(\partial/\partial z_1)F_1(y^*|\mathbf{z}, x) \neq 0$, $(\partial/\partial z_2)F_1(y^*|\mathbf{z}, x) = 0$, $(\partial/\partial z_1)F_2(y^*|\mathbf{z}, x) = 0$ and $(\partial/\partial z_2)F_2(y^*|\mathbf{z}, x) \neq 0$.*

Let $\mathcal{W}|(x, \mathbf{z})$ be the support of the conditional distribution of W given $(X, \mathbf{Z}) = (x, \mathbf{z})$.

ASSUMPTION 2.4. *For each $(x, \mathbf{z}) \in \mathcal{XZ}$ there exists $w^* \in \mathcal{W}|(x, \mathbf{z})$ such that $(\partial/\partial w)\lambda(w^*, x) \neq 0$.*

Assumption 2.3 requires that, for some value in the support of Y, each component conditional CDF be affected by one of the covariates in the $\mathbf{Z}$ vector, but not by the other one. In Section 2.4.3, we relax this condition and consider the case where it holds for only one of the component CDFs.

Assumption 2.4 is simply requiring the covariate W to enter model (2.6). In fact, if Assumption 2.4 did not hold, then W would not affect either the mixing weights or the component CDFs and thus it would not even enter the conditioning on the left-hand side of (2.6). Further comments on the identifying assumptions and how they can be substantiated in economic applications are provided in Remarks 2.3 and 2.4.

We are now ready to state the first new result.

THEOREM 2.1. *Consider model* (2.6) *and let Assumptions 2.1–2.4 hold. Then, the component CDFs $F_i(y|\mathbf{z}, x)$, for $i = 1, 2$, are nonparametrically identified for every $(y, x, \mathbf{z}) \in R^p \times \mathcal{XZ}$ and $\lambda(w, x)$ is nonparametrically identified for every $(x, w) \in \mathcal{XW}_x^*$, where $\mathcal{XW}_x^* \equiv \{(x, w) \in \mathcal{XW}$ such that $(\partial/\partial w)\lambda(w, x) \neq 0\}$. Moreover, the results are constructive. Specifically,*

$$F_1(y|\mathbf{z}, x) = K(y|\mathbf{z}, w^*, x) - \frac{(\partial/\partial z_2)K(y^*|\mathbf{z}, w^*, x)}{(\partial^2/\partial w \partial z_2)K(y^*|\mathbf{z}, w^*, x)} \frac{\partial}{\partial w} K(y|\mathbf{z}, w^*, x)$$

and

$$F_2(y|\mathbf{z}, x) = K(y|\mathbf{z}, w^*, x) - \frac{(\partial/\partial z_1)K(y^*|\mathbf{z}, w^*, x)}{(\partial^2/\partial w \partial z_1)K(y^*|\mathbf{z}, w^*, x)} \frac{\partial}{\partial w} K(y|\mathbf{z}, w^*, x).$$

Further,

$$\lambda(w, x) = \frac{\zeta(w, x)}{1 + \zeta(w, x)}$$

where

$$\zeta(w, x) \equiv -\frac{(\partial/\partial z_1)K(y^*|\mathbf{z}, w, x)(\partial^2/\partial w \partial z_2)K(y^*|\mathbf{z}, w, x)}{(\partial/\partial z_2)K(y^*|\mathbf{z}, w, x)(\partial^2/\partial w \partial z_1)K(y^*|\mathbf{z}, w, x)}.$$

A few remarks are in order.

REMARK 2.1. The identification strategy used to prove Theorem 2.1 does not impose any support assumptions on the covariates. This is in contrast to several contributions in the literature,

which rely on 'identification at infinity'-type restrictions – see, e.g. Tamer (2003) – for the case of multiple equilibria in discrete games. It is a well-documented fact that identification at infinity of a parameter tends to be associated with slow rates of convergence of its estimators; see, for instance, Chamberlain (1986) and Andrews and Schafgans (1998). Therefore, our constructive identification results can be used to obtain estimators with better finite-sample properties than existing estimators.

REMARK 2.2. Unlike the models presented in Section 2.1, no restrictions are imposed on the dimensionality of the outcome. In particular, we do not need Y to be at least three-dimensional, as required by the framework of Allman et al. (2009).

REMARK 2.3. The variable W is required to affect the mixing weights (Assumption 2.4), but not the component distributions. This is the same exclusion restriction that was imposed in Henry et al. (2014); see Section 2.2 for a discussion of a range of economic applications where this condition may be plausibly justified. Note that Henry et al. (2014) only relied on the exclusion restriction on W and obtained a partial identification result. Here, we achieve point identification because of the additional assumptions on the Z covariates. Moreover, Assumption 2.4 is testable. Using, e.g. (A.5) evaluated at a candidate value of W, one can test the hypothesis that $(\partial/\partial w)\lambda(w)$ is zero by looking at $(\partial^2/\partial w \partial z_1)K(y^*|\mathbf{z}, w)$, which is estimable from the data.

REMARK 2.4. A sufficient condition for Assumption 2.3 is that, for a given value of X, $F_1(y|\mathbf{z}, x) = F_1(y|z_1, x)$ and $F_2(y|\mathbf{z}, x) = F_2(y|z_2, x)$. This constraint is similar in spirit to exclusion restrictions that are often imposed in discrete games of complete information in presence of multiple equilibria. For example, Tamer (2003) and Bajari et al. (2010) consider entry games where the exclusion restriction may be satisfied by variables that enter a firm's profit function, but not those of its competitors.

REMARK 2.5. Assumption 2.3 imposes restrictions on the CDFs of the component (conditional) distributions. These conditions may be reformulated using any other linear functional of the CDFs, such as the characteristic functions evaluated at a given point or the moments of a given order.

REMARK 2.6. Assumption 2.3 is essential for identification of the model. However, one can relax this restriction and still identify some of the parameters of interest. In Section 2.4.3, we consider the case where Assumption 2.3 holds for only one of the component CDFs and we show that the corresponding CDF is point identified under an additional assumption.

REMARK 2.7. Assumption 2.1 may be relaxed with minor changes in the statement of the theorem. Specifically, Section 2.4.2 discusses the case where $\mathbf{Z}$ and W have a discrete distribution.

2.4.1. Mixture model with an arbitrary number of components. The results in Theorem 2.1 extend to the case of mixtures with an arbitrary number of components. Consider the model

$$K(y|\mathbf{z}, w, x) = \sum_{j=1}^{J} \lambda_j(w, x) F_j(y|\mathbf{z}, x). \tag{2.7}$$

It is useful to rewrite this as

$$K(y|\mathbf{z}, w, x) = F_J(y|\mathbf{z}, x) + \sum_{j=1}^{J-1} \lambda_j(w, x)(F_j(y|\mathbf{z}, x) - F_J(y|\mathbf{z}, x)). \tag{2.8}$$

In the two-component case, we assumed that $\mathbf{Z}$ was a bivariate random variable. In order to achieve identification, we now need $\mathbf{Z}$ to be J-valued.

As in the bivariate mixture case, we make some assumptions that are not essential, but simplify the analysis.

ASSUMPTION 2.5. *The random variables Y, $\mathbf{Z}$ and W are continuously distributed, with support $\mathcal{Y}$, $\mathcal{Z}$ and $\mathcal{W}$, respectively.*

ASSUMPTION 2.6. *The functions $\lambda_j(\cdot, x)$ and $F_j(y|\cdot, x)$ for $j = 1, \ldots, J$, are differentiable for every $(x, y) \in \mathcal{X} \times R^p$.*

As in the two-component case, the key identifying assumptions concern the way in which $\mathbf{Z}$ and W affect the component CDFs and the mixing weights, respectively.

Concerning the component distributions, the generalization of Assumption 2.3 is given by the following.

ASSUMPTION 2.7. *For every $(\mathbf{z}, x) \in \mathcal{Z}\mathcal{X}$, there exists $y^* \in R^p$ such that, for $j = 1, \ldots, J$, $(\partial/\partial z_j)F_j(y^*|\mathbf{z}, x) \neq 0$ and $(\partial/\partial z_i)F_j(y^*|\mathbf{z}, x) = 0$ for all $i \neq j$.*

As for the restrictions imposed on the mixing weights, we first need some definitions. For a vector $\mathbf{w} = (w_1, \ldots, w_{J-1})$ in $\mathcal{W}^{J-1}$ and $x \in \mathcal{X}$, define $\Lambda(\mathbf{w}, x)$ as the $(J-1)$-by-$(J-1)$ matrix with its ijth element given by $\lambda_i(w_j, x)$. In other words, each column of $\Lambda(\mathbf{w}, x)$ contains the $J-1$ independent mixing weights corresponding to a given value of W. Similarly, define $D_w\Lambda(\mathbf{w}, x)$ as the $(J-1)$-by-$(J-1)$ matrix with its ijth element being $(\partial/\partial w)\lambda_i(w_j, x)$.

Given the definitions above, the generalization of Assumption 2.4 takes the following form.

ASSUMPTION 2.8. *For every $x \in \mathcal{X}$, there exists a vector $\mathbf{w}^* = (w_1^*, \ldots, w_{J-1}^*)$ in $\mathcal{W}^{J-1}$ such that $\Lambda(\mathbf{w}^*, x)$ and $D_w\Lambda(\mathbf{w}^*, x)$ are nonsingular.*

Assumption 2.8 requires W to affect the mixing weights to a sufficient degree and thus is analogous to the rank condition in instrumental variables models.

In this set-up, the model is point identified, as formalized below.

THEOREM 2.2. *Consider model (2.7) and let Assumptions 2.5–2.8 hold. Then, the component CDFs $F_j(y|\mathbf{z}, x)$, for $j = 1, \ldots, J$, are identified for every $(y, \mathbf{z}, x) \in \mathcal{Y}\mathcal{Z}\mathcal{X}$. Further, the matrix $\Lambda(\mathbf{w}, x)$ is identified for every $(x, \mathbf{w}) \in \mathcal{X}\mathcal{W}_x^{J-1}$, where $\mathcal{X}\mathcal{W}_x^{J-1} \equiv \{(x, \mathbf{w}) \in \mathcal{X}\mathcal{W}^{J-1}$ such that $\Lambda(\mathbf{w}, x)$ and $D_w\Lambda(\mathbf{w}, x)$ are nonsingular$\}$.*

2.4.2. Discrete covariates. A second extension of the basic results covers the case of discrete covariates.

For a function $g : R^d \to R$, let

$$\Delta_{x_i', x_i} g(x_1, \ldots, x_d) \equiv g(x_1, \ldots, x_i', \ldots, x_d) - g(x_1, \ldots, x_i, \ldots, x_d) \tag{2.9}$$

denote the difference with respect to the ith argument. We work under the following assumptions.

ASSUMPTION 2.9. *The variables $\mathbf{Z}$ and W have discrete distributions with support $\mathcal{Z}$ and $\mathcal{W}$, respectively.*

ASSUMPTION 2.10. *For every* $(\mathbf{z}, x) \in \mathcal{ZX}$, *there exists* $y^* \in \mathcal{Y}$ *and* $\mathbf{z}' \in \mathcal{Z}$ *such that* $\Delta_{z_1', z_1} F_1(y^*|\mathbf{z}, x) \neq 0$, $\Delta_{z_2', z_2} F_1(y^*|\mathbf{z}, x) = 0$, $\Delta_{z_1', z_1} F_2(y^*|\mathbf{z}, x) = 0$ *and* $\Delta_{z_2', z_2} F_2(y^*|\mathbf{z}, x) \neq 0$.

ASSUMPTION 2.11. *For every* $x \in \mathcal{X}$, *there exist* $w^*, w^{*'} \in \mathcal{W}$ *such that* $\Delta_{w^{*'}, w^*} \lambda(w^*, x) \neq 0$.

THEOREM 2.3. *Consider model (2.6) and let Assumptions 2.9, 2.10 and 2.11 hold. Then, for every* $(y, x, \mathbf{z}) \in \mathcal{YXZ}$, *the component CDFs* $F_i(y|\mathbf{z}, x)$, *for* $i = 1, 2$, *are nonparametrically identified and* $\lambda(w, x)$ *is nonparametrically identified for every* $(x, w) \in \mathcal{XW}^*_{x,d}$, *where* $\mathcal{XW}^*_{x,d} \equiv \{(x, w) \in \mathcal{XW}$ *such that* $\Delta_{w', w} \lambda(w^*, x) \neq 0$ *for some* $w' \in \mathcal{W}\}$. *Moreover, the results are constructive. Specifically,*

$$F_1(y|\mathbf{z}, x) = K(y|\mathbf{z}, w^*, x) - \frac{\Delta_{z_2', z_2} K(y^*|\mathbf{z}, w^*, x)}{\Delta_{w^{*'}, w^*} \Delta_{z_2', z_2} K(y^*|\mathbf{z}, w^*, x)} \Delta_{w^{*'}, w^*} K(y|\mathbf{z}, w^*, x)$$

$$F_2(y|\mathbf{z}, x) = K(y|\mathbf{z}, w^*, x) - \frac{\Delta_{z_1', z_1} K(y^*|\mathbf{z}, w^*, x)}{\Delta_{w^{*'}, w^*} \Delta_{z_1', z_1} K(y^*|\mathbf{z}, w^*, x)} \Delta_{w^{*'}, w^*} K(y|\mathbf{z}, w^*, x)$$

Further,

$$\lambda(w, x) = \frac{\zeta(w, x)}{1 + \zeta(w, x)}$$

where

$$\zeta(w, x) \equiv -\frac{\Delta_{z_1', z_1} K(y^*|\mathbf{z}, w, x) \Delta_{w^{*'}, w^*} \Delta_{z_2', z_2} K(y^*|\mathbf{z}, w, x)}{\Delta_{z_2', z_2} K(y^*|\mathbf{z}, w, x) \Delta_{w^{*'}, w^*} \Delta_{z_1', z_1} K(y^*|\mathbf{z}, w, x)}.$$

REMARK 2.8. Even when $\mathbf{Z}$ and W are continuously distributed, the result in Theorem 2.3 is potentially useful in implementing nonparametric estimation, as it allows the econometrician to avoid nonparametric estimation of derivatives, which have a notoriously slow rate of convergence. Note that applying the results above to the case where $\mathbf{Z}$ and W are continuous would still require maintaining Assumptions 2.10 and 2.11, which involve discrete differences.

2.4.3. Relaxing Assumption 2.3. The identification strategy for the baseline model crucially relies on Assumption 2.3. We now try to relax this restriction. In particular, we focus on the case in which Assumption 2.3 holds for one of the component CDFs, but may be violated for the other one. For fixed values of X and $\mathbf{Z}$, the model is again

$$K(y|\mathbf{z}, w, x) = \lambda(w, x) F_1(y|\mathbf{z}, x) + (1 - \lambda(w, x)) F_2(y|\mathbf{z}, x),$$

and we impose the following milder restriction.

ASSUMPTION 2.12. *For every* $(\mathbf{z}, x) \in \mathcal{ZX}$, *there exists* $y^* \in \mathcal{Y}$ *such that* $(\partial/\partial z_1) F_1(y^*|\,\mathbf{z}, x) \neq 0$ *and* $(\partial/\partial z_2) F_1(y^*|\mathbf{z}, x) = 0$.

No conditions are imposed on F_2 except for the standard requirements that define a CDF. In this setting, the model is not identified in general. However, under an additional restriction, it is possible to point identify the component CDF on which the conditions are imposed. The following formalizes this idea.

ASSUMPTION 2.13. *For every* $(\mathbf{z}, x) \in \mathcal{ZX}$, *the cross-derivative* $(\partial^2/\partial z_1 \partial z_2) F_1(y^*|\mathbf{z}, x)$ *exists and is nonzero.*

THEOREM 2.4. *Consider model (2.6) and let Assumptions 2.1, 2.2, 2.4, 2.12 and 2.13 hold. Then, $F_1(y|\mathbf{z}, x)$ is nonparametrically identified for every $y, \mathbf{z}, x \in R^p \times \mathcal{Z}\mathcal{X}$.*

2.4.4. Generalizing the mixture weight function λ. In this subsection we consider a model of the form

$$K(y|\mathbf{z}, w, x) = \lambda(\mathbf{z}, w, x)F_1(y|\mathbf{z}, x) + (1 - \lambda(\mathbf{z}, w, x))F_2(y|\mathbf{z}, x). \tag{2.10}$$

As before, $K(\cdot|\cdot, \cdot, \cdot)$ is the conditional distribution function of Y given $(\mathbf{Z}, W, X) = (\mathbf{z}, w, x)$, and is available from the joint distribution of the vector valued random variable $(Y, \mathbf{Z}, W, X)$, which is observable. The component distributions $F_1(\cdot|\cdot, \cdot)$, $F_2(\cdot|\cdot, \cdot)$ and the mixing weight $\lambda(\cdot, \cdot, \cdot)$ are unknown functions to be identified from the knowledge of $K(\cdot|\cdot, \cdot, \cdot)$. We maintain Assumption 2.1, thus $\mathbf{Z}$ and w are continuously distributed.

ASSUMPTION 2.14. *The functions $\lambda(\cdot, \cdot, x)$ and $F_i(y|\cdot, x)$ for $i = 1, 2$, are differentiable for every $(y, x) \in \mathcal{Y}\mathcal{X}$.*

As before, it is possible to drop the continuity/differentiability assumptions; see the discussion in Remark 2.10. The following additional assumptions are made. Recall $\mathcal{Y}$ denotes the support of Y.

ASSUMPTION 2.15. *For each $(x, w, \mathbf{z}) \in \mathcal{X}\mathcal{W}\mathcal{Z}$, there exist nonempty, nonsingleton sets $\mathcal{Y}_1(x, w, \mathbf{z})$ and $\mathcal{Y}_2(x, w, \mathbf{z})$ in $\mathcal{Y}$ such that (a) $(\partial/\partial z_2)F_1(\cdot|\mathbf{z}, x) = 0$ and $(\partial/\partial z_1)F_1(\cdot|\mathbf{z}, x) \neq 0$ on $\mathcal{Y}_1(x, w, \mathbf{z})$; (b) $(\partial/\partial z_1)F_2(\cdot|\mathbf{z}, x) = 0$ and $(\partial/\partial z_2)F_2(\cdot|\mathbf{z}, x) \neq 0$ on $\mathcal{Y}_2(x, w, \mathbf{z})$; (c) the functions $F_1(\cdot|\mathbf{z}, x) - F_2(\cdot|\mathbf{z}, x)$ and $(\partial/\partial z_1)F_1(\cdot|\mathbf{z}, x)$ are linearly independent on $\mathcal{Y}_1(x, w, \mathbf{z})$; (d) the functions $F_1(\cdot|\mathbf{z}, x) - F_2(\cdot|\mathbf{z}, x)$ and $(\partial/\partial z_2)F_2(\cdot|\mathbf{z}, x)$ are linearly independent on $\mathcal{Y}_2(x, w, \mathbf{z})$.*

As before, let $\mathcal{W}|(x, \mathbf{z})$ be the support of the conditional distribution of W given $(X, \mathbf{Z}) = (x, \mathbf{z})$.

ASSUMPTION 2.16. *For each $(x, \mathbf{z}) \in \mathcal{X}\mathcal{Z}$ there exists $w^* \in \mathcal{W}|(x, \mathbf{z})$ such that $(\partial/\partial w)$ $\lambda(\mathbf{z}, w^*, x) \neq 0$.*

The new model (2.10) differs from the original model (2.6) in that it allows the covariate $\mathbf{Z}$ to enter into the mixture weight λ in a completely free manner. In this aspect, therefore, it generalizes the model (2.6). Note, however, that the assumptions made for the model (2.6) and the current assumption (i.e. Assumptions 2.15 and 2.16) are non-nested, reflecting a major difference in the identification strategies for the two models. Therefore, neither of the two models is a special case of the other. For the current theorem, we assume that there is more than one value of y for which Assumptions 2.15(a) and (b) hold. However, Assumption 2.3 holding at one value of y suffices for Theorem 2.1.

Once again, the main application would be a model with unobserved heterogeneity or a model with multiple equilibria. For example, take the generic demand model described in Henry et al. (2014) (p. 124). If consumer preferences have two types, say a 'high' type and a 'low' type, each corresponding to F_1 and F_2, then as far as there exists a variable (e.g. geographical variable) that is correlated with the types but does not enter the utility functions, then such a variable works as w. Moreover, if there are covariates Z_1 and Z_2 such that each only affects each type of consumers, then they satisfy Assumption 2.15. The previous result in Theorem 2.1 was proved under the formulation (2.6), which, applied to the current example, would mean that individual types and the covariates $\mathbf{Z} = (Z_1, Z_2)$ are uncorrelated; this seems a strong restriction. The current theorem removes this restriction, by allowing for a fully unrestricted and nonparametric dependence of

the mixing weights λ on $\mathbf{Z}$. Alternatively, suppose we have a model with two equilibria at hand. If there exist covariates Z_1 and Z_2 such that Z_1 only affects one of the two equilibrium distributions and Z_2 the other, the model falls into the category discussed here. In this case, the function $\lambda(\cdot, \cdot, \cdot)$ determines the equilibrium selection mechanism, and the specification in this section allows the equilibrium selection mechanism to depend on the covariate $\mathbf{Z}$ in a fully general, nonparametric manner.

The following theorem shows that the component distribution functions $F_i(\cdot|\cdot, \cdot)$ and the mixture weight function $\lambda(\cdot, \cdot, \cdot)$ are nonparametrically identified. Define

$$\mathcal{K}_1(x, w, \mathbf{z}) := \Big\{(y_a, y_b) \in R^{2p} : \frac{\partial}{\partial w} K(y_a|\mathbf{z}, w, x) \frac{\partial^2}{\partial w \partial z_1} K(y_b|\mathbf{z}, w, x) - \frac{\partial^2}{\partial w \partial z_1} K(y_a|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_b|\mathbf{z}, w, x) \neq 0\Big\}$$

and

$$\mathcal{K}_2(x, w, \mathbf{z}) := \Big\{(y_c, y_d) \in R^{2p} : -\frac{\partial}{\partial w} K(y_c|\mathbf{z}, w, x) \frac{\partial^2}{\partial w \partial z_2} K(y_d|\mathbf{z}, w, x) + \frac{\partial^2}{\partial w \partial z_2} K(y_c|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_d|\mathbf{z}, w, x) \neq 0\Big\}.$$

It is straightforward to see that the sets $\mathcal{K}_1(x, w, \mathbf{z}) \cap (\mathcal{Y}_1(x, w, \mathbf{z}) \times \mathcal{Y}_1(x, w, \mathbf{z}))$ and $\mathcal{K}_2(x, w, \mathbf{z}) \cap (\mathcal{Y}_2(x, w, \mathbf{z}) \times \mathcal{Y}_2(x, w, \mathbf{z}))$ are nonempty under Assumptions 2.15 and 2.16. Define

$$\mathcal{XW}_x^* \equiv \Big\{(x, w) \in \mathcal{XW} : \frac{\partial}{\partial w} \lambda(\mathbf{z}, w, x) \neq 0\Big\}$$

and

$$\mathcal{XW}_x^* \mathcal{Z} \equiv \Big\{(x, w, \mathbf{z}) \in \mathcal{XWZ} : \frac{\partial}{\partial w} \lambda(\mathbf{z}, w, x) \neq 0\Big\}.$$

THEOREM 2.5. *Consider model* (2.10) *and let Assumptions 2.1, 2.15 and 2.16 hold. Then, the component CDFs* $F_i(y|\mathbf{z}, x)$*, for* $i = 1, 2$*, are nonparametrically identified for every* $(y, x, \mathbf{z}) \in R^p \times \mathcal{XZ}$ *and* $\lambda(\mathbf{z}, w, x)$ *is nonparametrically identified for every* $(\mathbf{z}, x, w) \in \mathcal{ZXW}_x^*$*. Moreover, the results are constructive. That is, for every* $(y, x, w, \mathbf{z}) \in R^p \times \mathcal{XW}_x^* \mathcal{Z}$*,*

$$F_1(y|\mathbf{z}, x) = K(y|\mathbf{z}, w, x) - \frac{-\frac{\partial}{\partial z_2} K(y_c|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_d|\mathbf{z}, w, x) + \frac{\partial}{\partial z_2} K(y_d|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_c|\mathbf{z}, w, x)}{-\frac{\partial}{\partial w} K(y_c|\mathbf{z}, w, x) \frac{\partial^2}{\partial w \partial z_2} K(y_d|\mathbf{z}, w, x) + \frac{\partial^2}{\partial w \partial z_2} K(y_c|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_d|\mathbf{z}, w, x)} \times \frac{\partial}{\partial w} K(y|\mathbf{z}, w, x),$$

$$F_2(y|\mathbf{z}, x) = K(y|\mathbf{z}, w, x) - \frac{-\frac{\partial}{\partial z_1} K(y_a|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_b|\mathbf{z}, w, x) + \frac{\partial}{\partial z_1} K(y_b|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_a|\mathbf{z}, w, x)}{\frac{\partial}{\partial w} K(y_a|\mathbf{z}, w, x) \frac{\partial^2}{\partial w \partial z_1} K(y_b|\mathbf{z}, w, x) - \frac{\partial^2}{\partial w \partial z_1} K(y_a|\mathbf{z}, w, x) \frac{\partial}{\partial w} K(y_b|\mathbf{z}, w, x)} \times \frac{\partial}{\partial w} K(y|\mathbf{z}, w, x),$$

and

$$\lambda(\mathbf{z}, w, x) = \frac{\zeta(\mathbf{z}, w, x)}{\zeta(\mathbf{z}, w, x) + 1},$$

where

$$\zeta(\mathbf{z}, w, x) := \frac{(-\frac{\partial}{\partial z_1} K(y_a|\mathbf{z}, w, x)\frac{\partial}{\partial w} K(y_b|\mathbf{z}, w, x) + \frac{\partial}{\partial z_1} K(y_b|\mathbf{z}, w, x)\frac{\partial}{\partial w} K(y_a|\mathbf{z}, w, x))}{(-\frac{\partial}{\partial z_2} K(y_c|\mathbf{z}, w, x)\frac{\partial}{\partial w} K(y_d|\mathbf{z}, w, x) + \frac{\partial}{\partial z_2} K(y_d|\mathbf{z}, w, x)\frac{\partial}{\partial w} K(y_c|\mathbf{z}, w, x))}$$
$$\times \frac{(-\frac{\partial}{\partial w} K(y_c|\mathbf{z}, w, x)\frac{\partial^2}{\partial w \partial z_2} K(y_d|\mathbf{z}, w, x) + \frac{\partial^2}{\partial w \partial z_2} K(y_c|\mathbf{z}, w, x)\frac{\partial}{\partial w} K(y_d|\mathbf{z}, w, x))}{(\frac{\partial}{\partial w} K(y_a|\mathbf{z}, w, x)\frac{\partial^2}{\partial w \partial z_1} K(y_b|\mathbf{z}, w, x) - \frac{\partial^2}{\partial w \partial z_1} K(y_a|\mathbf{z}, w, x)\frac{\partial}{\partial w} K(y_b|\mathbf{z}, w, x))}$$

if $(y_a, y_b) \in \mathcal{K}_1(x, w, \mathbf{z})$, $(y_c, y_d) \in \mathcal{K}_2(x, w, \mathbf{z})$, $y_a, y_b \in \mathcal{Y}_1(x, w, \mathbf{z})$ *and* $y_c, y_d \in \mathcal{Y}_2(x, w, \mathbf{z})$.

REMARK 2.9. Once again, the identification result is constructive. Note that the sets $\mathcal{K}_1(x, w, \mathbf{z})$ and $\mathcal{K}_2(x, w, \mathbf{z})$ are identifiable from the knowledge of the joint distribution of $(y, \mathbf{z}, w, x)$.

REMARK 2.10. Analogously to our treatment of discrete covariates in Theorem 2.3, it is possible to replace differential operators with difference operators to obtain nonparametric identification of the model (2.10) when $\mathbf{Z}$ and W are discrete random variables. Again, such a result can be useful in nonparametric estimation even when $\mathbf{Z}$ and W are continuously distributed (see the discussion in Remark 2.8).

3. GENERAL MIXTURES

Section 2 dealt with the case where the distribution of the latent variable is supported on a finite number of points. We now turn to the more general setting in which the latent variable is allowed to have infinite support. An important difference with respect to the case of finite mixtures is that now the dimension of the latent variable giving rise to the mixture plays a key role. The models considered below illustrate this point in greater detail. Attention is focused on four main settings: panel data models, mixed proportional hazard models, random coefficients models and fully nonparametric random utility models.

3.1. Panel data models

A class of models in which general mixtures are used to model unobserved heterogeneity is panel data models. While there is a vast literature on parametric and semiparametric panel data, the nonparametric treatment of these models has developed more recently and is still an active area of research. We focus on the setting considered by Evdokimov (2010). The model takes the form

$$y_{it} = m(x_{it}, \alpha_i) + \epsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{3.1}$$

where the scalar outcome Y is modelled as a function of observable covariates X, unobserved time-invariant heterogeneity α and unobserved idiosyncratic heterogeneity ϵ. Crucially, the function m is not specified parametrically and the time-invariant heterogeneity α is not assumed to enter the outcome equation additively. The latter feature allows for heterogeneous marginal effects of the covariates on the outcome Y across individuals with different values of α but

same value of X. In other words, observationally equivalent individuals are allowed to respond differently to a change in the covariates. The data in several microeconometric applications suggest that this may be the case, which makes this framework particularly relevant for applied work.[8] Note that the model can be interpreted both as a random effects model and as a fixed effects model, depending on whether α_i is assumed to be independent of X_i or not, respectively.

Under the random effects assumption, the function m is treated nonparametrically, with a normalization for the distribution of α. Therefore, the unknown parameters are m and the distribution of ϵ conditional on covariates. However, in the fixed effects case, the distribution of α given covariates can be identified and estimated from the data. This is especially relevant for policy analysis.

Evdokimov (2010) provides sufficient conditions for the nonparametric identification of the parameters of interest in each of these two settings. A key assumption in each case is the monotonicity of m in α.[9] Importantly, this imposes the implicit restriction that α be scalar-valued. While this requirement does limit the flexibility of the model, the presence of the idiosyncratic shock ϵ makes the framework suitable for modelling a variety of economic situations. For instance, if the goal is to estimate returns to education, ϵ could be thought of as the luck and measurement error component of wages, while α could be interpreted as (time-invariant) unobserved ability. Another important tool for identification is Kotlarski's Lemma: for random variables X_0, X_1 and X_2 with nonvanishing characteristic functions and $E[X_0] = m_0$ being finite, the characteristic functions of X_0, X_1 and X_2 are recovered from the joint distribution of $X_0 + X_1$ and $X_0 + X_2$, up to m_0, with convenient explicit formulas; see, e.g. Prakasa Rao (1983). With the additive structure with respect to ϵ, (3.1), applications of Kotlarski's Lemma are useful in terms of both identification and estimation.

Two other features make this model appealing to applied researchers. First, only two time periods ($T = 2$) are required for identification, which allows the results above to be used with very short panels. Second, the estimation procedure proposed by Evdokimov (2010) is easy to implement, as it avoids numerical optimization and only relies on computation of empirical CDFs and quantile functions.

Finally, it may be useful to compare the results in this paper to those in Kitamura (2003) discussed in Section 2.3. In both cases, the goal is to nonparametrically identify a regression model while accounting for unobserved heterogeneity via mixtures. However, Kitamura (2003) focuses on cross-sectional data and assumes the mixture is finite; each component distribution is a nonparametric object and therefore unobserved heterogeneity is function-valued (i.e. infinite-dimensional). Evdokimov (2010) considers panel data and allows the distribution of the latent variable to be continuous, and possibly with a fixed effects assumption. However, the unobserved heterogeneity α is assumed to be scalar-valued.

3.2. Mixed proportional hazard model

Another notable example where mixtures play a fundamental role is mixed proportional hazard models. These models are often used in labour economics to investigate the duration of unemployment spells. The standard set-up is as follows

$$\theta(t, x) = \phi(x)\psi(t)\alpha \tag{3.2}$$

[8] Evdokimov (2010) provides a number of references on this point.
[9] Strict monotonicity is required in the fixed effects case, while in the random effects case weak monotonicity suffices.

$$F(t|x) = 1 - \int_0^{\infty} \mathrm{e}^{-\phi(x)\int_0^t \psi(s)\,\mathrm{d}s\alpha}\,\mathrm{d}G(\alpha). \tag{3.3}$$

Equation (3.2) defines the hazard rate, i.e. the 'intensity' with which the event of interest – leaving unemployment, in the labour literature example – occurs. The hazard rate is a function of t, the time elapsed from a given initial point (e.g. the start of the unemployment spell), covariates X and unobserved (by the econometrician) α. Equation (3.3) states that the survival function is given by a function of the hazard rate and the distribution of α. Note that once again α is scalar-valued.

In single-spell duration models, the researcher is assumed to be able to recover the distribution $F(t|x)$ from the data. The unknown parameters are the functions ϕ and ψ, and the latent variable's CDF G. In their seminal contribution, Elbers and Ridder (1982) show that variation in the covariates X nonparametrically identifies all of the parameters of interest under regularity conditions. The important paper by Heckman and Singer (1984a) expands on these results in two main directions.

First, the authors provide an alternative set of assumptions that ensure identification. While these alternative restrictions are not globally weaker than the ones proposed by Elbers and Ridder (1982), the conditions imposed on the distribution of unobserved heterogeneity are indeed less restrictive. In particular, Heckman and Singer (1984a) allow the latent variable α not to have finite moments of any order, thus extending the analysis to fat-tail distributions.

Second, the authors show that it is possible to achieve identification even when the researcher does not have access to continuous covariates data. This result comes at the cost of imposing stronger parametric assumptions on the hazard rate. However, the latent variable distribution can still be treated nonparametrically. In a different paper, Heckman and Singer (1984b) propose a nonparametric ML procedure to estimate G.

3.3. Random coefficients models

3.3.1. Linear models. Random coefficient linear regression models have been widely used in estimating economic models incorporating unobserved heterogeneity. The basic model is

$$Y = \beta' \mathbf{X}, \tag{3.4}$$

where Y is a scalar, continuously distributed outcome, $\mathbf{X}$ is a d-dimensional vector of observed covariates and β is a random vector with density f_β. It is often assumed that $\mathbf{X} = (1, X_2 \ldots, X_d)$, so that the model can be rewritten as

$$y = \beta_2 x_2 + \ldots + \beta_d x_d + \epsilon, \tag{3.5}$$

with $\epsilon = \beta_1$. Equation (3.5) shows clearly that a random coefficients model can be viewed as a standard regression model in which the marginal effects of the covariates on the outcome are allowed to differ across individuals. This feature also characterized the models of Kitamura (2003) and Evdokimov (2010) discussed in previous sections (equations (2.4) and (3.1), respectively), though here β is multi-dimensional and continuously distributed. However, while in those models the regression function was treated nonparametrically, here we assume a linear relationship.

In this set-up, the structural unknown parameter becomes the distribution of the unobserved heterogeneity β. Traditionally, the problem has been tackled in one of two ways. One approach is to impose parametric assumptions (e.g. normality for β and sometimes independence across its elements as well). These restrictions greatly simplify the problem, but of course misspecification

becomes a serious concern. Alternatively, one may choose to rely on relatively mild assumptions on the relationship between covariates and unobserved heterogeneity, such as mean independence and conditional homoscedasticity of β given X; see, e.g. the survey by Hsiao and Pesaran (2004). Under these restrictions, the average marginal effect $E[\beta]$ and the variance $\text{Var}(\beta)$ are identified. However, as pointed out by Hoderlein et al. (2010), identification is not achieved for other important features of the distribution of β. For instance, the quantiles of the marginal distributions of β and properties depending on moments of order higher than two (such as skewness and kurtosis) are not uniquely identified from the data.

More recently, the literature in both statistics and econometrics has tried to find an approach alternative to the above two. The goal is to identify the entire distribution of β while avoiding parametric assumptions. As the discussion above makes clear, this requires stronger conditions than mean independence and conditional homoscedasticity. In particular, full independence between the random coefficients and the covariates $\mathbf{X}$ – or some instruments $\mathbf{Z}$ – suffices for nonparametric identification.

A key tool for nonparametric identification of the linear random coefficients regression is the theory of Radon transforms. The Radon transform $\mathcal{R}g$ of a function $g : R^d \to R$ is defined by

$$(\mathcal{R}g)(\mathbf{z}, w) = \int_{\mathbf{s}'\mathbf{z}=w} g(\mathbf{s})\, \mathrm{d}\mathbf{s}, \quad (\mathbf{z}, w) \in S^{d-1} \times R, \tag{3.6}$$

where S^{d-1} denotes a $(d-1)$-dimensional unit sphere in R^d. The above definition applied to the random coefficients model (3.4) leads to

$$f(y|\mathbf{x}) = \int_{\beta'\mathbf{x}=y} f_\beta(\beta)\mathrm{d}\beta = (\mathcal{R}f_\beta)(\mathbf{x}, y). \tag{3.7}$$

That is, under the assumption of independence between $\mathbf{X}$ and β, the Radon transform of f_β is equal to the conditional density of Y given $\mathbf{X}$. It is known that in (3.6) if g is in $L^1(R^d)$, then $\mathcal{R}g$ exists for almost all $(\mathbf{z}, w) \in S^{d-1} \times R$ and the map $g \to \mathcal{R}g$ is injective on $L^1(R^d)$; see Proposition 3.4 of Helgason (2009). The left-hand side of (3.7) can be recovered from the data, which suggests that one could apply the inverse Radon transform to an estimator of $f(y|\mathbf{x})$ in order to estimate the unknown f_β. However, the inverse of $\mathcal{R}$ is not a continuous operator and small changes in its argument may lead to big changes in the value of the resulting estimator. In other words, this is an ill-posed inverse problem. Some regularization is thus required.

Beran et al. (1996) work with characteristic functions and obtain a consistent and asymptotically normal estimator for f_β. However, Hoderlein et al. (2010) use (3.7) directly along with a regularized inverse of $\mathcal{R}$ with a kernel smoother; see Cavalier (2000) and Bissantz et al. (2014) for other applications of kernel smoothing in inverting the Radon operator. The estimator they propose achieves the optimal rate of convergence in a Sobolev class of functions. Further, they extend the model to allow for endogeneity of $\mathbf{X}$ and for nonlinearities in the way $\mathbf{X}$ affects the outcome Y. Masten (2014) studies nonparametric identification and estimation of the distribution of random coefficients in a linear simultaneous equation models. Hoderlein et al. (2015) explore the identifiability of the distribution of random coefficients in a linear triangular model and provide bounds when point identification is not available, together with semiparametric and nonparametric estimators.

3.3.2. Discrete choice models. The previous subsection analysed linear random coefficient models where the outcome variable was assumed to be continuously distributed. However, in many economic applications, one may be interested in modelling how a discrete variable

(e.g. the choice whether or not to buy a good) depends on covariates, and to allow for heterogeneity in this relation across individual observations.[10]

More specifically, the model we consider is

$$y = I\left\{\beta'\mathbf{x} \geq 0\right\}, \tag{3.8}$$

where Y is a binary outcome, $\mathbf{X} = (1, X_2, \ldots, X_d)$ is a d-dimensional vector of covariates and β is again assumed to be random with unknown density f_β.

As in the linear case, the literature has traditionally employed a parametric approach to the identification and estimation of the unobserved heterogeneity distribution. For instance, Train (2003) discusses the popular choice of mixed logit specifications. The nonparametric treatment of model (3.8) is a more recent development. An important contribution is the paper by Ichimura and Thompson (1998), who propose a nonparametric ML estimator (NPMLE) for the distribution of β. They provide a consistency proof for their NPMLE, which can be regarded as a generalization of Cosslett's NPMLE for binary choice models; see Cosslett (1983). Their procedure, however, is computationally very intensive, as it requires high-dimensional numerical optimization. Note that numerical optimization for MLE in mixture models can be difficult even when the model is tackled parametrically.

A recent paper by Gautier and Kitamura (2013) proposes a constructive nonparametric identification strategy leading to a plug-in estimator that avoids numerical optimization and integration, and thus is easy to implement and computationally parsimonious. In the basic model, the paper assumes that $\mathbf{X}$ and β are statistically independent. This allows one to write

$$r(\mathbf{x}) \equiv P(Y = 1|\mathbf{X} = \mathbf{x}) = E_\beta[I\{\beta'\mathbf{X} \geq 0\}]. \tag{3.9}$$

Further, since all that matters is the angle between $\mathbf{X}$ and β, without loss of generality we can normalize $\mathbf{X}$ and β so that they both belong to the unit sphere S^{d-1}. Letting the density f_β of β be defined with respect to the uniform spherical measure σ on S^{d-1}, we can then write

$$r(\mathbf{x}) = \int_{b\in S^{d-1}} I\{b'\mathbf{x} \geq 0\} f_\beta(b)\mathrm{d}\sigma(b) = \int_{b\in H(\mathbf{x})} f_\beta(b)\mathrm{d}\sigma(b) \equiv \mathcal{H}(f_\beta)(\mathbf{x}), \tag{3.10}$$

where $H(\mathbf{x}) \equiv \{b \in S^{d-1} : b'\mathbf{x} \geq 0\}$, which is a hemisphere, and the mapping $\mathcal{H}$ is called hemispherical transform.

We now see that analogously to the linear case considered in Section 3.3.1, we obtain an equation where the left-hand side is estimable from the data ($r(\mathbf{x})$ is the choice probability given $\mathbf{X} = \mathbf{x}$) and the right-hand side is the image of the unknown parameter f_β under the operator $\mathcal{H}$. However, in the binary choice case, the model is not identified unless further restrictions are imposed. The reason is that the operator $\mathcal{H}$ is in general not injective. In particular, Gautier and Kitamura (2013) show that the even part of the function f_β, denoted f_β^+, is not identified from (3.10).[11] The strategy adopted in the paper consists of two steps. First, conditions are provided under which f_β^- is identified. Second, it is shown that under certain assumptions one can find a one-to-one mapping between f_β^- and f_β, so that f_β is identified as well.

[10] For instance, Berry and Haile (2010) discuss nonparametric identification in discrete choice models with random coefficients which are often used in the industrial organization literature.

[11] The even part of a function g, denoted g^+, is defined as $g^+(x) = (g(x) + g(-x))/2$. Similarly, the odd part is defined as $g^-(x) = (g(x) - g(-x))/2$.

As far as the first step is concerned, a sufficient condition is that the nonconstant covariates $(X_2, \ldots, X_d)$ before normalization be supported on the entire space R^{d-1}. This rules out discrete regressors and regressors with bounded support, although Gautier and Kitamura (2013) also discuss a possible extension to limited-support covariates.

Turning to the second step, a sufficient condition that allows one to recover f_β from f_β^- is that the support of β is a subset of some hemisphere. More precisely, we assume that there exists a vector $\mathbf{c} \in S^{d-1}$ such that $P\{\beta'\mathbf{c} > 0\} = 1$, which is precisely the identification condition introduced by Ichimura and Thompson (1998). This assumption does not seem to be too restrictive in a number of economic applications. For instance, the condition is satisfied with $\mathbf{c}$ being a vector of zeros and one 1 if the researcher is willing to assume that one element of the β vector is positive. Note that knowledge of which specific element is positive is not required. A case where the sign of a coefficient may be reasonably assumed to be known is the effect of own price on demand for a good. Moreover, the support assumption on the distribution of β imposes constraints on f_β^-, which is identified under weak conditions in the first step. Therefore, the second-step assumption can be tested based on the observables.

Even assuming identification, estimation of f_β is a nontrivial task. Similarly to the linear random coefficient model, the reason is that the operator $\mathcal{H}$ in (3.10) has a discontinuous inverse. Thus, some regularization procedure is required. Indeed, Gautier and Kitamura (2013) show that the operator $\mathcal{H}$ is an analogue of a convolution operator in R^d, which implies that inverting it amounts to deconvolution. The formula for the estimator is obtained following the constructive identification proof. Further, they show that the estimator is consistent, they derive its rate of convergence and they provide a pointwise asymptotic normality proof.

The extension to the case where some of the covariates $\mathbf{X}$ are endogenous and suitable instruments are available is also discussed. On a related note, as pointed out by Hoderlein et al. (2010), the assumption of full independence between $\mathbf{X}$ and β is indeed strong, but it can be tested, at least in principle. One could split the support of $\mathbf{X}$ into two subregions and estimate f_β for each subregion separately. One could then compare the resulting estimates to obtain a specification test for the model. This procedure could be applied both to the linear model and to the discrete choice model.

Estimation of random coefficients distributions in more complicated models is of great interest. For example, Hoderlein et al. (2012) consider identification and estimation of the joint distribution of vector-valued random coefficients in various structural models under completeness conditions. Arellano and Bonhomme (2012) consider estimation of moments of the distribution of random coefficients conditional on observables (thus fixed effects) with panel data. See Graham and Powell (2012) for further developments in identification and estimation in such settings.

3.4. Random utility models, revealed preference and nonparametrics

Here, we discuss the analysis of random utility models using revealed preference restrictions, based on the analysis in Kitamura and Stoye (2013). The treatment is fully nonparametric and, as such, unobserved heterogeneity is infinite-dimensional.

In the models so far considered in Section 3, the latent variable giving rise to the mixture has been constrained to be vector-valued (and in some set-ups even scalar-valued). This is in contrast with the identification analysis for finite mixtures in Section 2, where often component

distributions are treated nonparametrically, corresponding to infinite-dimensional unobserved heterogeneity.

The analysis in Kitamura and Stoye (2013) concerns a fully nonparametric analysis of RUMs. In their analysis, for one thing, unobserved heterogeneity is represented by utility functions drawn from a general distribution. Therefore, it is clear that the unobserved heterogeneity has infinite dimension in this context, without assuming finiteness of the support of the distribution. Given the results presented so far in this paper, one would not expect to obtain point identification results under the described level of generality. Moreover, unlike standard nonparametric approaches, little smoothness conditions are imposed, and the focus is to obtain partial identification results.

In their set-up, the researcher has access to data on price $\mathbf{p} \in R^k$ (after normalizing income to be 1) and consumption choices $\mathbf{y} \in \mathcal{Y}$, where $\mathcal{Y} \subset R^k$ is a commodity space. Let $B(\mathbf{p})$ denote the budget set associated with $\mathbf{p}$. We maintain throughout this section that the number of budget sets (i.e. $\mathbf{p}_1, \ldots, \mathbf{p}_J$) is finite. Extending the results to the case of infinitely many budgets is possible at least on a theoretical level (see McFadden, 2005). Each individual is associated with a utility function u, which is drawn from the population distribution. As mentioned above, this is often used as a modelling device to account for heterogeneity in the preferences across individuals, though here the function u itself is regarded as a random parameter.

For a given price vector $\mathbf{p}$, the consumer choice $\mathbf{y}$ solves the standard maximization problem

$$\mathbf{y} \in \arg\max_{\mathbf{x} \in B(\mathbf{p})} u(\mathbf{x}). \tag{3.11}$$

The researcher can estimate choice probabilities $P\{\mathbf{y} \in A | \mathbf{p}\}$, for $A \subset \mathcal{Y}$, from the available data on prices and consumption choices. The question is then whether the estimated choice probabilities can be rationalized as an outcome of the RUM outlined above. In other words, the problem is to find a distribution over individual preferences (i.e. over utility functions) such that, in the aggregate, the behaviour implied by this distribution is consistent with the observed consumer choices. This is the essence of the stochastic revealed preference problem considered by McFadden and Richter (1991).[12] It can be viewed as a random utility version of the (deterministic) revealed preference theory pioneered by Samuelson (1938). The key difference is that, in the stochastic setting, preferences are allowed to be random draws from the population, which means that the object that the researcher is concerned with is the distribution over utility functions rather than their realized values.

Once the identified set for the distribution of the heterogeneous latent preferences has been obtained, the econometrician can use it to perform counterfactual analysis. Specifically, one can compute (bounds on) the expected choice probabilities in settings for which no observations are available. In the case of consumer demand, this amounts to estimating (bounds on) choice probabilities and expected demand for budget sets that are not in the data.

Traditionally, empirical analysis of the RUM often has been carried out parametrically. This approach imposes a functional form on the utility function of each consumer; probit/logit models and their variants have been popular choices in the literature. To allow for heterogeneity across consumers, the utility function is assumed to depend on a finite-dimensional

[12] McFadden and Richter (1991) show that a necessary and sufficient condition for the existence of such a distribution over preferences is the so-called axiom of revealed stochastic preferences (ARSP). Thus, the stochastic revealed preference problem may be reformulated as a test of the ARSP.

parameter, which is modelled as random and plays the role of the latent variable. A second parametric restriction is usually imposed on the distribution of the latent variable, which is assumed to belong to a parametric family or to be discrete; see, e.g. McFadden and Train (2000).

More recently, econometricians have sought approaches that do not rely on the restrictive parametric assumptions. As is often the case, more flexibility comes at a cost and most of the papers in this body of research only achieve partial identification of the unobserved heterogeneity distribution. However, in special settings, it may still be possible to obtain point identification, as in Cosslett (1983) and Matzkin (1992).

The RUM framework can be studied from a fully nonparametric standpoint. In particular, in Kitamura and Stoye (2013), no restrictions are imposed on the form of the utility function, except for the basic rationality and local nonsatiation conditions. As mentioned above, allowing for such a high degree of flexibility in general precludes point identification. However, it is still possible to test the existence of a distribution over consumer types (i.e. utilities) that rationalizes the observed choice behaviour. While the task of testing rationality at the individual level has been undertaken by many authors, usually they do not use the RUM formulation.

The procedure proposed by Kitamura and Stoye (2013) involves nuisance parameters with very high dimensions. Moreover, the implied inequalities take an 'indirect' form that makes it problematic to apply the rich literature on moment inequalities. To see this point in the simplest possible setting, consider a case with two (intersecting) budget lines (say budget 1 and 2) and two goods. Denote the segments that are on budget 1 and above (below) budget 2 by segment A (B), and those on budget 2 and above (below) budget 1 by segment C (D). A consumer's choice pattern is represented by a 4×1 vector, each element corresponding to segments A, B, C and D, respectively, and taking the value of 1 (0) if a consumption bundle on the segment is (not) chosen. For example, a consumer who chooses a bundle on segment A when facing budget 1 and a bundle on segment C when facing budget 2 is represented by $(1, 0, 1, 0)'$. The basic revealed preference theory implies that $(1, 0, 1, 0)'$, $(1, 0, 0, 1)'$ and $(1, 0, 0, 1)'$, labelled as a_1, a_2 and a_3, are valid whereas $(0, 1, 0, 1)'$ is not. Denoting the conditional choice probabilities for segments A and B given budget 1 by π_A and π_B, and defining π_C and π_D analogously, let $\pi = (\pi_A, \pi_B, \pi_C, \pi_D)'$. According to the McFadden–Richter theory, π is rationalizable (as an outcome of the RUM) if and only if it belongs to the cone spanned by a_1, a_2 and a_3, i.e.

$$\pi = \lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3, \quad \lambda_1, \lambda_2, \lambda_3 \geq 0. \tag{3.12}$$

As this example has a simple structure, it is easy to solve the indirect form of constraints (3.12) to obtain a system of direct inequalities in terms of the choice probability (i.e. $\pi_B + \pi_D \leq 1$), in addition to trivial constraints automatically satisfied by every valid conditional probability vector. The condition $\pi_B + \pi_D \leq 1$ is a special case of a moment inequality. This operation of obtaining a moment inequality from the expression (3.12) is theoretically guaranteed by Weyl's theorem; see, e.g. Ziegler (1995). Unfortunately, implementing it in a problem with an empirically relevant scale is computationally prohibitive. Therefore, one cannot rely on standard methods to obtain critical values, such as moment selection. Kitamura and Stoye (2013) thus employ a modified bootstrap procedure to compute the critical values for the test. The tools that are developed may also be used to perform policy evaluation and prediction, and obtain identified regions for various counterfactuals.

4. CONCLUSION

This paper considered several recent developments in the econometrics literature on mixture models. The emphasis has been placed on how these advances can contribute to the usefulness and flexibility of mixtures in a variety of economics applications. The applications in which these results may be employed include: multiple equilibria in discrete games (such as games of entry or technology adoption), analysis of panel data, random utility models and consumer behaviour analysis, measurement error, switching regression models and proportional hazard models.

We also presented some new results on nonparametric identification of finite mixture models under exclusion restrictions. Our identification strategy requires very mild conditions on the support of the covariates. This is in contrast with the identification at infinity approach often used in the existing literature. It is shown that the basic results can be extended to deal with a mixture model with an arbitrary number of components and models with discrete covariates.

ACKNOWLEDGEMENTS

Y. Kitamura acknowledges financial support from the National Science Foundation via grant SES-1156266. The authors thank Tim Armstrong, Phil Haile, Arthur Lewbel, Lorenzo Magnolfi, Camilla Roncoroni, participants at the *Econometrics Journal* Special Session on heterogeneity at the 2013 Royal Economic Society Conference and the National University of Singapore and, in particular, Lars Nesheim for very helpful comments.

REFERENCES

Abbring, J. H. and G. J. van den Berg (2003). The non-parametric identification of treatment effects in duration models. *Econometrica 71*, 1491–517.

Allman, E., C. Matias and J. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics 37*, 3099–132.

Andrews, D. W. and M. M. A. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies 65*, 497–517.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *Review of Economic Studies 79*, 987–1020.

Arellano, M., R. Blundell and S. Bonhomme (2014). Household earnings and consumption: a nonlinear framework. Discussion paper.

Bajari, P., H. Hong and S. P. Ryan (2010). Identification and estimation of a discrete game of complete information. *Econometrica 78*, 1529–68.

Beran, R., A. Feuerverger and P. Hall (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Annals of Statistics 24*, 2569–692.

Berry, S. and P. Haile (2010). Identification of a heterogeneous generalized regression model with group effects. Cowles Foundation Discussion Paper 1732, Yale University.

Berry, S. and E. Tamer (2006). Identification in models of oligopoly entry. In R. Blundell, W. K. Newey and T. Persson (Eds.), *Advances in Economics and Econometrics. Ninth World Congress, Volume 2*, 46–85. Cambridge: Cambridge University Press.

Berry, S., M. Carnall and P. Spiller (2006). Airline hubbing, costs and demand. In D. Lee (Ed.), *Advances in Airline Economics, Volume 1: Competition Policy and Anti-Trust*, 183–214. Amsterdam: Elsevier.

Bissantz, N., H. Holzmann and K. Proksch (2014). Confidence regions for images observed under the radon transform. *Journal of Multivariate Analysis 128*, 86–107.

Bollinger, C. R. (2006). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics 73*, 387–99.

Bonhomme, S. and J.-M. Robin (2014). Generalized non-parametric deconvolution with an application to earnings dynamics. *Review of Economic Studies 77*, 491–533.

Bonhomme, S., K. Jochmans and J.-M. Robin (2016a). Estimating multivariate latent-structure models. *Annals of Statistics 44*, 540–63.

Bonhomme, S., K. Jochmans and J.-M. Robin (2016b). Nonparametric estimation of finite mixtures from repeated measurements. *Journal of Royal Statistical Society*, Series B *78*, 211–29.

Bound, J., C. Brown and N. Mathiowetz (2001). Measurement error in survey data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 5*, 3705–3843. Amsterdam: North-Holland.

Cameron, S. V. and J. J. Heckman (1998). Life cycle schooling and dynamic selection bias: models and evidence for five cohorts. *Journal of Political Economy 106*, 262–311.

Carroll, R. J., D. Ruppert, L. A. Stefanski and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, (2nd ed.). London: Chapman and Hall/CRC.

Cavalier, L. (2000). Efficient estimation of a density in a problem of tomography. *Annals of Statistics*, 630–47.

Chamberlain, G. (1986). Asymptotic efficiency in semiparametric models with censoring. *Journal of Econometrics 32*, 189–218.

Chen, X., H. Hong and D. Nekipelov (2011). Nonlinear models of measurement errors. *Journal of Economic Literature 49*, 901–37.

Cho, J. S. and H. White (2007). Testing for regime switching. *Econometrica 75*, 1671–720.

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in airline markets. *Econometrica 77*, 1791–828.

Cooper, R. W. (2002). Estimation and identification of structural parameters in the presence of multiple equilibria. *Annales d'Economie et de Statistique 6*, 1–26.

Cosslett, S. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica 51*, 765–82.

Cosslett, S. R. and L.-F. Lee (1985). Serial correlation in latent discrete variable models. *Journal of Econometrics 27*, 79–97.

Elbers, C. and G. Ridder (1982). True and spurious duration dependence: the identifiability of the proportional hazard model. *Review of Economic Studies 49*, 403–09.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data models with unobserved heterogeneity. Working paper, Princeton University.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Berlin: Springer.

Gautier, E. and Y. Kitamura (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica 81*, 581–607.

Graham, B. and J. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel model. *Econometrica 80*, 2105–52.

Hall, P. and X. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics 31*, 201–24.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–84.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *Journal of Economic Perspectives 15*, 57–67.

Heckman, J. J. and B. Singer (1984a). The identifiability of the proportional hazard model. *Review of Economic Studies 51*, 231–41.

Heckman, J. J. and B. Singer (1984b). A method of minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*, 271–320.

Helgason, S. (2009). *Integral Geometry and Radon Transforms*. Berlin: Springer.

Henry, M., K. Jochmans and B. Salanié (2013). Inference on mixtures under tail restrictions. Working paper, Columbia University, New York.

Henry, M., Y. Kitamura and B. Salanié (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics 5*, 123–44.

Hoderlein, S., J. Klemelä and E. Mammen (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory 26*, 804–37.

Hoderlein, S., L. Nesheim and A. Simoni (2012). Semiparametric estimation of random coefficients in structural economic models. CWP 09/12, Centre for Microdata Methods and Practice, Insitute for Fiscal Studies and University College London.

Hoderlein, S., H. Holzmann and A. Meister (2015). The triangular model with random coefficients. CWP 33/15, Centre for Microdata Methods and Practice, Insitute for Fiscal Studies and University College London.

Hohmann, D. and H. Holzmann (2013). Two-component mixtures with independent coordinates as conditional mixtures: nonparametric identification and estimation. *Electronic Journal of Statistics 7*, 859–80.

Horowitz, J. L. and C. Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica 63*, 281–302.

Hsiao, C. and M. H. Pesaran (2004). Random coefficient panel data models. IZA Discussion Paper 1236, Institute for the Study of Labor, Bonn.

Hu, Y., D. McAdams and M. Shum (2013). Identification of first-price auctions with non-separable unobserevd heterogeneity. *Journal of Econometrics 174*, 186–93.

Ichimura, H. and T. S. Thompson (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics 86*, 269–95.

Kasahara, H. and K. Shimotsu (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica 77*, 135–75.

Kasahara, H. and K. Shimotsu (2011). Sequential estimation of dynamic programming models with unobserved heterogeneity. Working paper, Graduate School of Economics, Hitotsubashi University.

Keane, M. P. and K. I. Wolpin (1997). The career decisions of young men. *Journal of Political Economy 105*, 473–522.

Kitamura, Y. (2003). Nonparametric identifiability of finite mixtures. Working paper, Yale University.

Kitamura, Y. and J. Stoye (2013). Nonparametric analysis of random utility models: testing. Working paper, Yale University.

Kruskal, J. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. *Linear Algebra and Its Applications 18*, 95–138.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association 73*, 805–11.

Lewbel, A. and X. Tang (2015). Identification and estimation of games with incomplete information using excluded regressors. *Journal of Econometrics 189*, 229–44.

Lindsay, B. G. (1988). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5*, 1–163. Beachwood, OH: Institute of Mathematical Statistics.

McFadden, D. L. (2005). Revealed stochastic preference: a synthesis. *Economic Theory 26*, 245–64.

McFadden, D. and K. Richter (1991). Stochastic rationality and revealed stochastic preference. In J. Chipman, D. McFadden and K. Richter (Eds.), *Preferences, Uncettainty and Rationality*, 161–186. Boulder, CO: Westview Press.

McFadden, D. L. and K. Train (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics 15*, 447–70.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York, NY: Wiley.

Mahajan, A. (2006). Identification and estimation of regression models with misclassification. *Econometrica 74*, 631–65.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Berlin: Springer.

Masten, M. (2014). Random coefficients on endogenous variables in simultaneous equations models. CWP 01/14, Centre for Microdata Methods and Practice, Insitute for Fiscal Studies and University College London.

Matzkin, R. (1992). Nonparametric and distribution free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*, 239–70.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics 144*, 81–117.

Prakasa Rao, B. (1983). *Nonparametric Functional Estimation*. New York, NY: Academic Press.

Qin, J. (1998). Semiparametric likelihood based method for goodness of fit tests and estimation in upgraded mixture models. *Scandinavian Journal of Statistics 25*, 681–91.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics 27*, 1368–84.

Samuelson, P. (1938). A note on the pure theory of consumer's behaviour. *Economica 5*, 61–71.

Schennach, S. M. (2013). Measurement error in nonlinear models – a review. In D. A. Acemoglu, M. Arellano and E. Dekel (Eds.), *Advances in Economics and Econometrics, Tenth World Congress, Volume III: Econometrics*, 296–337. Cambridge: Cambridge University Press.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies 70*, 147–65.

Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Ziegler, G. M. (1995). *Lectures on Polytopes, Volume 152*. Berlin: Springer.

Zou, F., J. P. Fine and B. S. Yandell (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika 89*, 61–75.

APPENDIX

Proof of Theorem 2.1: Because the entire proof is conditional on $X = x$, we drop this conditioning for notational convenience.

Equation (2.6) can be rewritten as

$$K(y|\mathbf{z}, w) = F_2(y|\mathbf{z}) + \lambda(w)(F_1(y|\mathbf{z}) - F_2(y|\mathbf{z})). \tag{A.1}$$

Differentiating (A.1) with respect to w, we obtain

$$\frac{\partial}{\partial w} K(y|\mathbf{z}, w) = \frac{\partial}{\partial w}\lambda(w)(F_1(y|\mathbf{z}) - F_2(y|\mathbf{z})) \tag{A.2}$$

for every $y \in \mathcal{Y}$. Moreover, if $y = y^*$, we can use Assumption 2.3 to obtain

$$\frac{\partial}{\partial z_1} K(y^*|\mathbf{z}, w) = \lambda(w)\frac{\partial}{\partial z_1} F_1(y^*|\mathbf{z}) \tag{A.3}$$

and

$$\frac{\partial}{\partial z_2} K(y^*|\mathbf{z}, w) = (1 - \lambda(w)) \frac{\partial}{\partial z_2} F_2(y^*|\mathbf{z}). \tag{A.4}$$

Differentiating (A.3) and (A.4) with respect to w, we obtain, respectively,

$$\frac{\partial^2}{\partial w \partial z_1} K(y^*|\mathbf{z}, w) = \frac{\partial}{\partial w} \lambda(w) \frac{\partial}{\partial z_1} F_1(y^*|\mathbf{z}) \tag{A.5}$$

and

$$\frac{\partial^2}{\partial w \partial z_2} K(y^*|\mathbf{z}, w) = -\frac{\partial}{\partial w} \lambda(w) \frac{\partial}{\partial z_2} F_2(y^*|\mathbf{z}). \tag{A.6}$$

By (A.3) and (A.4), we obtain

$$\frac{(\partial/\partial z_1) K(y^*|\mathbf{z}, w)}{(\partial/\partial z_2) K(y^*|\mathbf{z}, w)} = \frac{\lambda(w)}{1 - \lambda(w)} \cdot \frac{(\partial/\partial z_1) F_1(y^*|\mathbf{z})}{(\partial/\partial z_2) F_2(y^*|\mathbf{z})} \tag{A.7}$$

and by (A.5) and (A.6), for every $w \in \mathcal{W}_x^*$, we obtain

$$\frac{(\partial^2/\partial w \partial z_1) K(y^*|\mathbf{z}, w)}{(\partial^2/\partial w \partial z_2) K(y^*|\mathbf{z}, w)} = -\frac{(\partial/\partial z_1) F_1(y^*|\mathbf{z})}{(\partial/\partial z_2) F_2(y^*|\mathbf{z})}. \tag{A.8}$$

Combining (A.7) and (A.8) yields

$$\frac{\lambda(w)}{1 - \lambda(w)} = -\frac{(\partial/\partial z_1) K(y^*|\mathbf{z}, w)}{(\partial/\partial z_2) K(y^*|\mathbf{z}, w)} \cdot \frac{(\partial^2/\partial w \partial z_2) K(y^*|\mathbf{z}, w)}{(\partial^2/\partial w \partial z_1) K(y^*|\mathbf{z}, w)} \equiv \zeta(w), \tag{A.9}$$

for every $w \in \mathcal{W}_x^* \equiv \{w \in \mathcal{W} : (\partial/\partial w)\lambda(w, x) \neq 0\}$. Therefore,

$$\lambda(w) = \frac{\zeta(w)}{1 + \zeta(w)} \tag{A.10}$$

for all $w \in \mathcal{W}_x^*$. Because ζ is only a function of observables, (A.10) shows that the mixing weight $\lambda(w)$ is identified for all $w \in \mathcal{W}_x^*$.

We now turn to showing identification of the component CDFs. The following relationship is useful:

$$\frac{\lambda(w)}{(\partial/\partial w)\lambda(w)} = \frac{(\partial/\partial z_1) K(y^*|\mathbf{z}, w)}{(\partial^2/\partial w \partial z_1) K(y^*|\mathbf{z}, w)}. \tag{A.11}$$

Equation (A.11) follows from (A.3) and (A.5).

Now let $w^* \in \mathcal{W}_x^*$ and let y be an arbitrary vector in $\mathcal{Y}$. By (A.1) and (A.2),

$$K(y|\mathbf{z}, w^*) = F_2(y|\mathbf{z}) + \frac{\lambda(w^*)}{(\partial/\partial w)\lambda(w^*)} \cdot \frac{\partial}{\partial w} K(y|\mathbf{z}, w^*), \tag{A.12}$$

which, by (A.11), implies

$$F_2(y|\mathbf{z}) = K(y|\mathbf{z}, w^*) - \frac{(\partial/\partial z_1) K(y^*|\mathbf{z}, w^*)}{(\partial^2/\partial w \partial z_1) K(y^*|\mathbf{z}, w^*)} \cdot \frac{\partial}{\partial w} K(y|\mathbf{z}, w^*). \tag{A.13}$$

Because the value y was arbitrary, (A.13) shows that F_2 is identified for every $y \in \mathcal{Y}$, given $Z = z$. F_1 can be treated symmetrically. □

Proof of Theorem 2.2: Because the entire proof is conditional on $X = x$, we drop the conditioning on $X = x$ for ease of exposition.

We first need some notation. For any $\mathbf{w} \in \mathcal{W}^{J-1}$, let $\mathbf{K}(y|\mathbf{z}, \mathbf{w}) \equiv (K(y|\mathbf{z}, w_1), \ldots, K(y|\mathbf{z}, w_{J-1}))'$. Let D_q denote the differential operator with respect to variable q, so that

$$D_w \mathbf{K}(y|\mathbf{z}, \mathbf{w}) \equiv \left(\frac{\partial}{\partial w} K(y|\mathbf{z}, w_1), \ldots, \frac{\partial}{\partial w} K(y|\mathbf{z}, w_{J-1}) \right)'.$$

Similarly, let $D_{\mathbf{z}_{-J}} \mathbf{K}(y|\mathbf{z}, \mathbf{w})$ be the $(J-1)$-by-$(J-1)$ matrix with ijth element $(\partial/\partial \mathbf{z}_j) K(y|\mathbf{z}, w_i)$, i.e. each row contains the derivatives with respect to the different elements of $\mathbf{Z}_{-J} \equiv (Z_1, \ldots, Z_{J-1})'$ for a given value of W, and let $D_{w\mathbf{z}_{-J}} \mathbf{K}(y|\mathbf{z}, \mathbf{w})$ be the $(J-1)$-by-$(J-1)$ matrix with ijth element $(\partial^2/\partial w \partial z_j) K(y|\mathbf{z}, w_i)$. Analogously, let $D_{z_J} \mathbf{K}(y|\mathbf{z}, \mathbf{w})$ be the $(J-1)$-dimensional vector with jth element $(\partial/\partial z_J) K(y|\mathbf{z}, w_j)$ and let $D_{wz_J} \mathbf{K}(y|\mathbf{z}, \mathbf{w})$ be the $(J-1)$-dimensional vector with jth element $(\partial^2/\partial w \partial z_J)$ $K(y|\mathbf{z}, w_j)$.

Finally, let $\mathbf{F}(y|\mathbf{z}) \equiv (F_1(y|\mathbf{z}) - F_J(y|\mathbf{z}), \ldots, F_{J-1}(y|\mathbf{z}) - F_J(y|\mathbf{z}))'$ and $\mathbf{F}_J(y|\mathbf{z}) \equiv F_J(y|\mathbf{z}) \cdot \mathbf{1}$, where $\mathbf{1}$ is the $(J-1)$-dimensional column vector of ones. Using the same notation as above, we have that $D_{\mathbf{z}_{-J}} \mathbf{F}(y|\mathbf{z})$ is the $(J-1)$-by-$(J-1)$ matrix with ijth element $(\partial/\partial \mathbf{z}_j)(F_i(y|\mathbf{z}) - F_J(y|\mathbf{z}))$ and $D_{z_J} \mathbf{F}_J(y|\mathbf{z})$ is the $(J-1)$-dimensional column vector $(\partial/\partial z_J) F_J(y|\mathbf{z}) \cdot \mathbf{1}$.

Given the notation above, we can write model (2.8) as

$$\mathbf{K}(y|\mathbf{z}, \mathbf{w}) = \Lambda(\mathbf{w})' \mathbf{F}(y|\mathbf{z}) + \mathbf{F}_J(y|\mathbf{z}). \tag{A.14}$$

Differentiating (A.14) with respect to W and evaluating at $\mathbf{w}^*$ yields

$$D_w \mathbf{K}(y|\mathbf{z}, \mathbf{w}^*) = D_w \Lambda(\mathbf{w}^*)' \mathbf{F}(y|\mathbf{z}). \tag{A.15}$$

Now we evaluate (A.14) at y^*. Differentiating with respect to $\mathbf{Z}_{-J}$ and using Assumption 2.7, we obtain

$$D_{\mathbf{z}_{-J}} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*) = \Lambda(\mathbf{w}^*)' D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z}). \tag{A.16}$$

Similarly, differentiation with respect to Z_J yields

$$D_{z_J} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*) = (\mathbf{I} - \Lambda(\mathbf{w}^*))' D_{z_J} \mathbf{F}_J(y^*|\mathbf{z}), \tag{A.17}$$

where $\mathbf{I}$ is the $(J-1)$-dimensional identity matrix. Differentiating (A.16) and (A.17) again with respect to W leads to

$$D_{w\mathbf{z}_{-J}} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*) = D_w \Lambda(\mathbf{w}^*)' D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z}), \tag{A.18}$$

and

$$D_{wz_J} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*) = -D_w \Lambda(\mathbf{w}^*)' D_{z_J} \mathbf{F}_J(y^*|\mathbf{z}). \tag{A.19}$$

Note that, by Assumption 2.7, the matrix $D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z})$ is a diagonal matrix with jjth entry $(\partial/\partial z_j) F_j(y^*|\mathbf{z}) \neq 0$, for $j = 1, \ldots, J-1$, and hence it is invertible. Thus, from (A.16) it follows

$$\Lambda(\mathbf{w}^*) = (D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z})')^{-1} D_{\mathbf{z}_{-J}} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*)', \tag{A.20}$$

which shows that, if $D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z})$ is identified, then $\Lambda(\mathbf{w}^*)$ is identified as well.

Further, (A.18) and (A.19) imply

$$(D_{w\mathbf{z}_{-J}} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*))^{-1} D_{wz_J} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*) = -(D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z}))^{-1} D_{z_J} \mathbf{F}_J(y^*|\mathbf{z}). \tag{A.21}$$

Note that the matrix $D_{w\mathbf{z}_{-J}} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*)$ is invertible because $D_w \Lambda(\mathbf{w}^*)$ is invertible by Assumption 2.8 and $D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z})$ is invertible because it is full-rank, as discussed above.

Together with the fact that $D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z})$ is a diagonal matrix, (A.21) shows that, if $D_{z_J} \mathbf{F}_J(y^*|\mathbf{z})$ were known, then $D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z})$ would be identified.

Now, (A.17) and (A.20) imply

$$D_{z_J} \mathbf{F}_J(y^*|\mathbf{z}) = D_{z_J} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*) + D_{\mathbf{z}_{-J}} \mathbf{K}(y^*|\mathbf{z}, \mathbf{w}^*)(D_{\mathbf{z}_{-J}} \mathbf{F}(y^*|\mathbf{z}))^{-1} D_{z_J} \mathbf{F}_J(y^*|\mathbf{z}), \tag{A.22}$$

which, by (A.21), becomes

$$D_{z_J}\mathbf{F}_J(y^*|\mathbf{z}) = D_{z_J}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*) - D_{\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*) \times (D_{w\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*))^{-1} D_{wz_J}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*). \tag{A.23}$$

Equation (A.23) shows that $D_{z_J}\mathbf{F}_J(y^*|\mathbf{z})$ is identified. In fact, because every element of the vector $D_{z_J}\mathbf{F}_J(y^*|\mathbf{z})$ is equal to $(\partial/\partial z_J)F_J(y^*|\mathbf{z})$, the latter partial derivative is identified by $J-1$ distinct equations.

As noted above, identification of $D_{z_J}\mathbf{F}_J(y^*|\mathbf{z})$ implies identification of $D_{\mathbf{z}_{-J}}\mathbf{F}(y^*|\mathbf{z})$, which in turn implies identification of $\Lambda(\mathbf{w}^*)$.

We now show how to recover the component CDFs. Combining (A.16) and (A.18), we obtain

$$D_{\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*)(D_{w\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*))^{-1} = \Lambda(\mathbf{w}^*)'(D_w\Lambda(\mathbf{w}^*)')^{-1}. \tag{A.24}$$

Now, given that $\Lambda(\mathbf{w}^*)$ is identified as shown above, we can use (A.15) and (A.24) to obtain

$$\mathbf{F}(y|\mathbf{z}) = (\Lambda(\mathbf{w}^*)')^{-1} D_{\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*)(D_{w\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*))^{-1} D_w\mathbf{K}(y|\mathbf{z},\mathbf{w}^*), \tag{A.25}$$

which shows that $\mathbf{F}(y|\mathbf{z})$ is identified. Note that the matrix $\Lambda(\mathbf{w}^*)$ is invertible by Assumption 2.8.

Finally, using (A.14) and (A.25), we obtain

$$\mathbf{F}_J(y|\mathbf{z}) = \mathbf{K}(y|\mathbf{z},\mathbf{w}^*) - D_{\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*)(D_{w\mathbf{z}_{-J}}\mathbf{K}(y^*|\mathbf{z},\mathbf{w}^*))^{-1} D_w\mathbf{K}(y|\mathbf{z},\mathbf{w}^*). \tag{A.26}$$

□

Proof of Theorem 2.3: The proof proceeds along the same lines as the proof of Theorem 2.1 with the difference operator defined in (2.9) replacing the differential operator. □

Proof of Theorem 2.4: Again, given that the entire analysis is conditional on $X = x$, we omit this conditioning for brevity. We can rewrite (2.6) as

$$K(y|\mathbf{z},w^*) = F_1(y|\mathbf{z}) + (1-\lambda(w^*))(F_2(y|\mathbf{z}) - F_1(y|\mathbf{z})), \tag{A.27}$$

which shows that, if $(1-\lambda(w^*))(F_2(y|\mathbf{z}) - F_1(y|\mathbf{z}))$ is identified, then $F_1(y|\mathbf{z})$ is identified as well. Differentiating (2.6) and using Assumption 2.12,

$$\frac{\partial}{\partial z_2}K(y^*|\mathbf{z},w^*) = (1-\lambda(w^*))\frac{\partial}{\partial z_2}F_2(y^*|\mathbf{z}). \tag{A.28}$$

Differentiating (A.28) with respect to z_1,

$$\frac{\partial^2}{\partial z_1\partial z_2}K(y^*|\mathbf{z},w^*) = (1-\lambda(w^*))\frac{\partial^2}{\partial z_1\partial z_2}F_2(y^*|\mathbf{z}). \tag{A.29}$$

The cross-derivative exists and is nonzero by Assumption 2.13.

Moreover, differentiating (A.29) with respect to w,

$$\frac{\partial^3}{\partial z_1\partial z_2\partial w}K(y^*|\mathbf{z},w^*) = -\frac{\partial}{\partial w}\lambda(w^*)\cdot\frac{\partial^2}{\partial z_1\partial z_2}F_2(y^*|\mathbf{z}). \tag{A.30}$$

Combining (A.29) and (A.30),

$$\frac{(\partial^3/\partial z_1\partial z_2\partial w)K(y^*|\mathbf{z},w^*)}{(\partial^2/\partial z_1\partial z_2)K(y^*|\mathbf{z},w^*)} = -\frac{(\partial/\partial w)\lambda(w^*)}{(1-\lambda(w^*))}, \tag{A.31}$$

which shows that the right-hand side of (A.31) is identified.

Now, differentiating (2.6) with respect to w yields

$$\begin{aligned}\frac{\partial}{\partial w}K(y|\mathbf{z}, w^*) &= -\frac{\partial}{\partial w}\lambda(w^*)(F_2(y|\mathbf{z}) - F_1(y|\mathbf{z})) \\ &= -\frac{(\partial/\partial w)\lambda(w^*)}{(1-\lambda(w^*))}\cdot(1-\lambda(w^*))(F_2(y|\mathbf{z}) - F_1(y|\mathbf{z})). \end{aligned} \tag{A.32}$$

Combining (A.31) and (A.32) shows that $(1-\lambda(w^*))(F_2(y|\mathbf{z}) - F_1(y|\mathbf{z}))$ is identified, which implies that $F_1(y|\mathbf{z})$ is identified, as claimed at the outset.

The final formula is

$$F_1(y|\mathbf{z}) = K(y|\mathbf{z}, w^*) - \frac{\partial}{\partial w}K(y|\mathbf{z}, w^*)\cdot\frac{(\partial^2/\partial z_1\partial z_2)K(y^*|\mathbf{z}, w^*)}{(\partial^3/\partial z_1\partial z_2\partial w)K(y^*|\mathbf{z}, w^*)}. \tag{A.33}$$

□

Proof of Theorem 2.5: Fix $(x, w, \mathbf{z}) \in \mathcal{X}\mathcal{W}^*_x\mathcal{Z}$. As in the proof of Theorem 2.1, all the arguments remain valid conditional on $X = x$, and we drop x from the notation. Letting $F_\Delta(y^*|\mathbf{z}) := F_1(y|\mathbf{z}) - F_2(y|\mathbf{z})$, (2.10) becomes

$$K(y|\mathbf{z}, w) = F_2(y|\mathbf{z}) + \lambda(\mathbf{z}, w)F_\Delta(y^*|\mathbf{z}) \tag{A.34}$$

Differentiating (A.34) with respect to w,

$$\frac{\partial}{\partial w}K(y|\mathbf{z}, w) = \frac{\partial}{\partial w}\lambda(\mathbf{z}, w)\cdot F_\Delta(y^*|\mathbf{z}) \tag{A.35}$$

holds for every $y \in \mathcal{Y}$. Similarly, differentiating (2.10) with z_1 and z_2 at $y_1^* \in \mathcal{Y}_1(x, w, \mathbf{z})$ and $y_2^* \in \mathcal{Y}_2(x, w, \mathbf{z})$, respectively, by Assumption 2.15 we obtain

$$\frac{\partial}{\partial z_1}K(y_1^*|\mathbf{z}, w) = \frac{\partial}{\partial z_1}\lambda(\mathbf{z}, w)\cdot F_\Delta(y_1^*|\mathbf{z}) + \lambda(\mathbf{z}, w)\frac{\partial}{\partial z_1}F_1(y_1^*|\mathbf{z}) \tag{A.36}$$

and

$$\frac{\partial^2}{\partial w\partial z_1}K(y_1^*|\mathbf{z}, w) = \frac{\partial^2}{\partial w\partial z_1}\lambda(\mathbf{z}, w)\cdot F_\Delta(y_1^*|\mathbf{z}) + \frac{\partial}{\partial w}\lambda(\mathbf{z}, w)\frac{\partial}{\partial z_1}F_1(y_1^*|\mathbf{z}), \tag{A.37}$$

and similar results for $(\partial/\partial z_1)K(y_1^*|\mathbf{z}, w)$ and $(\partial^2/\partial w\partial z_2)K(y_2^*|\mathbf{z}, w)$.

By (A.37) and (A.35)

$$\begin{aligned}\frac{\partial}{\partial z_1}F_1(y_1^*|\mathbf{z}) &= \frac{1}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\Big(\frac{\partial^2}{\partial w\partial z_1}K(y_1^*|\mathbf{z}, w) - \frac{\partial^2}{\partial w\partial z_1}\lambda(\mathbf{z}, w)\cdot F_\Delta(y_1^*|\mathbf{z})\Big) \\ &= \frac{1}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\Big(\frac{\partial^2}{\partial w\partial z_1}K(y_1^*|\mathbf{z}, w) - \frac{\partial^2}{\partial w\partial z_1}\lambda(\mathbf{z}, w)\frac{(\partial/\partial w)K(y_1^*|\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\Big).\end{aligned}$$

Using this and (A.35) in (A.36),

$$\begin{aligned}\frac{\partial}{\partial z_1}K(y_1^*|\mathbf{z}, w) = &\frac{(\partial/\partial z_1)\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\frac{\partial}{\partial w}K(y_1^*|\mathbf{z}, w) + \frac{\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)} \\ &\times\Big(\frac{\partial^2}{\partial w\partial z_1}K(y_1^*|\mathbf{z}, w) - \frac{\partial^2}{\partial w\partial z_1}\lambda(\mathbf{z}, w)\frac{(\partial/\partial w)K(y_1^*|\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\Big)\end{aligned}$$

$$= \frac{1}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\left(\frac{\partial}{\partial z_1}\lambda(\mathbf{z}, w) - \frac{\lambda(\mathbf{z}, w)(\partial^2/\partial w \partial z_1)\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\right)$$
$$\times \frac{\partial}{\partial w}K(y_1^*|\mathbf{z}, w) + \frac{\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\frac{\partial^2}{\partial w \partial z_1}K(y_1^*|\mathbf{z}, w)$$
$$:= \beta_1(\mathbf{z}, w)\frac{\partial}{\partial w}K(y_1^*|\mathbf{z}, w) + \frac{\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\frac{\partial^2}{\partial w \partial z_1}K(y_1^*|\mathbf{z}, w),$$

with

$$\beta_1(\mathbf{z}, w) := \frac{1}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\left(\frac{\partial}{\partial z_1}\lambda(\mathbf{z}, w) - \frac{\lambda(\mathbf{z}, w)(\partial^2/\partial w \partial z_1)\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}\right).$$

Evaluating this at $(y_a, y_b) \in \mathcal{K}_1 \cap (\mathcal{Y}_1(x, w, \mathbf{z}) \times \mathcal{Y}_1(x, w, \mathbf{z}))$

$$\begin{pmatrix} \frac{\partial}{\partial z_1}K(y_a|\mathbf{z}, w) \\ \frac{\partial}{\partial z_1}K(y_b|\mathbf{z}, w) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial w}K(y_a|\mathbf{z}, w) & \frac{\partial^2}{\partial w \partial z_1}K(y_a|\mathbf{z}, w) \\ \frac{\partial}{\partial w}K(y_b|\mathbf{z}, w) & \frac{\partial^2}{\partial w \partial z_1}K(y_b|\mathbf{z}, w) \end{pmatrix} \begin{pmatrix} \beta_1(\mathbf{z}, w) \\ \frac{\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)} \end{pmatrix}.$$

Under Assumptions 2.15 and 2.16

$$\Delta_1 = \frac{\partial}{\partial w}K(y_a|\mathbf{z}, w)\frac{\partial^2}{\partial w \partial z_1}K(y_b|\mathbf{z}, w) - \frac{\partial^2}{\partial w \partial z_1}K(y_a|\mathbf{z}, w)\frac{\partial}{\partial w}K(y_b|\mathbf{z}, w) \neq 0$$

and

$$\frac{\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)} = \frac{N_1}{\Delta_1}, \tag{A.38}$$

where

$$N_1 = -\frac{\partial}{\partial z_1}K(y_a|\mathbf{z}, w)\frac{\partial}{\partial w}K(y_b|\mathbf{z}, w) + \frac{\partial}{\partial z_1}K(y_b|\mathbf{z}, w)\frac{\partial}{\partial w}K(y_a|\mathbf{z}, w).$$

Proceeding similarly, with the roles of the first and the second mixture components switched,

$$\frac{1 - \lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)} = \frac{N_2}{\Delta_2}, \tag{A.39}$$

where

$$N_2 = -\frac{\partial}{\partial z_2}K(y_c|\mathbf{z}, w)\frac{\partial}{\partial w}K(y_d|\mathbf{z}, w) + \frac{\partial}{\partial z_2}K(y_d|\mathbf{z}, w)\frac{\partial}{\partial w}K(y_c|\mathbf{z}, w)$$

and

$$\Delta_2 = -\frac{\partial}{\partial w}K(y_c|\mathbf{z}, w)\frac{\partial^2}{\partial w \partial z_2}K(y_d|\mathbf{z}, w) + \frac{\partial^2}{\partial w \partial z_2}K(y_c|\mathbf{z}, w)\frac{\partial}{\partial w}K(y_d|\mathbf{z}, w) \neq 0.$$

By (A.38) and (A.39),

$$\frac{1}{(\partial/\partial w)\lambda(\mathbf{z}, w)} = \frac{\lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)} + \frac{1 - \lambda(\mathbf{z}, w)}{(\partial/\partial w)\lambda(\mathbf{z}, w)}$$
$$= \frac{N_1}{\Delta_1} + \frac{N_2}{\Delta_2},$$

yielding

$$\frac{\partial}{\partial w}\lambda(\mathbf{z}, w) = \frac{\Delta_1\Delta_2}{N_1\Delta_2 + N_2\Delta_1}.$$

By (A.38)

$$\lambda(\mathbf{z}, w) = \frac{N_1\Delta_2}{N_1\Delta_2 + N_2\Delta_1}.$$

Moreover, for every $y \in \mathcal{Y}$, by (A.35)

$$F_\Delta(y|\mathbf{z}) = \frac{N_1\Delta_2 + N_2\Delta_1}{\Delta_1\Delta_2}\frac{\partial}{\partial w}K(y|\mathbf{z}, w) \tag{A.40}$$

and by (A.34)

$$\begin{aligned} F_2(y|\mathbf{z}) &= K(y|\mathbf{z}, w) - \lambda(\mathbf{z}, w)F_\Delta(y|\mathbf{z}) \\ &= K(y|\mathbf{z}, w) - \frac{N_1}{\Delta_1}\frac{\partial}{\partial w}K(y|\mathbf{z}, w), \end{aligned}$$

also yielding

$$\begin{aligned} F_1(y|\mathbf{z}) &= F_\Delta(y|\mathbf{z}) + F_2(y|\mathbf{z}) \\ &= \frac{N_1\Delta_2 + N_2\Delta_1}{\Delta_1\Delta_2}\frac{\partial}{\partial w}K(y|\mathbf{z}, w) + K(y|\mathbf{z}, w) - \frac{N_1}{\Delta_1}\frac{\partial}{\partial w}K(y|\mathbf{z}, w) \\ &= K(y|\mathbf{z}, w) - \frac{N_2}{\Delta_2}\frac{\partial}{\partial w}K(y|\mathbf{z}, w). \end{aligned}$$

□